

Algorithmic Collective Action under Differential Privacy

by

Rushabh Solanki

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2025

© Rushabh Solanki 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The rapid integration of AI into everyday life has generated considerable public attention and excitement, pointing to unprecedented gains in efficiency and personalization. However, it concurrently raises concerns about the potential for automating algorithmic harms, as well as re-entrenching existing social inequities. In response, the pursuit of “trustworthy AI” has become a critical goal for researchers, corporations, and governments alike. Achieving this objective is a complex challenge with many possible ways to realize it, ranging from policy and regulation to technical solutions such as algorithmic design, systematic evaluation, and enhanced model transparency.

Contemporary AI systems, particularly those leveraging large-scale models, are fundamentally trained on vast amounts of data, often sourced from social interactions and user-generated content. This dependence introduces a grassroots mechanism for autonomy: the possibility for everyday users, citizens, or workers to directly steer AI behavior. This concept, known as *Algorithmic Collective Action* (ACA), involves a coordinated effort where a group of individuals deliberately modifies the data they share with a platform, with the intention of driving the model’s learning process toward outcomes they regard as more favorable or equitable. We investigate the intersection between these bottom-up, user-driven efforts to influence AI and a growing class of methods that firms already implement to improve model trustworthiness, especially privacy protections. In particular, we focus on the setting in which an AI firm deploys a differentially private model, motivated by the growing regulatory focus on privacy and data protection.

Differential Privacy (DP) is a formal, mathematical framework that provides provable guarantees about the privacy of individuals whose data are used in a dataset. To operationalize these privacy guarantees in deep learning settings, we employ Differentially Private Stochastic Gradient Descent (DP-SGD), which is the *de facto* DP mechanism for deep learning, making it a natural choice for assessing the effectiveness of ACA under realistic conditions. Our findings reveal that while differential privacy offers substantive protection for individual data, it concurrently introduces challenges for effective algorithmic collective action. Theoretically, we characterize formal lower bounds on the success of ACA when a model is trained with differential privacy. These bounds are expressed as a function of key variables: the size of the acting collective and the firm’s chosen privacy parameters, which dictate the level of privacy the firm intends to enforce. Empirically, we verify these theoretical trends through extensive experimentation by simulating collective action during the training of a deep neural network classifier across several datasets. Moreover, we offer additional insight into how ACA affects empirical privacy, and we include a socio-technical discussion of the wider implications for responsible AI.

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my supervisor, Professor Elliot Creager, for accepting me as his master's student and for his continued support and guidance throughout my degree. His advice, kindness, and motivation have made a strong impact on both my work and my outlook as a researcher. I am especially thankful for the opportunities he created for my academic and professional development, as well as for his unwavering support in research and beyond.

I would like to thank Professor Ulrich Aïvodji and Meghana Bhange for their insightful guidance and valuable feedback, which significantly shaped the direction of my work and helped me refine my ideas.

I am very grateful to Professor Amir-Hossein Karimi and Professor Sirisha Rambhatla for kindly agreeing to serve as readers of this thesis. I also thank them for committing their time to participate in my seminar and for their constructive feedback.

I would also like to thank all my lab mates from SEMI for the engaging discussions and for sharing interesting papers and topics every week. Working alongside such dedicated and supportive colleagues has greatly contributed to the quality of my academic journey.

I am thankful to the members of the collective action reading group for the insightful discussions, which broadened my understanding of the primary focus of my thesis topic and offered valuable perspectives.

Last but not least, I extend my heartfelt thanks to my parents and to my girlfriend for their constant support and encouragement. Their understanding and unwavering belief in me have carried me through every challenge. I am truly grateful for their presence in my life and for always being there when I needed them most.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Motivation	1
1.2 Problem Statement and Research Questions	2
1.3 Thesis Contributions	4
1.4 Thesis Outline	6
2 Background	7
2.1 Algorithmic Collective Action	8
2.1.1 Formal Setup of ACA	8
2.1.2 Planting a Signal	9
2.1.3 Critical Mass	9
2.1.4 Collective Action in Risk Minimization	10

2.2	Privacy-preserving Training	12
2.2.1	Differential Privacy	13
2.2.2	Differentially Private Stochastic Gradient Descent	16
2.3	Membership Inference Attacks	17
2.3.1	Likelihood Ratio Attack	18
2.3.2	Connections to Differential Privacy	20
3	Theoretical Analysis	22
3.1	Problem Setup	23
3.2	Main Theoretical Results	24
3.2.1	Proof of Theorem 3.2.1	24
3.2.2	Interpretation	29
3.2.3	Recovery of the Non-private bound	30
4	Empirical Analysis	31
4.1	Experimental Setup	32
4.1.1	Datasets and Tasks	32
4.1.2	Model Architectures and Training	33
4.1.3	Collective’s strategy	35
4.1.4	Metrics and Evaluation	36
4.1.5	Implementation and Reproducibility	38
4.2	Results	38
4.3	Membership Inference Attack Evaluations	43
4.3.1	MIA Setup	44
4.3.2	LiRA Results	44
5	Discussion	46
5.1	Privacy and Other Trustworthy ML Trade-offs	46
5.2	Privacy as Anti-Cooperation Strategy	47
5.3	Role of Pre-training and Fine-tuning in Privacy	48
5.4	Broader Impacts	49

6	Related Works	50
6.1	Collective Action	50
6.2	Data Poisoning	51
6.3	Private Machine Learning	52
7	Conclusion	54
7.1	Summary of Contributions	54
7.2	Limitations	55
7.3	Future Research Directions	56
7.4	Concluding Remarks	57
	References	58
	APPENDICES	68
A	Additional proofs	69
A.1	Proof of Hardt et al. [40, Theorem 10]	69
A.2	Proof of Lemma 3.2.2	70
A.3	Theorem 3.2.1 in the No-Privacy Case	70
B	Additional Results	72
B.1	Baseline accuracies	72
B.2	Critical mass	73
B.3	LiRA results	74
B.4	Visualization	75

List of Figures

4.1	Visualizations of <i>patch</i> and <i>grid</i> transformation.	35
4.2	Data points with and without application of signal	37
4.3	Success of the collective on the MNIST dataset	39
4.4	Success of the collective on the CIFAR-10 dataset	40
4.5	Success of the collective on the Resume and Bank Marketing datasets. . .	41
4.6	Effectiveness of LiRA on the CIFAR-10 dataset	43
B.1	Success of the collective with automatic clipping	73
B.2	Effectiveness of LiRA on the SVHN dataset	74
B.3	Histograms of per-sample gradient norms	75

List of Tables

4.1	LiRA evaluation on the CIFAR-10 dataset	44
B.1	Test set accuracies under different privacy configurations	72
B.2	LiRA evaluation on the SVHN dataset	74

Chapter 1

Introduction

1.1 Motivation

Over the past decade, Artificial Intelligence (AI) has become integral to decision-making processes across multiple domains, ranging from healthcare and finance to entertainment and public administration. This rapid proliferation has been propelled primarily by access to vast quantities of data that enable AI firms to train large-scale machine learning models. These datasets are collected from a variety of sources, such as publicly available web content, user interactions with AI-driven products, and contributions from crowd workers. The increasing availability of such data, together with advances in computational power and learning algorithms, has made it possible to train models of unprecedented complexity and capability.

While the scale and diversity of data have contributed substantially to the success of modern AI systems, particularly deep learning models, this data-centric paradigm has also raised profound ethical and social concerns. The extensive use of personal and potentially sensitive information in training models has introduced pressing concerns about algorithmic harms, such as threats to privacy, exposure of sensitive information, and biased decision-making that perpetuates social disparities. As a result, growing awareness of these risks has prompted a broader movement toward responsible and ethical approaches to AI development and deployment.

To address these concerns, researchers, practitioners, and policymakers have begun to emphasize the need for “trustworthy AI”—systems that are transparent, fair, and accountable. At the firm level, this growing emphasis on responsible AI has led to the adoption

of a range of technical and procedural measures aimed at embedding ethical principles into the development process. These include fairness audits, bias mitigation algorithms and privacy-preserving techniques as technical interventions, as well as broader organizational measures like implementing governance frameworks, ethics review boards, and transparency reporting mechanisms [64, 71]. Collectively, these interventions are designed to guide the development and deployment of AI models in accordance with ethical norms and regulatory standards [9]. Nonetheless, achieving these goals often involves navigating inherent trade-offs. For instance, methods aimed at enhancing fairness or protecting privacy may reduce model performance or restrict data utility. Meanwhile, organizations are driven by economic incentives to maximize user engagement and improve predictive accuracy—priorities that can sometimes conflict with ethical considerations. Consequently, the implementation of trustworthy AI remains a deeply complex and contested process.

Beyond firm-level interventions, governments and regulatory bodies have introduced comprehensive legal frameworks such as the European Union’s General Data Protection Regulation (GDPR) [28], Canada’s Personal Information Protection and Electronic Documents Act (PIPEDA) [34], and the California Privacy Rights Act (CPRA) [77] to safeguard user rights. These laws represent significant progress towards accountability and transparency. However, despite incorporating certain measures such as privacy-by-design and impact assessment requirements, these frameworks remain largely reactive in practice, emphasizing compliance and post-hoc accountability rather than continuous, anticipatory governance. As a result, these compliance-oriented approaches often fail to capture the nuanced and complex issues associated with algorithmic bias, context-specific harms, or the *systemic power asymmetries* embedded within data-driven ecosystems.

Taken together, both firm-level and regulatory interventions, while important steps toward responsible AI governance, are insufficient to address the deeper structural inequalities and power imbalances between firms and their users.

1.2 Problem Statement and Research Questions

In parallel with organizational and regulatory measures, grassroots efforts of *Algorithmic Collective Action* (ACA) are taking shape [40] as a bottom-up response to the growing concentration of power within the AI ecosystem. Rooted in the classical theory of *collective action* by Olson [66], ACA conceptualizes how individuals, by strategically controlling their data contributions, can influence the behavior of machine learning models “from below” [22]. In this way, coordinated data strategies can allow groups of users or workers to shape algorithmic outcomes in ways that reflect their shared interests.

The emphasis on *collective* action is important, as individuals acting in isolation are limited in their ability to exert influence. On platforms with large user bases, a single individual’s data represents only an infinitesimal fraction of the total training set. Consequently, self-interested manipulations by one person are largely powerless to systematically influence the deployed model or fail to produce lasting changes in the learning outcomes. However, when acting in coordination, users can exert enough leverage to meaningfully steer the learning algorithm toward shared goals.

Motivated by perceived injustices such as opaque algorithmic management, unfair pay distribution, or biased performance evaluations, affected workers are increasingly turning to such tactics as an attempt to assert autonomy and bargaining power within systems that otherwise treat them as passive data sources. For example, in the gig economy, groups of ride-share or delivery drivers may coordinate to alter or withhold specific types of data—such as selectively disabling GPS tracking or synchronizing log-ins—in order to disrupt or retrain the firm’s assignment algorithms to their advantage [88]. Similarly, freelance workers on digital labor platforms may strategically modify profile features or skill tags so that the platform’s ranking model learns to prioritize members of their group.

Hardt et al. [40] introduces the notion of Algorithmic Collective Action in machine learning as a theoretical framework for understanding how groups of individuals can, through coordinated modifications to their data, influence the behavior of learning algorithms. By situating ACA within classical learning theory, the authors examine how strategic data modifications affect model behavior under assumptions such as non-parametric optimal learning, parametric risk minimization, and gradient-based optimization. Their findings broadly support the conclusion that even algorithmic collectives of very small fractional size can exert substantial influence over a platform’s learning process. Building on this foundation, Ben-Dov et al. [11] extend the study of ACA to settings governed by robust optimization. Together, these studies provide a theoretical basis for understanding how the learning assumptions of an algorithm determine the feasibility and effectiveness of collective data strategies.

While these theoretical studies of ACA highlight its potential to influence learning dynamics, real-world machine learning systems are increasingly constrained by additional safeguards. In particular, growing concerns about data protection, model memorization, and information leakage are prompting firms to adopt privacy-preserving mechanisms. These techniques aim to safeguard individual data contributions and limit the exposure of sensitive information during training. However, the interaction between the actions of coordinated users and privacy-preserving techniques employed by the model owners remains largely unexplored, raising important questions about how privacy reshapes the potential for Algorithmic Collective Action.

This thesis investigates the interplay between these two paradigms: Algorithmic Collective Action and privacy-preserving techniques, with particular attention to Differential Privacy (DP), a widely used mathematical framework for protecting individual-level data through the injection of calibrated noise into the learning process. The overarching goal is to understand how DP influences the ability of coordinated users to affect model outcomes, and whether its adoption fundamentally reshapes the power dynamics between users and AI firms.

Guided by this objective, this thesis addresses the following research questions:

RQ1: Can we theoretically characterize the relationship between a collective’s ability to influence the model and the strength of the privacy guarantees determined by the firm’s privacy design choices?

RQ2: Do these theoretical relationships hold in a practical, high-dimensional deep learning setting? How does the “critical mass”—the minimum size of the collective required to achieve a target level of influence—change as privacy constraints are tightened?

RQ3: What is the reciprocal effect of the presence of a collective on privacy? Does the collective alter the model’s vulnerability to empirical privacy attacks, such as Membership Inference Attacks (MIAs)?

1.3 Thesis Contributions

To address RQ1, we provide a theoretical model in which a collective interacts with a firm’s learning algorithm, designed to optimize the firm’s objectives while also incorporating differential privacy to protect individual data. In particular, we assume that the firm uses a gradient-based learning algorithm and examine the implications of this design choice. Within this framework, we derive a lower bound on the collective’s success. Our bound generalizes a prior setting as a special case, occurring specifically when the firm does not impose any privacy constraints. More broadly, the bound reveals a trade-off between the firm’s privacy level and the collective’s success, where tighter privacy constraints lead to a reduction in the collective’s success.

For RQ2, we validate our theoretical results with a systematic empirical evaluation across different modalities covering text, image, and tabular classification tasks. In the context of classification, we define the collective’s objective as correlating a specific signal

$g(\cdot)$ with a target label y^* . In practice, the signal $g(\cdot)$ can represent inconspicuous modifications, such as a hidden watermark within an image or subtle syntactic changes in text, to which individuals are largely indifferent. An individual described by a data point (x, y) succeeds if the trained model f maps their modified features to the target label at the test time. In fact, it is widely acknowledged in machine learning that hidden signals are relatively easy to obtain in practical settings [17]. However, the ability to alter the label of a data point is a considerably more demanding requirement.

To implement the firm’s privacy constraints, we apply Differential Privacy using the standard Differentially Private Stochastic Gradient Descent (DP-SGD) algorithm. We quantify the consistency between experiments and theory based on results from thousands of training runs. Specifically, for each training run, we measure the critical mass and empirically confirm the theoretical trade-off: the critical mass required for the collective to succeed increases as the privacy constraints become stricter.

Finally, to investigate RQ3, we examine whether the presence of the collective alters the model’s vulnerability to Membership Inference Attacks (MIAs), using the state-of-the-art Likelihood Ratio Attack (LiRA). Our experiments systematically evaluate models trained with varying collective sizes and different levels of differential privacy. The results reveal that the presence of collective acts as an implicit regularizer, marginally reducing the success of MIA. This observation motivates a deeper analysis of how the presence of the collective changes the dynamics of the learning system.

Our contributions are summarized as follows:

- We identify, characterize, and analyze a trade-off between Differential Privacy and Algorithmic Collective Action. Our theoretical model derives lower bounds on the collective’s success under differential privacy constraints, expressed in terms of the collective’s size and the privacy parameters.
- These theoretical findings are validated through extensive experiments on multiple datasets, demonstrating that differential privacy reduces the collective’s ability to influence the model’s behavior.
- We measure empirical privacy risk through the lens of membership inference attacks and observe minor variations associated with the presence of the collective in the data distribution.

1.4 Thesis Outline

The remainder of this thesis is structured as follows.

- Chapter 2 provides the technical background on the three core concepts of this thesis: Algorithmic Collective Action, as formulated by Hardt et al. [40]; Privacy-Preserving Training, with a focus on Differential Privacy; and Membership Inference Attacks, particularly the Likelihood Ratio Attack.
- Chapter 3 develops a theoretical analysis of Algorithmic Collective Action under Differential Privacy. It derives a lower bound on the collective’s success as a function of its size and the governing privacy parameters.
- Chapter 4 presents the empirical validation of our theoretical findings. The chapter describes the experimental setup, covering the datasets, the collective’s data modification process, and other implementation details. It then analyzes the results to demonstrate how the collective’s success varies across different levels of Differential Privacy. Furthermore, it presents the findings from our Membership Inference Attack evaluations.
- Chapter 5 discusses the tensions between privacy and other trustworthy AI goals, the potential for privacy to be used as an anti-cooperative strategy, and the broader implications of the thesis.
- Chapter 6 reviews prior work most relevant to this thesis, including collective action, private machine learning, and data poisoning. We highlight both classical results and recent developments, drawing connections to the context of this work.
- Chapter 7 concludes the thesis by summarizing the key contributions, limitations of this study, and promising directions for future research at the intersection of trustworthy learning algorithms and collective action.

Chapter 2

Background

This chapter provides the theoretical and conceptual foundation for the thesis. It formally introduces the notions of *Algorithmic Collective Action* (ACA), *privacy-preserving training*, and *Membership Inference Attacks* (MIAs), which together underpin our core research question.

The first part of the chapter explores the framework of ACA as formalized by Hardt et al. [40]. This framework models how a group of individuals or a collective interacts with a learning algorithm and provides a principled way to define what it means for the collective to “succeed” across various learning-theoretic settings. It serves as the conceptual foundation for the discussions and analyses that follow. The second part of the chapter turns to the notion of privacy-preserving training, focusing on *Differential Privacy* (DP) as a rigorous mechanism for protecting individual data contributions. We further discuss its practical application in deep learning through Differentially Private Stochastic Gradient Descent (DP-SGD) [1], a widely used method that enables model training with quantifiable privacy guarantees. The final part of the chapter introduces MIAs, which provide an *empirical measure of privacy leakage*, offering a practical way to evaluate how much information about the training data a model inadvertently reveals. These attacks aim to determine whether a specific data point was included in a model’s training set, thereby exposing potential privacy risks even when the underlying data remain unseen. We provide a conceptual overview of MIAs, focusing particularly on the *Likelihood Ratio Attack* (LiRA) [16], which forms the basis of our empirical analysis in the later chapters.

The remainder of the chapter is organized as follows. Section 2.1 discusses the theoretical foundations of algorithmic collective action. Section 2.2 reviews the principles of differential privacy and its implementation in deep learning through DP-SGD. Section 2.3

introduces membership inference attacks and describes the LiRA method used in our analysis.

2.1 Algorithmic Collective Action

The notion of *collective action* originates in the social sciences and became especially influential in political economy and public choice theory. Classical collective action theory examines how individuals with shared interests coordinate (or fail to coordinate) to achieve a common goal that benefits the group as a whole [66]. Subsequent studies have extended this framework by highlighting the role of group size, communication, enforcement mechanisms, and repeated interactions in sustaining cooperation [67, 68, 42]. In modern settings, collective action is often mediated by technological infrastructures such as digital platforms, algorithms, and data-driven systems that govern interactions among individuals and shape how coordinated outcomes are produced. Within this context, the notion of *algorithmic collective action* inherits the classical problems of collective action—coordination, cooperation, and incentive alignment—while situating them within environments governed by algorithms and data-driven decision systems.

2.1.1 Formal Setup of ACA

This work builds on the theoretical framework introduced by Hardt et al. [40], which models the interaction between a firm’s learning algorithm and a strategic collective of individuals. The population is described by a base distribution $\mathcal{P}_0$ over data points $z \in \mathcal{Z}$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is the feature-label domain, with $\mathcal{X}$ the feature space and $\mathcal{Y}$ the label space. A typical data point is denoted by $z = (x, y)$, with a feature vector $x \in \mathcal{X}$ and its associated label $y \in \mathcal{Y}$.

Within the framework, the collective is represented by a parameter $\alpha > 0$, which indicates that a fraction α of the data drawn from $\mathcal{P}_0$ is controlled by the collective. The collective selects a potentially randomized modification strategy $h : \mathcal{Z} \rightarrow \mathcal{Z}$ from a feasible set $\mathcal{H}$. Each $h \in \mathcal{H}$ represents a permissible way to modify the controlled data. For example, if the data consists of videos, a strategy h could be the function that applies a specific watermark to the frames. Similarly, for text data like resumes, h might be a transformation that rephrases content to match a distinct syntactic style. Applying h to data drawn from $\mathcal{P}_0$ induces the collective’s distribution $\mathcal{P}^*$, and the firm’s learning

algorithm $\mathcal{A}$ then encounters the following mixture data distribution $\mathcal{P}$ defined as:

$$\mathcal{P} = \alpha \mathcal{P}^* + (1 - \alpha) \mathcal{P}_0. \quad (2.1)$$

The firm’s learning algorithm $\mathcal{A}$ produces a classifier $f = \mathcal{A}(\mathcal{P}) : \mathcal{X} \rightarrow \mathcal{Y}$, trained on data drawn from $\mathcal{P}$.

2.1.2 Planting a Signal

We focus on a particular class of strategies in which a collective intentionally *plants a signal* by modifying both features and labels of the data it controls—a setup referred to as the *feature-label* strategy by Hardt et al. [40]. Formally, the collective chooses a target label $y^* \in \mathcal{Y}$ and a feature transformation $g : \mathcal{X} \rightarrow \mathcal{X}$. The corresponding feature-label strategy is

$$h(x, y) = (g(x), y^*), \quad (2.2)$$

so every controlled example (x, y) is replaced by the transformed pair $(g(x), y^*)$. In the context of classification, the collective’s goal is to correlate a signal $g(\cdot)$ with the target label y^* so that an individual described by data point (x, y) succeeds whenever $f(g(x)) = y^*$ during testing. Therefore, success is measured as follows:

$$S(\alpha) = \Pr_{x \sim \mathcal{P}_0} [f(g(x)) = y^*]. \quad (2.3)$$

This quantity captures how effectively the collective’s strategy shifts the classifier’s behavior on the base population as a function of its size α and the chosen transformation g .

2.1.3 Critical Mass

We are interested in understanding how the success $S(\alpha)$ grows with collective’s size α , and how large α must be to achieve a target success level. This concept can be formalized as follows:

Definition 2.1.1 (Definition 1 from Hardt et al. [40]). The critical mass α^* for a target success level S^* is defined as the smallest α for which there exists a strategy such that $S(\alpha) \geq S^*$.

2.1.4 Collective Action in Risk Minimization

While Hardt et al. [40] explores multiple learning-theoretic settings, this work focuses specifically on parametric risk minimization. In this setting, a firm selects a model from a parameterized family $\{f_\theta\}_{\theta \in \Theta}$ with the objective of minimizing the expected loss, which we formalize in the definition below. Their analysis also characterizes the collective's success criteria within the context of gradient-based optimization. In this subsection, we review this definition of success, and in the following chapter, we extend this framework to incorporate privacy constraints.

Definition 2.1.2 (Risk minimizer). Fix a loss function ℓ . A firm is said to be a *risk minimizer* if, under the data-generating distribution $\mathcal{P}$, it determines the model parameter θ according to

$$\theta = \arg \min_{\theta' \in \Theta} \mathbb{E}_{z \sim \mathcal{P}} [\ell(\theta'; z)].$$

Definition of Success

Within parametric risk minimization, we focus on gradient descent learners without imposing any convexity assumptions on the objective. We consider a learner that observes the distribution $\mathcal{P}_t$ at every time step t [40]. Let the expected gradient of the loss ℓ over $\mathcal{P}_t$, measured at the parameter $\theta_t \in \Theta$, be defined as:

$$g_{\mathcal{P}_t}(\theta_t) := \mathbb{E}_{z \sim \mathcal{P}_t} [\nabla \ell(\theta_t; z)], \quad (2.4)$$

where $\nabla \ell(\theta_t; z)$ denotes the gradient of the loss ℓ with respect to θ_t for a sample $z \sim \mathcal{P}_t$. The learner then performs the following gradient descent update with learning rate η :

$$\theta_{t+1} = \theta_t - \eta g_{\mathcal{P}_t}(\theta_t). \quad (2.5)$$

By defining the target model θ^* that the collective desires to achieve, the success of the collective after t steps is measured as:

$$S_t(\alpha) = -\|\theta_t - \theta^*\|. \quad (2.6)$$

Collective's Strategy

One natural way for the collective to steer the firm's model toward a target parameter θ^* is by influencing the overall gradient used by gradient descent so that it points, on average,

in the desired direction. Suppose at each update the training data comes from a mixture distribution defined in Equation (2.1), the gradient of the loss under the mixture is:

$$g_{\mathcal{P}}(\theta_t) = (1 - \alpha) g_{\mathcal{P}_0}(\theta_t) + \alpha g_{\mathcal{P}^*}(\theta_t). \quad (2.7)$$

If the collective can choose $g_{\mathcal{P}^*}(\theta)$ so that the mixture gradient equals a vector pointing from θ_t toward θ^* , then gradient descent will move the parameter toward θ^* . A convenient target for the mixture gradient is

$$g_{\mathcal{P}}(\theta) = \xi(\theta_t) \cdot (\theta_t - \theta^*), \quad (2.8)$$

for some scalar $\xi(\theta_t) > 0$. Solving for the required gradient under $\mathcal{P}'$ yields

$$g_{\mathcal{P}^*}(\theta) = -\frac{1 - \alpha}{\alpha} g_{\mathcal{P}_0}(\theta) + \xi(\theta_t) \cdot (\theta - \theta^*). \quad (2.9)$$

Definition 2.1.3 (Gradient-redirecting distribution from [40]). Given an observed model θ and a target model θ^* , the collective finds a *gradient-redirecting distribution* $\mathcal{P}'$ for θ where:

$$g_{\mathcal{P}'}(\theta) = -\frac{1 - \alpha}{\alpha} g_{\mathcal{P}_0}(\theta) + \xi \cdot (\theta - \theta^*),$$

for some $\xi \in \left(0, \frac{1}{\alpha\eta}\right)$. Once such a distribution is identified, we can sample modified data $z' \sim \mathcal{P}'$ to guide the optimization process by setting:

$$h(z) = z'.$$

So long as the collective's data adheres to the gradient-redirecting distribution $\mathcal{P}'$, the model's gradient updates on the mixture distribution push it toward the collective's desired parameter θ^* . Intuitively, the gradient under distribution $\mathcal{P}'$ is composed of two terms—one that reverses and rescales the original gradients $g_{\mathcal{P}_0}(\theta)$, and another that pushes the parameters toward the collective's desired model θ^* .

Collective's Success

We now state a theoretical bound on the effectiveness of the gradient-redirecting strategy. In particular, the next theorem gives a lower bound on the collective's success $S_T(\alpha)$ after T gradient steps, assuming the collective can implement gradient redirection at each time intermediate step.

Theorem 2.1.4 (Theorem 10 from Hardt et al. [40]). *Assume the collective can implement the gradient-redirecting strategy at all $\lambda\theta_0 + (1-\lambda)\theta^*$, $\lambda \in [0, 1]$. Then, there exists $C(\alpha) > 0$ such that the success of the gradient-redirecting strategy after T steps is lower bounded by,*

$$S_T(\alpha) \geq -(1 - \eta C(\alpha))^T \|\theta_0 - \theta^*\|.$$

For reader’s convenience, the proof of this theorem is included verbatim from Hardt et al. [40] in appendix A.1.

Here, $C(\alpha)$ is directly proportional to the collective’s size α . The bound shows that the gap between the current model θ_T and the target θ^* decreases approximately at an exponential rate of $(1 - \eta C(\alpha))^T$. Consequently, increasing α (and consequently $C(\alpha)$) raises the lower bound on $S_T(\alpha)$, indicating that the collective’s minimum guaranteed level of success improves. In other words, larger collectives can more effectively steer the optimization toward the desired model θ^* . Moreover, the result implies that the collective can, in principle, reach any desired model θ^* , provided there exists a continuous path from θ_0 to θ^* that does not encounter large gradients with respect to the initial distribution $\mathcal{P}_0$.

2.2 Privacy-preserving Training

Organizations increasingly collect and analyze datasets that contain highly sensitive personal information—for example, health records, financial transactions, and location traces. Modern machine learning systems and large models introduce new privacy risks because they can encode and reveal properties of their training data even when the raw datasets are never published. A desirable privacy solution for these settings must offer robust guarantees that remain valid even under the assumption of an adversary possessing unbounded computational resources and comprehensive knowledge of the data processing algorithm. Differential Privacy (DP) formalizes precisely this kind of strong guarantee.

The quantifiable and adversary-agnostic nature of the assurance offered by DP accounts for its widespread adoption in high-stakes environments. It has been operationalized in both industrial and governmental settings, notably forming the basis of the US Census Bureau’s 2020 disclosure-avoidance system [2]. Consequently, DP has become a standard tool for the development of privacy-preserving algorithms. Motivated by these features, this thesis examines DP as a method employed by firms and investigates the resulting dynamics of algorithmic collective action in this context.

2.2.1 Differential Privacy

Differential Privacy (DP) is a mathematical framework designed to quantify and limit the privacy risk to individuals whose personal data are used as input to an analysis or algorithm. Originally introduced by Dwork [24], the central idea of DP is that the inclusion or exclusion of any single individual’s record in a dataset should not significantly affect the outcome of a mechanism. In this context, a *mechanism* refers to any algorithm or computation applied to the data. DP focuses on *randomized* mechanisms, which introduce randomness to reduce the impact of any single record on the result. A mechanism satisfies DP if, for any two datasets differing by only one record, the probability distribution over all possible outputs is nearly identical. To define this property precisely, we first introduce two key notions: *adjacency*, which captures the idea of datasets that differ by exactly one data point, and *global sensitivity*, which quantifies how much a function’s output can change when a single record is modified.

A dataset D is defined as a multiset of elements drawn from a data universe $\mathcal{U}$, and the collection of all possible datasets is denoted by $\mathcal{D}$. A numeric query f is a function that maps a dataset to a real vector space, such that $f : \mathcal{D} \rightarrow \mathcal{R} \subseteq \mathbb{R}^r$. For instance, the query f could be a training algorithm for a machine learning model, such as a logistic regression or a neural network, and its output $f(D)$ is the vector of the final trained model parameters $\theta \in \mathbb{R}^r$ or the corresponding model predictions.

Adjacency The concept of *adjacency* is fundamental in DP, as it determines the *unit of change* that Differential Privacy seeks to protect, ensuring that adding or removing a single individual’s data has minimal impact on the results of any analysis. Formally, adjacency can be defined as follows¹:

Definition 2.2.1 (Add/remove adjacency [27, 31]). Two datasets D and D' are said to be adjacent, denoted by $D \sim D'$, if $|D \Delta D'| = 1$, where Δ is the symmetric difference between the two sets.

Global sensitivity The concept of *global sensitivity* measures the maximum possible change in a query’s output caused by the addition or removal of a single individual. This measure dictates the amount of noise that must be added to the query output to achieve privacy.

¹An alternative notion of adjacency, known as *exchange adjacency*, defines two datasets as adjacent if they differ in the data of exactly one individual while maintaining the same dataset size. Conceptually, this can be viewed as the removal of one individual followed by the addition of another.

Definition 2.2.2 (Global sensitivity). The global sensitivity of a function $f : D \rightarrow \mathcal{R}$ is defined as the maximum difference in the output of f over all pairs of adjacent datasets $D \sim D' \in \mathcal{D}$, measured with respect to p -norm:

$$\Delta_p f = \max_{D \sim D'} \|f(D) - f(D')\|_p.$$

Formal definition of Differential Privacy The extent to which adding or removing an individual’s data affects the outcome depends on the specific function under consideration, whether it involves calculating sums, averages, or other measures. This sensitivity measure indicates how much the output of the target function must be adjusted to protect an individual’s privacy. DP accomplishes this by adding noise to the function’s output in an amount calibrated to the function’s sensitivity. To formalize this intuition, DP is defined in probabilistic terms, specifying how much the output distribution of a mechanism may change when an individual’s data is added or removed.

Definition 2.2.3 $((\epsilon, \delta)$ -Differential Privacy from Dwork et al. [26]). A randomized mechanism $\mathcal{M} : D \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ϵ, δ) -differentially private if, for any event $S \subseteq \mathcal{R}$ and any pair $D, D' \in \mathcal{D}$ of adjacent datasets:

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta,$$

where the probability is calculated over the randomness of $\mathcal{M}$.

A differentially private mechanism maps a dataset to a probability distribution over possible outputs, for example, by adding noise or making randomized choices. The released output corresponds to one random draw from this distribution. This inherent randomness is precisely what provides privacy. The privacy guarantee is quantified by two parameters, ϵ and δ , which together determine the level of indistinguishability between outputs from adjacent datasets.

Privacy loss ϵ The parameter ϵ , referred to as the *privacy loss*, bounds the multiplicative difference between the output distributions of adjacent datasets. Conceptually, ϵ can be viewed as a knob that controls the level of privacy. Smaller values of ϵ correspond to stronger privacy guarantees as they constrain how much the output probabilities can differ when a single individual’s data is modified. When $\epsilon = 0$ and $\delta = 0$, $\Pr[\mathcal{M}(D) \in S] = \Pr[\mathcal{M}(D') \in S]$ for all S , providing *perfect* privacy, since the output distribution is identical for all datasets. In this extreme case, no information about any individual can be

inferred, and the output simultaneously conveys no utility because it must remain constant regardless of the data. Conversely, as $\varepsilon \rightarrow \infty$, the privacy constraints become vacuous, and the mechanism may reveal the data without restriction.

In practice, achieving a balance between privacy and utility requires selecting an appropriate value of ε . Smaller values of ε (such as $\varepsilon \leq 1$) provide stronger privacy guarantees but may degrade data accuracy, whereas larger values (e.g., $\varepsilon \geq 5$) offer higher utility at the expense of weaker privacy protection. This trade-off is reflected in real-world deployments of differential privacy. Apple’s implementation of differential privacy in its data collection systems employs a per-user, per-day privacy budgets typically in the range of $\varepsilon = 2$ to $\varepsilon = 16$ [23]. The U.S. Census Bureau adopted an overall privacy-loss budget of $\varepsilon = 19.61$ for the 2020 Census redistricting data [82]. In contrast, Google’s Community Mobility Reports, published during the COVID-19 pandemic, used a considerably smaller per-user daily budget of $\varepsilon = 2.64$ [5].

Failure probability δ The second parameter, δ , represents a margin of error, defining the probability with which the privacy guarantee may not hold. It is typically chosen as a small value close to zero, often much smaller than $1/N$, where N denotes the size of the dataset. Selecting such a negligible δ ensures that the risk of disclosing sensitive information about any individual remains exceedingly low. Allowing a small positive δ is often necessary to achieve reasonable utility, as enforcing $\delta = 0$ (known as *pure differential privacy*) generally requires injecting excessive noise, rendering the results less informative. When $\delta > 0$, the mechanism satisfies (ε, δ) -DP, which is known as *approximate differential privacy*.

The Gaussian mechanism The canonical mechanism for (ε, δ) -Differential Privacy is the Gaussian mechanism [27]. It operates by adding noise calibrated to the l_2 sensitivity of a query, denoted by $\Delta_2 f$. For a numerical query $f : \mathcal{D} \rightarrow \mathcal{R}$, the mechanism is formally defined as $\mathcal{M}_{\text{Gauss}}(D; f, \varepsilon) = f(D) + z$, where $z \in \mathcal{R}$ is a vector of i.i.d.² samples drawn from $\mathcal{N}(0, \sigma^2 I)$ ³. While this standard calibration restricts ε to the interval $(0, 1]$, generalized variations such as the Analytical Gaussian Mechanism provide precise (ε, δ) -DP guarantees for any $\varepsilon > 0$ [8].

²i.i.d. stands for independent and identically distributed.

³To guarantee (ε, δ) -DP for $\varepsilon \in (0, 1]$, the standard deviation σ must satisfy $\sigma \geq \sqrt{2 \ln(1.25/\delta)} (\Delta_2 f / \varepsilon)$.

An important feature of differential privacy is its composition property: when several differentially private algorithms are applied to the same data, the overall privacy guarantee degrades in a controlled way. This property allows us to quantify the cumulative privacy loss over multiple computations. The compositional nature of DP makes it well-suited for iterative algorithms, forming the foundation of Differentially Private Stochastic Gradient Descent (DP-SGD), where composition governs how privacy loss accumulates over the course of model training.

2.2.2 Differentially Private Stochastic Gradient Descent

Training modern deep learning models typically involves iterative optimization algorithms such as Stochastic Gradient Descent (SGD). However, standard SGD does not preserve privacy, as a model’s parameters θ can be heavily influenced by individual training examples and may even “memorize” them [90, 15]. Applying the formal definition of DP to the entire training function is challenging, since its global sensitivity is often complex to analyze and can even be unbounded.

A more practical approach is to privatize each step of the optimization process and then analyze the cumulative privacy loss using the composition properties of differential privacy. Differentially Private Stochastic Gradient Descent (DP-SGD), proposed by Abadi et al. [1], achieves (ε, δ) -Differential Privacy by extending the standard SGD algorithm and modifying the gradient computation at each training iteration, as outlined in Algorithm 1.

Algorithm 1 DP-SGD algorithm from Abadi et al. [1]

Input: Dataset $\mathcal{D}$, loss function ℓ , learning rate η , batch size $\mathcal{B}$, noise scale σ , clipping threshold C , initial model θ_0

for $t \in [T]$ **do**

 Uniformly draw mini-batch $\mathcal{B}_t$ from $\mathcal{D}$

 For each $z_i \in \mathcal{B}_t$, $g_i^{\text{clip}}(\theta_t) = \text{clip}(\nabla \ell(\theta; z_i), C)$

$g^{\text{DP}}(\theta_t) = \frac{1}{|\mathcal{B}_t|} \left(\left(\sum_i g_i^{\text{clip}}(\theta_t) \right) + \mathcal{N}(0, \sigma^2 C^2 I_d) \right)$

$\theta_{t+1} = \theta_t - \eta g^{\text{DP}}(\theta_t)$

Return: θ_T and the overall privacy cost (ε, δ)

The DP-SGD algorithm ensures privacy by introducing two key modifications to the standard gradient computation. The first modification addresses the issue of unbounded

sensitivity: in standard SGD, an individual data point z may yield an arbitrarily large gradient $\nabla\ell(\theta_t; z)$, resulting in unbounded sensitivity of the aggregated gradient. To mitigate this, DP-SGD applies per-sample gradient clipping, wherein the 2-norm of each individual gradient is restricted to a fixed threshold C . This operation ensures that the contribution of any single example to the sum $\sum_i g_i^{\text{clip}}(\theta_t)$ is limited to C . Consequently, this clipping step establishes a finite sensitivity bound.

The second modification introduces calibrated noise into the mechanism designed to ensure differential privacy. Specifically, DP-SGD employs the Gaussian mechanism [27], which adds Gaussian noise $\mathcal{N}(0, \sigma^2 C^2 I_d)$ to the sum of clipped gradients. This noise is drawn from a distribution whose scale is determined by the sensitivity (or clipping threshold) C and a noise multiplier σ . The noisy sum is then normalized by the batch size $|\mathcal{B}_t|$ to yield the final private gradient estimate $g^{\text{DP}}(\theta_t)$, which is then used to update the model. A larger noise multiplier σ provides stronger privacy guarantees (i.e., a smaller ε), but it also introduces more randomness into the gradient updates. This added randomness can reduce the accuracy of the optimization and, consequently, the model’s overall utility.

Privacy accounting While each iteration of DP-SGD satisfies a certain level of privacy on its own, model training typically involves thousands of such iterations, each consuming a portion of the overall privacy loss. To rigorously track how privacy degrades across these steps, *privacy accounting* techniques are used. A standard and powerful tool for this is Rényi Differential Privacy (RDP), which extends the standard DP definition by expressing privacy loss in terms of Rényi divergence between the probability distributions of outputs on adjacent datasets [62]. In the context of DP-SGD, the privacy accountant first computes the RDP cost for a single training step, which depends on the noise multiplier σ and the batch sampling probability $|\mathcal{B}_t|/|\mathcal{D}|$. These per-step RDP values are then accumulated over all training iterations and subsequently converted into an overall (ε, δ) guarantee using standard transformations, resulting in the final reported privacy loss of the model. In practice, DP-SGD implementations use RDP-based privacy accountants to obtain tight bounds on the total privacy budget, allowing practitioners to balance privacy and model performance more effectively.

2.3 Membership Inference Attacks

While Differential Privacy (see section 2.2) provides a formal, provable guarantee of privacy, it is also essential to empirically measure the potential privacy leakage of machine learning

models. A primary tool for this empirical evaluation is the Membership Inference Attacks (MIAs). The goal of an MIA is to determine, given a data point z and black-box or white-box access⁴ to a trained model f , whether z was included in the model’s training dataset D . This inference task is integral to privacy because it directly tests whether a model “remembers” individual training examples. For instance, if an adversary can confirm that a patient’s specific record was used to train a model designed to predict a particular disease, it could reveal sensitive information about the patient’s health status.

The vulnerability of models to MIAs arises from their tendency, especially high-capacity deep neural networks, to “memorize” or overfit to their training data [90]. As a result of this memorization, the model often exhibits different behavior on inputs it was trained on (*members*) compared to unseen inputs (*non-members*). For instance, the model may assign a lower loss or higher confidence score to members than to similar non-member points.

The seminal work by Shokri et al. [75] formalized this threat and introduced a general attack framework based on *shadow* models. In this approach, the adversary trains a set of shadow models $f'_i(\theta)$, where each model is trained on a dataset D_i sampled from the same underlying distribution as the target model’s data. By observing the outputs (e.g., confidence scores) of the shadow models on their respective members and non-members, the adversary creates a new dataset to train a binary classifier. The classifier then learns to distinguish the output patterns associated with member and non-member examples. Finally, the adversary can infer the membership of the target point z by feeding the target model’s output into the trained classifier.

The method we use in our analysis, the Likelihood-Ratio Attack (LiRA) [16] builds upon this approach to membership inference. While it also relies on shadow models to simulate the adversary’s knowledge, it differs from the earlier framework by replacing the general attack classifier with a more direct and statistically grounded membership test.

2.3.1 Likelihood Ratio Attack

The Likelihood-Ratio Attack (LiRA), introduced by Carlini et al. [16], is a widely used benchmark that reframes membership inference as a direct statistical hypothesis test. The goal is to determine which of two hypotheses is more likely to be correct for a given data point z :

⁴In a black-box attack, the adversary can only query the model and observe its outputs (for example, predicted labels or confidence scores) without having access to the model’s internal parameters or architecture. In a white-box attack, the adversary has full access to the model’s internals, including parameters, gradients, and architecture, thus providing stronger adversarial access.

- H_{in} : The data point z was *in* the training set D .
- H_{out} : The data point z was *not in* the training set D .

LiRA operates by comparing the distribution of a scalar score $s(f, z)$ extracted from the model’s output. The central idea of LiRA is that s will follow two distinct probability distributions depending on whether z is a member or not. Typical choices for this score include the per-example loss, $\ell(f(x), y)$, or the model’s predicted probability (confidence) of the true label y , where $z = (x, y)$.

The underlying intuition is that this score s will follow two different probability distributions depending on the membership status of the data point z . We are interested in estimating:

- $P(s \mid H_{\text{in}})$ (the IN distribution): The distribution of scores for point z given that z was used to train the model.
- $P(s \mid H_{\text{out}})$ (the OUT distribution): The distribution of scores for point z given that z was not used to train the model.

Intuitively, since the model may memorize its training examples, we expect the IN distribution to be concentrated around lower losses or higher confidence values than the OUT distribution. Having been trained on z , the model produces a score s that is likely to be an outlier relative to the scores it would generate for non-member points.

As the adversary lacks access to these true distributions, LiRA approximates them by training several *shadow models* on datasets independently sampled from the same population as the target model’s data. To build a statistical test specific to the point z , LiRA trains two distinct ensembles of shadow models

$$\begin{aligned}\mathcal{F}_{\text{in}}(z) &= \left\{ f_{\text{in}}^{(1)}, \dots, f_{\text{in}}^{(m)} \right\} && \text{each trained on a dataset that } \textit{includes} \ z, \\ \mathcal{F}_{\text{out}}(z) &= \left\{ f_{\text{out}}^{(1)}, \dots, f_{\text{out}}^{(m)} \right\} && \text{each trained on a dataset that } \textit{excludes} \ z.\end{aligned}$$

The adversary then computes the score for z using each of these shadow models, giving two sets of empirical scores: $\{s(f_{\text{in}}^{(j)}, z)\}_{j=1}^m$ and $\{s(f_{\text{out}}^{(j)}, z)\}_{j=1}^m$. These samples serve as empirical estimates of the score distributions IN and OUT *for the specific* z . LiRA models these two distributions as Gaussian, $\mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}}^2)$ and $\mathcal{N}(\mu_{\text{out}}, \sigma_{\text{out}}^2)$, whose parameters are estimated from the sample means and variances of these two sets of scores.

Finally, to perform the attack, the adversary obtains the score $s^* = s(f^*, z)$ from the target model f^* . They then compare its likelihood under the two estimated distributions using the log-likelihood ratio:

$$\Lambda(z) = \log \frac{p(s^* | H_{\text{in}})}{p(s^* | H_{\text{out}})} = \log \frac{\mathcal{N}(s^*; \mu_{\text{in}}(z), \sigma_{\text{in}}^2(z))}{\mathcal{N}(s^*; \mu_{\text{out}}(z), \sigma_{\text{out}}^2(z))}.$$

Thresholding $\Lambda(z)$ yields the attack decision. A large positive value implies that s^* is significantly more likely under the IN distribution, leading the adversary to conclude that z was a member. Intuitively, when the target model’s score s^* falls within a region of higher likelihood under the estimated IN distribution relative to the OUT distribution, the attack infers that the point is a member.

The effectiveness of LiRA is typically evaluated using two primary metrics: the Area Under the Receiver Operating Characteristic Curve (AUC) and the True Positive Rate (TPR) at a low False Positive Rate (FPR). The AUC provides a threshold-independent measure of the attack’s overall discriminative power. It can be interpreted as the probability that a randomly selected member receives a higher attack score $\Lambda(z)$ than a randomly selected non-member. An AUC of 0.5 is equivalent to random guessing, while scores at the extremes (1.0 for perfectly correct classification and 0.0 for perfectly inverted classification) both signify a perfect attack. However, in the context of privacy, TPR @ low FPR (for instance, TPR @ 0.1% FPR) is often the more critical metric. This metric reflects a more realistic and potentially more harmful attack scenario, in which the adversary aims for high precision and seeks to be very certain that any point identified as a “member” is indeed a member. By fixing the FPR at a low value, this metric evaluates the attack’s success rate (TPR) under a high-confidence operating point. A high TPR at a low FPR indicates a significant privacy vulnerability, as it demonstrates that the adversary can correctly identify a substantial proportion of training examples while making very few false identifications.

2.3.2 Connections to Differential Privacy

As outlined in the previous section, an algorithm that satisfies (ϵ, δ) -differential privacy provides a theoretical guarantee that the inclusion of any single example z can only minimally change the model’s output distribution. A direct consequence is that, for any membership test (including LiRA), the adversary’s advantage is bounded in terms of (ϵ, δ) . By definition, DP constrains the likelihood ratio between a model trained with z and one trained without it. This means that for a DP-trained model, the true IN distribution $P(s | H_{\text{in}})$ and OUT distribution $P(s | H_{\text{out}})$ are forced to be statistically similar. As the

DP guarantee becomes stronger, the two distributions become increasingly similar, leading to substantial overlap. This, in turn, shrinks the expected log-likelihood ratio $\Lambda(z)$ toward zero, making it impossible for any adversary to confidently distinguish members from non-members.

Chapter 3

Theoretical Analysis

This chapter develops the theoretical framework for analyzing the efficacy of collective action under the constraints of Differential Privacy (DP). Building upon the work of Hardt et al. [40], our analysis moves beyond the non-private setting to address the practical case where firms must deploy models trained with learning algorithms that rigorously protect user privacy. The primary objective is to formally characterize the fundamental trade-offs between the level of privacy guaranteed to users and the potential success of the collective.

The central contribution of this chapter is the formulation and proof of the main theorem that quantifies the success of the collective action. This theorem provides a precise mathematical relationship between the collective’s success and its key structural and privacy-related parameters. Specifically, we show how the collective’s success is a function of the collective’s size and the parameters of the DP mechanism—the noise multiplier σ and clipping threshold C . This result offers a clear interpretation of the “cost of privacy” by quantifying how stringent privacy requirements may limit the collective’s success, and how increasing the collective’s size can, in turn, compensate for the statistical noise introduced by DP.

This chapter is organized as follows. We begin in section 3.1 by formally defining the problem setup, outlining the theoretical assumptions underlying our analysis, and describing the properties of gradient updates under DP. Following this, in Section 3.2, we present our main theorem together with a rigorous proof and a detailed discussion of its interpretation.

3.1 Problem Setup

We assume that the firm deploys a private learning algorithm $\mathcal{A}$ with the objective of preserving user data privacy. Given a data distribution $\mathcal{P}$ (modeled as a mixture distribution as defined in Equation (2.1)), the model chosen by the firm is represented by $f = \mathcal{A}(\mathcal{P})$. Since we are concerned with gradient-based parametric risk minimization, we focus on learning algorithms that rely on (stochastic) gradient descent to optimize the model parameters—both with and without privacy constraints.

More specifically, we consider Differentially Private Stochastic Gradient Descent (DP-SGD), a widely used algorithm for training models under (ε, δ) -differential privacy constraints. DP-SGD proceeds by (i) clipping per-example gradients to bound their sensitivity and (ii) adding calibrated Gaussian noise to the aggregated clipped-gradients before performing the parameter update. We now formalize the dynamics of a differentially private learning process under algorithmic collective action. We consider a realistic learning scenario that does not assume convexity of the objective function, where the learner observes a batch of $\mathcal{B}$ sampled i.i.d. from the distribution $\mathcal{P}$ and aims to minimize the empirical loss $(\sum_{z \in \mathcal{B}} \ell(\theta; z)) / |\mathcal{B}|$ using gradient-based optimization methods. Here, $\ell(\theta; z)$ denotes the loss evaluated using model parameters θ on an example z . We define the average gradient over batch $\mathcal{B}$ as $\bar{g}_{\mathcal{B}}(\theta) = (\sum_{z \in \mathcal{B}} \nabla \ell(\theta; z)) / |\mathcal{B}|$. This empirical formulation differs from previous research [40], which analyzed expected gradients rather than the average over a finite batch.

At each time step t , the learner observes samples $\mathcal{B}_t$ from the current data distribution $\mathcal{P}_t$, allowing the collective to adaptively interact with the learner by choosing $\mathcal{B}_t^*$ from the distribution $\mathcal{P}_t^*$. This setup models the best-case scenario for the collective, enabling us to analyze the potential effectiveness of its strategy under ideal conditions. Given a clipping threshold C and a noise scale σ , the model parameters are updated by taking a gradient step computed according to the DP-SGD:

$$\theta_{t+1} = \theta_t - \eta \frac{1}{|\mathcal{B}_t|} \left(\left(\sum_{z \in \mathcal{B}_t} \text{clip}(\nabla \ell(\theta_t; z), C) \right) + \mathcal{N}(0, \sigma^2 C^2 I_d) \right) \quad (3.1)$$

where η is the learning rate, $\text{clip}(\mathbf{u}, C) = \mathbf{u} \min(1, C/\|\mathbf{u}\|)$ denotes the gradient clipping operation, which scales the gradient to have a norm of at most C , and I_d denotes the $d \times d$ identity matrix. For notational convenience, we first define the average clipped gradient as $\bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) := (\sum_{z \in \mathcal{B}_t} \text{clip}(\nabla \ell(\theta_t; z), C)) / |\mathcal{B}_t|$, which gives:

$$\theta_{t+1} = \theta_t - \eta \left(\bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_t|^2} I_d\right) \right) \quad (3.2)$$

Next, we define the final noisy (differentially private) gradient as $\bar{g}_{\mathcal{B}_t}^{\text{DP}}(\theta_t) := \bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) + \mathcal{N}(0, (\sigma^2 C^2 / |\mathcal{B}_t|^2) I_d)$, resulting in:

$$\theta_{t+1} = \theta_t - \eta \bar{g}_{\mathcal{B}_t}^{\text{DP}}(\theta_t), \quad (3.3)$$

With this formulation in place, we now proceed to the theoretical analysis of how differentially private learning dynamics influence algorithmic collective action.

3.2 Main Theoretical Results

This section presents our main theoretical result, which quantifies the effectiveness of the collective’s gradient-redirecting strategy (as defined in Definition 2.1.3) under DP-SGD. This result links the collective’s size α , the privacy parameters (the noise multiplier σ and the clipping threshold C), and the optimization dynamics through a lower bound on the collective’s success. The theorem below formalizes the bound.

Theorem 3.2.1. *Assume that the collective can implement the gradient-redirecting strategy from Definition 2.1.3 at all $\lambda\theta_0 + (1 - \lambda)\theta^*$, where $\lambda \in [0, 1]$ and $\theta_0, \theta^* \in \mathbb{R}^d$. Then, for a given clipping threshold C and noise multiplier σ , and letting $|\mathcal{B}|_{\min}$ denote the minimum batch size observed over T steps, there exists $B_{\alpha, C} > 0$, such that the success of the gradient-control strategy after T steps is lower-bounded with probability greater than $1 - \delta$ by,*

$$S_T(\alpha, \sigma, C) \geq -(1 - \eta B_{\alpha, C})^T \|\theta_0 - \theta^*\| - \frac{\sigma C}{|\mathcal{B}|_{\min}^2} \Gamma_T \Delta_{d, \delta}. \quad (3.4)$$

Here, $B_{\alpha, C}$ is directly proportional to the collective’s size α and clipping threshold C , the function Γ_T denotes convergence-dependent noise-accumulation factor, and $\Delta_{d, \delta}$ is a high-probability upper bound on the Euclidean norm of a d -dimensional standard Gaussian vector. See subsection 3.2.1 for the proof, Equation (3.10) for the closed-form expressions, and subsection 3.2.2 for the interpretation.

3.2.1 Proof of Theorem 3.2.1

Before proving Theorem 3.2.1, we present a lemma and several technical details that will be used in the proof.

Tail bound on scaled Gaussian norm In order to analyze the effect of the Gaussian noise introduced by the DP-SGD updates, we require a high-probability bound¹ on the norm of a multidimensional Gaussian vector. The following lemma provides such a bound, which will be used in the proof of Theorem 3.2.1 to control the stochasticity introduced by the Gaussian noise. The proof of this lemma is provided in appendix A.2.

Lemma 3.2.2. *Let $Y_1, \dots, Y_d \sim \mathcal{N}(0, \sigma^2)$ be independent Gaussian random variables, and define the scaled chi-squared distribution as,*

$$Z = \|Y\|_2 = \sqrt{\sum_{i=1}^d Y_i^2}$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$Z \leq \sigma \left(\sqrt{d} + \sqrt{2 \log \left(\frac{1}{\delta} \right)} \right) \quad (3.5)$$

Decomposition of gradients At each iteration t , the model parameters are denoted by $\theta_t \in \Theta$, and the learner observes samples $\mathcal{B}_t$ from the data mixture $\mathcal{P}_t = \alpha \mathcal{P}'_t + (1 - \alpha) \mathcal{P}_0$, where $\mathcal{P}'_t$ is the collective's data distribution. The corresponding average clipped gradient over $\mathcal{B}_t$,

$$\bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) = \frac{1}{|\mathcal{B}_t|} \sum_{z \in \mathcal{B}_t} \text{clip}(\nabla \ell(\theta_t; z), C),$$

which can then be decomposed as

$$\begin{aligned} \bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) &= \frac{1}{|\mathcal{B}_t|} \left(\sum_{z \in \mathcal{B}'_t} \text{clip}(\nabla \ell(\theta_t; z), C) + \sum_{z \in \mathcal{B}_0} \text{clip}(\nabla \ell(\theta_t; z), C) \right) \\ &= \frac{|\mathcal{B}'_t|}{|\mathcal{B}_t|} \left(\frac{1}{|\mathcal{B}'_t|} \sum_{z \in \mathcal{B}'_t} \text{clip}(\nabla \ell(\theta_t; z), C) \right) + \frac{|\mathcal{B}_0|}{|\mathcal{B}_t|} \left(\frac{1}{|\mathcal{B}_0|} \sum_{z \in \mathcal{B}_0} \text{clip}(\nabla \ell(\theta_t; z), C) \right) \\ &= \alpha \bar{g}_{\mathcal{B}'_t}^{\text{clip}}(\theta_t) + (1 - \alpha) \bar{g}_{\mathcal{B}_0}^{\text{clip}}(\theta_t). \end{aligned}$$

¹A high-probability bound ensures that a random quantity deviates from its typical value (often its expectation) by at most a specified amount with probability at least $1 - \delta$, for a small $\delta > 0$ [6].

Adding the per-step DP noise $\mathcal{N}(0, (\sigma^2 C^2 / |\mathcal{B}_t|^2) I_d)$ to both sides makes the left-hand side $\bar{g}_{\mathcal{B}_t}^{\text{DP}}(\theta_t)$ (refer Equation (3.3)), yielding

$$\bar{g}_{\mathcal{B}_t}^{\text{DP}}(\theta_t) = \alpha \bar{g}_{\mathcal{B}_t'}^{\text{clip}}(\theta_t) + (1 - \alpha) \bar{g}_{\mathcal{B}_0}^{\text{clip}}(\theta_t) + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_t|^2} I_d\right)$$

where C is the clipping threshold, σ is the noise multiplier, and I_d denotes the $d \times d$ identity matrix.

Gradient redirection strategy under differential privacy Assuming that the collective implements the gradient-redirecting strategy (Definition 2.1.3), we can express the average clipped gradient under the collective's distribution as²

$$\bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) = \alpha \xi^c(\theta_t) (\theta_t - \theta^*) \quad (3.6)$$

$$\alpha \bar{g}_{\mathcal{B}_t'}^{\text{clip}}(\theta_t) + (1 - \alpha) \bar{g}_{\mathcal{B}_0}^{\text{clip}}(\theta_t) = \alpha \xi^c(\theta_t) (\theta_t - \theta^*) \quad (3.7)$$

where $\xi^c(\theta_t)$ is a scalar, similar to ξ defined in collective's strategy of subsection 2.1.4 but when considering clipped gradients at time step t . Rearranging the above expression gives the clipped variant of Definition 2.1.3:

$$\xi^c(\theta_t) (\theta_t - \theta^*) = \bar{g}_{\mathcal{B}_t'}^{\text{clip}}(\theta_t) + \frac{1 - \alpha}{\alpha} \bar{g}_{\mathcal{B}_0}^{\text{clip}}(\theta_t). \quad (3.8)$$

We can further get the expression of $\xi^c(\theta_t)$ by taking norms on both sides and isolating $\xi^c(\theta_t)$, which yields

$$\xi^c(\theta_t) = \frac{\left\| \bar{g}_{\mathcal{B}_t'}^{\text{clip}}(\theta_t) + \frac{1 - \alpha}{\alpha} \bar{g}_{\mathcal{B}_0}^{\text{clip}}(\theta_t) \right\|}{\|\theta_t - \theta^*\|}. \quad (3.9)$$

Proof. We begin with DP-SGD update (refer Equation (3.2))

$$\theta_{t+1} = \theta_t - \eta \left(\bar{g}_{\mathcal{B}_t}^{\text{clip}}(\theta_t) + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_t|^2} I_d\right) \right),$$

²This implicitly assumes that the collective has some knowledge of the firm's clipping operation when selecting the gradient-redirecting distribution.

where $\eta > 0$ is the learning rate. Starting with the final error after T iterations $\|\theta_T - \theta^*\|$, we unroll one step of the recurrence using the above DP-SGD update equation:

$$\|\theta_T - \theta^*\| \leq \left\| \theta_{T-1} - \eta \left(\bar{g}_{\mathcal{B}_T}^{\text{clip}}(\theta_{T-1}) + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_T|^2} I_d\right) \right) - \theta^* \right\|$$

Substituting the expected clipped gradients under the gradient-redirection strategy from Equation (3.6), we obtain

$$\leq \left\| \theta_{T-1} - \eta \left(\alpha \xi^c(\theta_{T-1}) (\theta_{T-1} - \theta^*) + \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_T|^2} I_d\right) \right) - \theta^* \right\|$$

Factoring out $(\theta_{T-1} - \theta^*)$ gives

$$= \left\| (1 - \eta \alpha \xi^c(\theta_{T-1})) (\theta_{T-1} - \theta^*) - \eta \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_T|^2} I_d\right) \right\|$$

We now relax the bound by replacing $\xi^c(\theta_{T-1})$ with its minimum value $\xi_{\min}^c = \min_{\lambda \in [0,1]} \xi^c(\lambda \theta_0 + (1-\lambda)\theta^*)$, leading to

$$\leq \left\| (1 - \eta \alpha \xi_{\min}^c) (\theta_{T-1} - \theta^*) - \eta \mathcal{N}\left(0, \frac{\sigma^2 C^2}{|\mathcal{B}_T|^2} I_d\right) \right\|$$

To make the recursive structure explicit, we expand the update one additional step to express θ_{T-1} in terms of θ_{T-2} :

$$= \left\| (1 - \eta \alpha \xi_{\min}^c)^2 (\theta_{T-2} - \theta^*) - \eta \mathcal{N}(0, \sigma^2 C^2 I_d) \left(\frac{1}{|\mathcal{B}_T|} + \frac{(1 - \eta \alpha \xi_{\min}^c)}{|\mathcal{B}_{T-1}|} \right) \right\|$$

Continuing this process, unrolling the linear recursion from θ_{T-1} down to θ_0 yields

$$= \left\| (1 - \eta \alpha \xi_{\min}^c)^T (\theta_0 - \theta^*) - \eta \mathcal{N}(0, \sigma^2 C^2 I_d) \sum_{k=0}^{T-1} \frac{(1 - \eta \alpha \xi_{\min}^c)^k}{|\mathcal{B}_{T-k}|} \right\|$$

Using the fact that adding or subtracting a zero-mean Gaussian random variable does not change the distribution, we obtain (where $\stackrel{d}{=}$ denotes equality in distribution):

$$\stackrel{d}{=} \left\| (1 - \eta \alpha \xi_{\min}^c)^T (\theta_0 - \theta^*) + \eta \mathcal{N}(0, \sigma^2 C^2 I_d) \sum_{k=0}^{T-1} \frac{(1 - \eta \alpha \xi_{\min}^c)^k}{|\mathcal{B}_{T-k}|} \right\|$$

By the triangle inequality, the norm of a sum is bounded above by the sum of the norms, and pulling constants outside the norm then gives³

$$\leq (1 - \eta \alpha \xi_{\min}^c)^T \|\theta_0 - \theta^*\| + \eta \sum_{k=0}^{T-1} \frac{(1 - \eta \alpha \xi_{\min}^c)^k}{|\mathcal{B}_{T-k}|} \|\mathcal{N}(0, \sigma^2 C^2 I_d)\|$$

Applying Lemma 3.2.2, the right-hand side can be further upper-bounded with probability at least $1 - \delta$, for θ with d degrees of freedom,

$$\leq (1 - \eta \alpha \xi_{\min}^c)^T \|\theta_0 - \theta^*\| + \eta \sum_{k=0}^{T-1} \frac{(1 - \eta \alpha \xi_{\min}^c)^k}{|\mathcal{B}_{T-k}|} \sigma C \left(\sqrt{d} + \sqrt{2 \log 1/\delta} \right)$$

Bounding the sum by assuming the minimum batch size $|\mathcal{B}|_{\min}$ for all T steps:

$$\leq (1 - \eta \alpha \xi_{\min}^c)^T \|\theta_0 - \theta^*\| + \eta \left(\sum_{k=0}^{T-1} (1 - \eta \alpha \xi_{\min}^c)^k \right) \frac{\sigma C}{|\mathcal{B}|_{\min}^2} \left(\sqrt{d} + \sqrt{2 \log 1/\delta} \right)$$

The terms inside the summation form a geometric progression—replacing the summation with its closed-form expression gives⁴

$$= (1 - \eta \alpha \xi_{\min}^c)^T \|\theta_0 - \theta^*\| + \eta \frac{1 - (1 - \eta \alpha \xi_{\min}^c)^T}{\alpha \xi_{\min}^c} \frac{\sigma C}{|\mathcal{B}|_{\min}^2} \left(\sqrt{d} + \sqrt{2 \log 1/\delta} \right)$$

Let $B_{\alpha,C} = \alpha \xi_{\min}^c$, $\Gamma_T = \eta \left(1 - (1 - \eta B_{\alpha,C})^T \right) / B_{\alpha,C}$ and $\Delta_{d,\delta} = \left(\sqrt{d} + \sqrt{2 \log 1/\delta} \right)$; then we have

$$\|\theta_T - \theta^*\| \leq (1 - \eta B_{\alpha,C})^T \|\theta_0 - \theta^*\| + \frac{\sigma C}{|\mathcal{B}|_{\min}^2} \Gamma_T \Delta_{d,\delta} \quad (3.10)$$

Multiplying both sides by -1 transforms the left-hand side into the collective's success (as defined in Equation (2.6)). This converts the upper bound on the parameter norm difference into a lower bound on collective's success, resulting in the final bound:

$$S_T(\alpha, \sigma, C) \geq - (1 - \eta B_{\alpha,C})^T \|\theta_0 - \theta^*\| - \frac{\sigma C}{|\mathcal{B}|_{\min}^2} \Gamma_T \Delta_{d,\delta}.$$

□

³For a geometric series $\sum_{k=0}^{T-1} r^k = \frac{1-r^T}{1-r}$, for $r \neq 1$.

⁴For any vectors a and b , $\|a + b\| \leq \|a\| + \|b\|$. More generally, for $\{a_i\}_{i=1}^n$, $\|\sum_{i=1}^n a_i\| \leq \sum_{i=1}^n \|a_i\|$.

3.2.2 Interpretation

The most intuitive factor that constrains the success of the collective when the firm uses DP-SGD is the algorithm’s inherent ability to limit the influence of any individual data point on the model’s output. Gradient clipping reduces the collective’s ability to align the gradients in their desired direction, while the injected noise further deflects this directional push. As a result, the signal that the collective is trying to correlate with the target label is also attenuated. This is equivalent to the collective introducing a noisy signal, which in turn increases the effort required for the collective to influence the outcome.

Theorem 3.2.1 provides a lower bound on the collective’s success, S_T , breaking its structure into two components: a *contraction term* and a *privacy penalty*.

$$S_T(\alpha, \sigma, C) \geq - \underbrace{(1 - \eta B_{\alpha, C})^T \|\theta_0 - \theta^*\|}_{\text{Contraction term}} - \underbrace{\frac{\sigma C}{|\mathcal{B}|_{\min}^2} \Gamma_T \Delta_{d, \delta}}_{\text{Privacy penalty}}.$$

The contraction term describes how quickly the collective’s gradient-redirecting strategy can pull the model parameters θ_t from their initial state θ_0 towards the collective’s target θ^* . As T increases (or as $B_{\alpha, C}$ or η increase), this factor decays toward zero. Because success is defined as the negative of the contraction term, larger T , $B_{\alpha, C}$, or η correspond to the greater success of the collective. On the other hand, the privacy penalty quantifies how the random noise added to ensure differential privacy pushes the model away from any fixed target, including the collective’s θ^* . Increasing the privacy penalty reduces the value on the right-hand side (due to the negative sign), which in turn decreases the collective’s success. We now analyze how α , σ , and C influence the collective’s success.

Impact of collective’s size α The collective’s size α appears in $B_{\alpha, C} = \alpha \xi_{\min}^c$ (see Equation (3.9)). For the first term of the bound, a larger collective increases $B_{\alpha, C}$, which in turn speeds up the geometric contraction rate $(1 - \eta B_{\alpha, C})^T$. Intuitively, a larger collective can pull the model more effectively toward the target. For the second term, $B_{\alpha, C}$ appears in the denominator of the $\Gamma_T = \eta(1 - (1 - \eta B_{\alpha, C})^T)/B_{\alpha, C}$, a function that generally decreases when $B_{\alpha, C}$ grows. As a result, an increase in $B_{\alpha, C}$ reduces the magnitude of the privacy penalty, thereby increasing the lower bound. Therefore, a larger collective’s size α unambiguously increases the lower bound on success by both strengthening the desired signal and reducing the relative impact of the noise.

Impact of noise multiplier σ The privacy penalty (the second term) is linear in σ , so increasing σ —corresponding to stronger privacy guarantees and a lower privacy budget ε —

directly reduces the collective’s success. This reflects the fundamental trade-off: the same mechanism that protects user privacy by adding noise also comes at the cost of reduction in performance of the collective.

Impact of clipping threshold C The clipping threshold C , set by the firm, influences the bound in two opposing ways, leading to a nuanced overall effect.

- Positive effect via $B_{\alpha,C}$: Larger C increases the clipped-gradient magnitude and therefore $B_{\alpha,C}$ (as $B_{\alpha,C}$ is proportional to the norm of the expected clipped gradient). As established in the analysis of α , a larger $B_{\alpha,C}$ strengthens the signal (improving the contraction rate) and reduces the relative noise impact (decreasing the Γ_T), thus improving the bound.
- Negative effect via the noise multiplier: A larger C increases the per-step injected noise (which has a standard deviation σC). This, in turn, harms the collective’s success by further lowering the second term.

As a result, the overall impact of the clipping threshold C on the success of the collective is determined by the interplay between these competing influences.

3.2.3 Recovery of the Non-private bound

Theorem 3.2.1 contains the non-private result as a single, natural corollary: by simultaneously removing both privacy mechanisms—taking the clipping threshold to infinity ($C \rightarrow \infty$) and setting the noise multiplier to zero ($\sigma = 0$)—the DP-specific terms vanish and the bound collapses to the previously established non-private bound. More precisely, as $C \rightarrow \infty$ the clipped gradient constant $B_{\alpha,C}$ converges to the non-private magnitude $C(\alpha)$ used in Theorem 2.1.4, while $\sigma = 0$ eliminates the second term; combining these limits yields exactly the contraction bound stated in Theorem 2.1.4. The recovery is outlined in appendix A.3.

Chapter 4

Empirical Analysis

The preceding chapter established the theoretical foundations necessary to understand the relationship between Algorithmic Collective Action and Differential Privacy. While foundational, the theoretical analysis largely abstracts from many practical aspects of the learning process. To bridge this gap between theory and practice, this chapter presents a comprehensive empirical investigation. Our primary objective is to validate whether the theoretically predicted trade-offs and relationships hold in the complex, high-dimensional, and non-convex optimization landscapes of modern deep neural networks.

We use Differentially Private Stochastic Gradient Descent (DP-SGD) to protect the privacy of the dataset and to examine how it affects the collective’s ability to influence the behavior of deep learning models trained on standard benchmark datasets. Specifically, we seek to quantitatively address the following research questions:

- How does the *critical mass*—the minimum collective’s size required to achieve a desired level of success—quantitatively shift as privacy guarantees ε are tightened?
- What is the practical impact of the firm’s choice of hyperparameters, particularly the gradient clipping threshold C , on the collective’s efficacy?
- In a reciprocal analysis, what effect does algorithmic collective action have on the model’s *empirical* privacy, as measured by its vulnerability to privacy attacks?

To address these questions, this chapter is structured as follows. Section 4.1 details the complete experimental setup. This includes information about the datasets, model architectures, training procedures, and the collective’s strategy. Section 4.2 presents the

primary results of our investigation. Through extensive experiments, we visualize the trade-off between the collective’s success and the privacy loss ϵ across different clipping thresholds. Finally, section 4.3 shifts the perspective to investigate the reciprocal effects of ACA on privacy, where we evaluate model’s vulnerability to Membership Inference Attacks in the presence of a collective.

4.1 Experimental Setup

This section outlines a comprehensive experimental framework designed to empirically investigate the effectiveness of Algorithmic Collective Action (ACA), under the assumption that the underlying learning algorithm enforces differential privacy. The primary objective of these experiments is to quantify how the introduction of Differentially Private Stochastic Gradient Descent (DP-SGD), affects a collective’s ability to influence a model’s behavior. We aim to identify the *critical mass*—minimum size of the collective required to achieve the desired outcome—under varying privacy configurations.

4.1.1 Datasets and Tasks

To ensure that our findings are generalizable, we selected four datasets spanning different data modalities and task complexities: one for multi-label, multi-class text classification, two for multi-class image classification, and one for binary tabular classification. This allows us to examine the collective’s influence across standard image classification benchmarks as well as real-world textual and tabular data.

Resume We use a real-world corpus of approximately 30,000 anonymous resumes collected from the web and annotated with high-level IT skill labels (for example, *web developer*, *database administrator*), introduced by Jiechieu and Tsopze [46]. The task is formulated as a multi-label, 10-class text classification problem, where each resume may have multiple labels associated with it. This dataset contains noisy natural-language inputs and entangled concepts, which are ideal for probing collective’s influence on model behavior [40].

MNIST A foundational dataset in image classification, consisting of 60,000 training images and 10,000 testing images of handwritten digits (0-9) [53]. Each image is a 28×28 -pixel grayscale grid. For the sake of simplicity, the training set was balanced to contain

5,000 samples per class, totaling 50,000 images. The task is 10-class image classification, used as a proof of concept for our approach.

CIFAR-10 A more complex image classification benchmark containing 50,000 training and 10,000 testing 32×32 color images across 10 mutually exclusive classes [49]. The task is a 10-class image classification problem. The higher dimensionality and intra-class variation of CIFAR-10 provides a more challenging environment to test the collective’s impact.

Bank Marketing A tabular dataset containing 41,188 samples with 20 features covers client demographic and marketing campaign details [63]¹. The task is a binary classification: predicting whether a client will subscribe to a term deposit. The data were balanced using undersampling, resulting in a total of 7,888 data points. This dataset allows us to study the dynamics of collective action in a real-world financial marketing context.

4.1.2 Model Architectures and Training

We select different model architectures and base model characteristics depending on the datasets, and adjust the base architecture to accommodate DP-SGD. We also apply different training procedures across datasets to account for their varying characteristics and modalities.

Model Architectures

For MNIST and CIFAR-10, we modified the standard ResNet-18 architecture. A key modification involved replacing all Batch Normalization layers with Group Normalization. This change was necessary for DP-SGD, as Batch Normalization’s reliance on batch-level statistics complicates the computation of per-sample gradients, which are essential for enforcing differential privacy. We also modified the initial layers of the model to better accommodate the smaller input dimensions. Specifically, the first convolutional layer was replaced by a 3×3 convolution (stride 1, padding 1, no bias) while preserving the original number of output channels, and the initial max-pooling layer was substituted with a 1×1 max-pooling layer (stride 1, padding 0) to effectively remove its downsampling effect.

¹Dataset (version 1) fetched from `OpenML` using `sklearn.datasets.fetch_openml`

Following Hardt et al. [40], we use the DistilBERT architecture, a smaller, distilled variant of BERT with 6 transformer layers, for training on the Resume dataset [73]. DistilBERT retains approximately 97% of BERT’s performance while using 40% fewer parameters, making it a lightweight yet effective model for text classification. The Bank Marketing dataset is modeled using a simpler feedforward neural network. This network consists of a single hidden layer with 128 units and a ReLU activation function [3], followed by a final output layer for binary classification.

Training Procedures

Design Overview Broadly, our experimental design varies two key factors: the collective’s size and the privacy configuration. The collective’s size determines the amount of data controlled by the collective, while the privacy configuration specifies the parameters and hyperparameters that govern the level of differential privacy. We test a spectrum of privacy levels, ranging from no privacy to increasingly strict constraints. For each distinct privacy configuration, we train multiple models with varying collective’s sizes. This approach allows use to identify the critical mass required to achieve success under each privacy setting. The exact configuration are detailed in section 4.2.

Non-private Baseline As a baseline, models were trained with standard Stochastic Gradient Descent (SGD) without the application of any privacy mechanisms. This non-private baseline serves as a reference point for both model utility and the critical mass required for collective action. For the Resume dataset, following the setup of prior work [40], we instead use the AdamW optimizer [56] with its default parameters as implemented in the `transformers` library.

Differentially Private Training We use DP-SGD for privacy-preserving training on all datasets except the Resume dataset. For the Resume dataset, we use DP-AdamW, the differentially private variant of AdamW, to enforce privacy constraints while maintaining consistency with the non-private baseline. To validate our theoretical results, we performed an extensive set of experiments, training multiple models while systematically varying the key privacy parameters: the per-sample gradient clipping norm C and the noise multiplier σ . For precise privacy accounting throughout training, we use Rényi Differential Privacy (RDP), as detailed in subsection 2.2.2.



Figure 4.1: Binary masks illustrating the spatial structure of the two image transformation signals applied to a 28×28 grid. The left mask corresponds to the *patch* transformation, where the 2×2 white block in the top corner is a single pixel. The right mask corresponds to the *grid* transformation, where each white pixel follows a periodic pattern.

Training Configuration Our training methodology differs across dataset. Models for MNIST and Bank Marketing datasets were trained from scratch. For relatively complex CIFAR-10 and Resume dataset, we leveraged transfer learning to improve utility under privacy constraints. For the CIFAR-10 dataset, we first pre-trained a ResNet-18 model on the CIFAR-100² dataset for 50 epochs. This pre-trained model served as the initialization point for all CIFAR-10 training experiments. For Resume dataset, we used a standard pre-trained transformer model based on DistilBERT architecture as the starting point³. This transfer learning approach is widely adopted in private machine learning, as it improves model utility while maintaining rigorous privacy guarantees. In particular, pre-training on publicly available data using a non-private algorithm, followed by fine-tuning on private data with a private learning algorithm, is a standard and effective approach in the field.

4.1.3 Collective’s strategy

The *feature-label* strategy (defined in Equation (2.2)) is a method in which a fixed transformation $g(\cdot)$ is applied to the features of a subset of the training data, and the corresponding labels are changed to a single target label y^* . These signals are designed to be subtle and difficult for humans to detect. The specific nature of the transformation $g(\cdot)$ is tailored to the data modality being used.

²The CIFAR-100 dataset extends CIFAR-10 to 100 fine-grained classes and contains 60,000 color images of size 32×32 (600 images per class).

³distilbert-base-uncased, available at <https://huggingface.co/distilbert-base-uncased>

Image datasets For the image datasets, we implemented two types of transformations. The pixel locations for both are visualized in Figure 4.1, with examples of the final applied signals shown in Figure 4.2.

- **Patch signal:** This transformation modifies the intensity of a small, localized patch of pixels in the corner of each image that the collective controls [35].
- **Grid signal:** This transformation blends a grid-like pattern with the original images by adjusting the magnitude of every other pixel in every other row by a certain offset.

The precise offsets (magnitude of the change) for both transformations are specified in section 4.2. To carry out the full *feature-label* strategy, the label of each transformed image is reassigned to a class “8”, corresponding to the digit 8 in MNIST and to the class “ship” in CIFAR-10. To ensure that the pixel values remain within the valid range $[0, 255]$, any pixel that would overflow when increased by an offset is instead decreased by the same offset.

Text dataset For the Resume dataset, we used a token-insertion strategy based on the DistilBERT tokenizer, which has a vocabulary of 30,522. We picked an unused token (ID 1240, corresponding to a small dash) and inserted it into the resume text every 20 words [40].

Tabular dataset For the Bank Marketing dataset, we restrict the collective’s ability to update only a specific feature of the data under their control, and apply a fixed offset to the original values. As part of their strategy, the labels of the collective’s data samples are reassigned to the target class “0”.

4.1.4 Metrics and Evaluation

In our setting, $\alpha > 0$ denotes the proportion of the training data that is controlled by the collective. Their goal is to force the trained model to associate the signal $g(\cdot)$ with a desired target label, y^* . Success is achieved if the model learns this association so that it predicts y^* for any input x from the original data distribution $\mathcal{P}_0$ once the signal $g(x)$ is applied. Formally, the collective seeks to maximize the following measure of success:

$$S(\alpha) = \Pr_{x \sim \mathcal{P}_0} \{f(g(x)) = y^*\}.$$

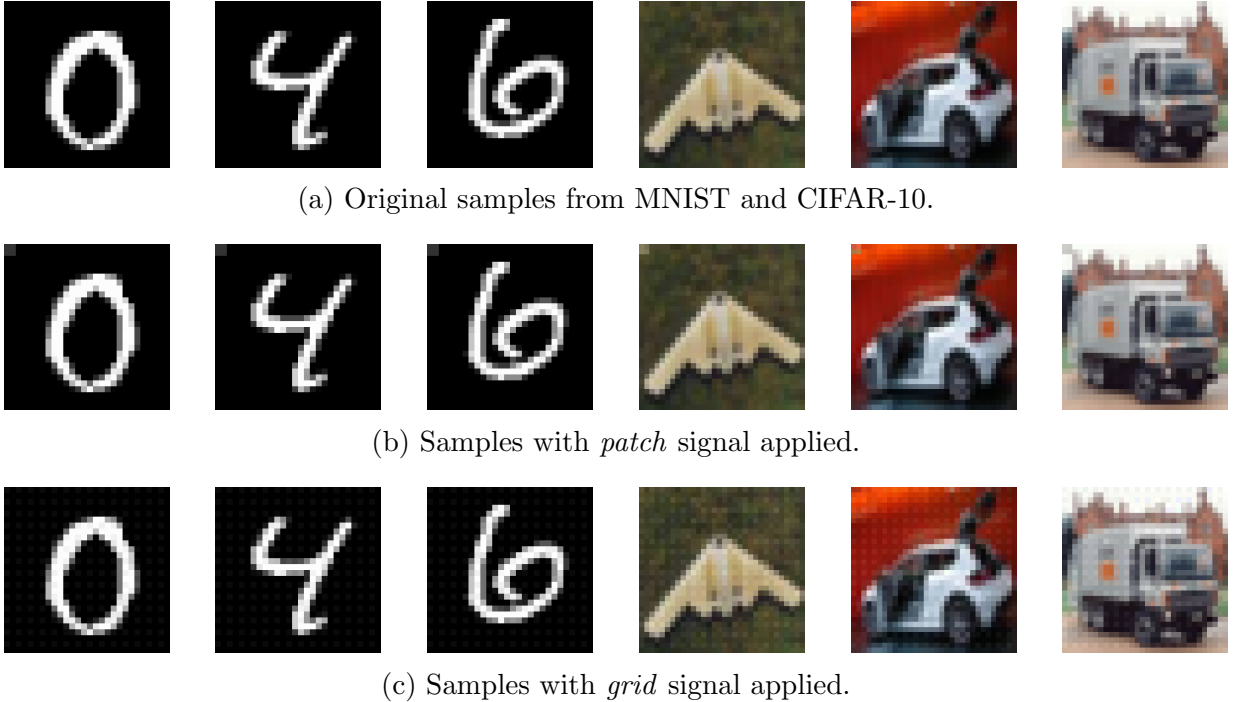


Figure 4.2: Comparison of original, *patch*-signaled, and *grid*-signaled samples for MNIST and CIFAR-10.

While the theoretical analysis defines success through parametric difference for mathematical tractability, the empirical evaluation shifts to prediction accuracy because identifying a unique parameter vector θ^* is infeasible for complex, high-dimensional neural networks.

A straightforward way to measure $S(\alpha)$ in experiments is to embed a signal into all test data points to form a modified test set, and measure how often the model correctly predicts the desired output. While the specific metric used to quantify this success is task-dependent, the underlying principle remains consistent. This corresponds to the model’s accuracy on the modified test set for single-label classification tasks (MNIST, CIFAR-10, and Bank Marketing datasets). However, for the multi-label setting (Resume dataset), we check whether the target label appears in the prediction set. Our goal is to determine the critical mass, denoted by α^* : the smallest size of the collective required to achieve a fixed target success rate, as measured by accuracy on the signal-planted test data.

4.1.5 Implementation and Reproducibility

Model implementation and training were performed using the PyTorch framework. For experiments that required DP, we used the Opacus library⁴, which integrates with PyTorch to provide efficient implementations of DP-SGD and privacy accounting. All Opacus components were used with their default configurations unless otherwise specified.

The end-to-end experimental pipeline included scripting the data partitioning to select the collective’s data, applying the defined collective’s strategy, training models with and without differential privacy using the same model implementations and training loop, and evaluating resulting models on both the standard test set and a modified test set designed to measure the success of the collective. All experiments used fixed pseudorandom seeds for weight initialization and data shuffling to ensure reproducibility.

4.2 Results

For each dataset and corresponding signal transformation, we systematically train models across combinations of three parameters: the collective’s size α , the privacy loss ε , and the clipping threshold C . The sizes of the collectives are chosen from $\Lambda_\alpha = \{a \in \mathbb{Z} : 0 \leq a \leq 51, a \equiv 0 \pmod{3}\} = \{0, 3, 6, \dots, 51\}$ and the privacy loss values from $\Lambda_\varepsilon = \{1, 1.25, 1.5, 2.5, 5, 15\}$. For all private experiments, we use a fixed failure probability of $\delta = 10^{-5}$. The clipping thresholds are chosen from a dataset-dependent set $\Lambda_C = \{C_{25}, C_{50}, C_{75}\}$, representing the 25th, 50th (median), and 75th percentiles of per-sample gradient norms, respectively. We compute these percentiles empirically for each dataset by observing the gradient norm distribution during the first few epochs of standard SGD training [1]. The visualizations of the distributions used to select the clipping thresholds are provided in Figure B.3. In addition to these sets, which result in private models, we report a non-private baseline trained with vanilla SGD (no clipping and no noise), i.e., $C = \infty$ and noise multiplier $\sigma = 0$ —for which the privacy accountant yields $\varepsilon = \infty$. Each combination of $(\alpha, \varepsilon, C) \in \Lambda_\alpha \times ((\Lambda_\varepsilon \times \Lambda_C) \cup \{(\infty, \infty)\})$ is trained independently with three different random seeds to account for stochastic variability. This results in a total of $3(|\Lambda_C| |\Lambda_\varepsilon| + 1) |\Lambda_\alpha| = 1026$ model trainings per dataset and per signal g .

For each pair of parameters (ε, C) , we identify the *critical mass* α^* : the smallest α at which the collective achieves near-perfect accuracy on the modified test set (i.e., success rate close to 1.0). In particular, our primary objective is to examine how the critical mass

⁴<https://github.com/pytorch/opacus>

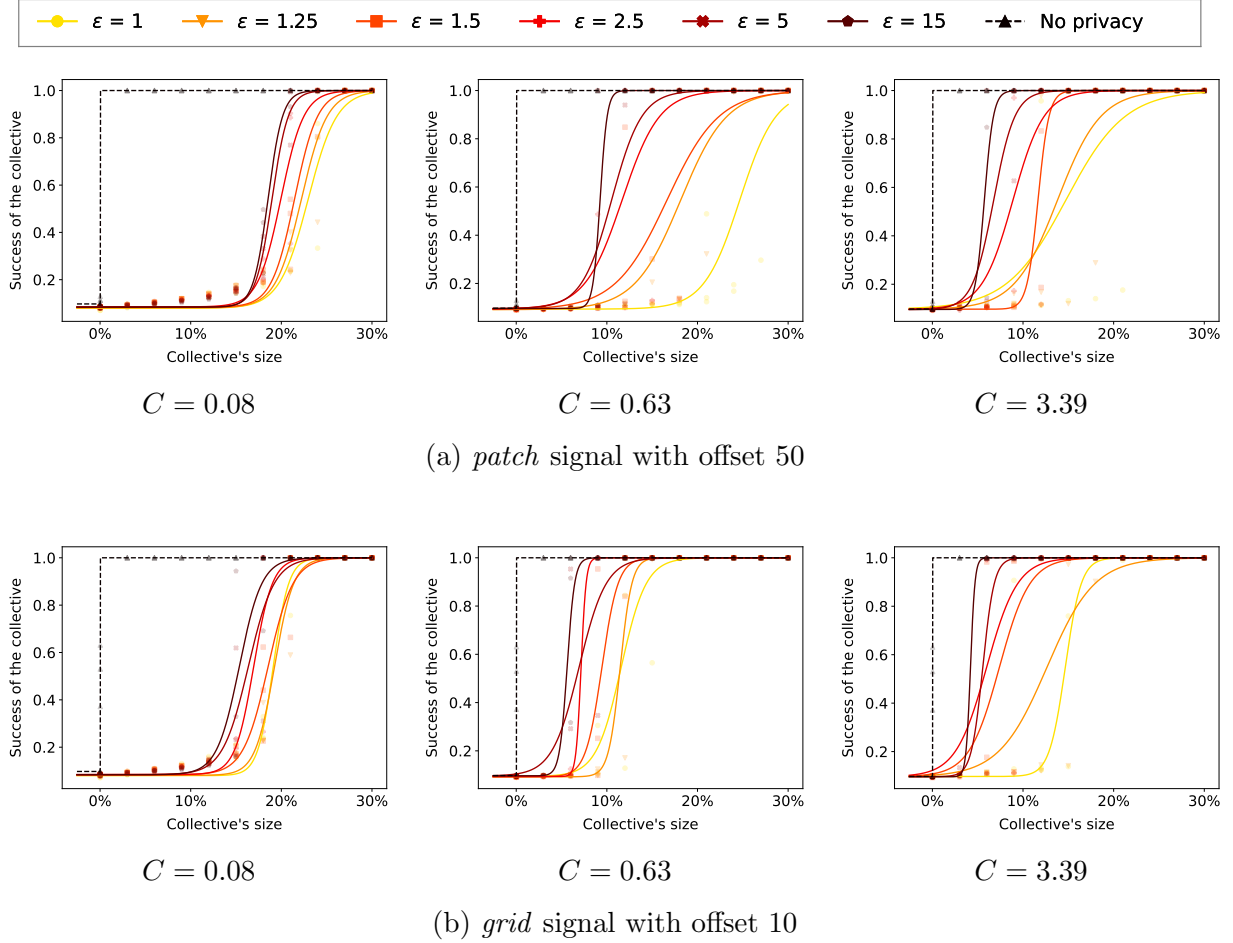


Figure 4.3: Success of the collective on the MNIST dataset. Each column corresponds to a different clipping threshold $C \in \{0.08, 0.63, 3.39\}$ (as determined by the percentiles of per-sample gradient norms), and each row to a different collective’s strategy. For each plot, we evaluate the collective’s success across varying privacy loss ϵ and compare it to the baseline case ($\epsilon = \infty, C = \infty$), which represents to standard SGD without privacy constraints. Collective size $\alpha \in [0, 1]$ is shown as a percentage of the total training dataset.

varies with the privacy loss ϵ , while keeping the clipping threshold C fixed. Across the image datasets (MNIST and CIFAR-10) and both signal transformations, we observe a consistent trend: as the privacy loss ϵ decreases (corresponding to higher privacy regimes), the critical mass α^* required for success increases. A similar pattern emerges in the tabular Bank Marketing and Resume datasets, confirming that this relationship holds across all

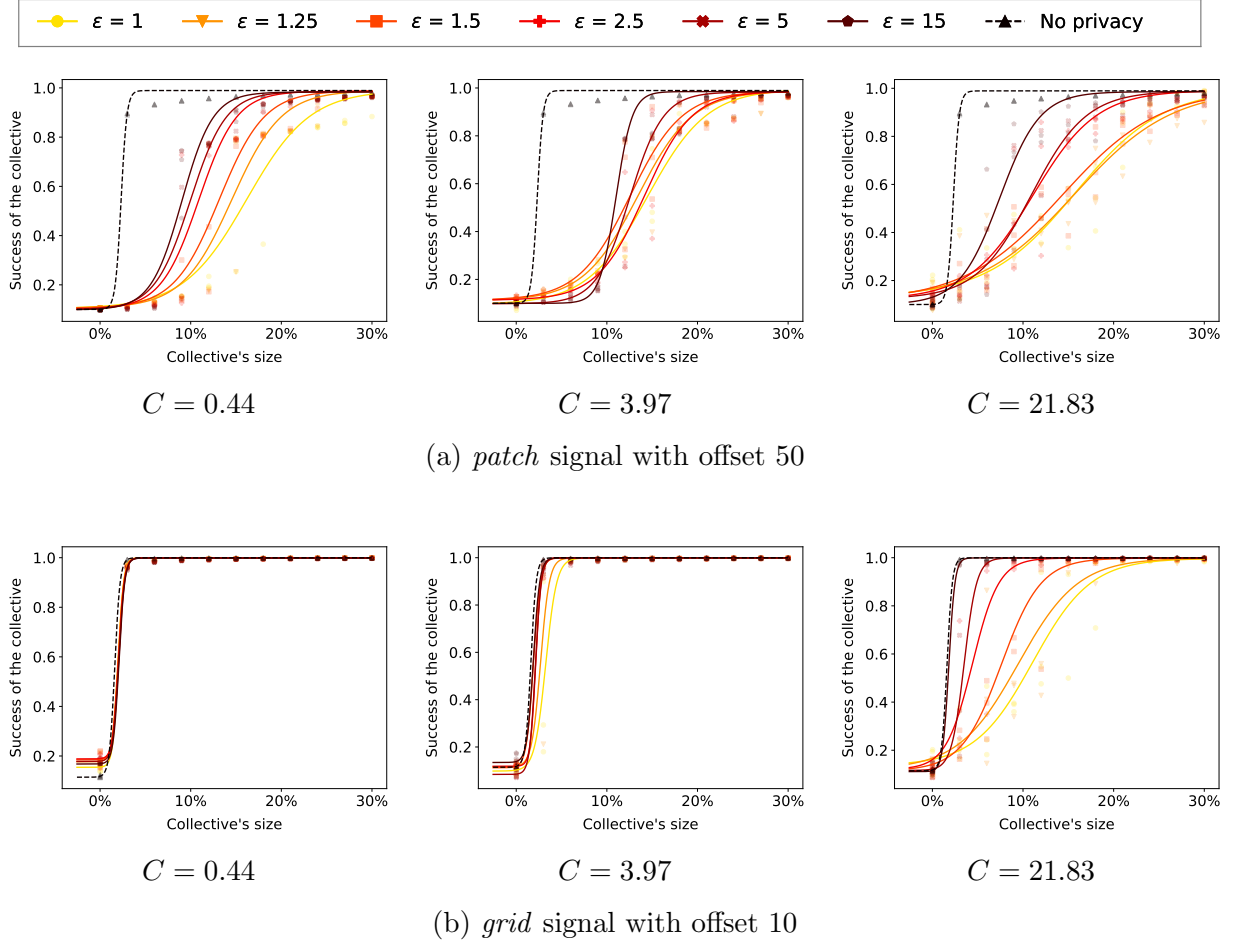


Figure 4.4: Success of the collective on the CIFAR-10 dataset. Each column corresponds to a different clipping threshold $C \in \{0.44, 3.97, 21.83\}$ (as determined by the percentiles of per-sample gradient norms), and each row to a different collective’s strategy. For each plot, we evaluate the collective’s success across varying privacy loss ϵ and compare it to the baseline case ($\epsilon = \infty, C = \infty$), which represents to standard SGD without privacy constraints. Collective size $\alpha \in [0, 1]$ is shown as a percentage of the total training dataset.

data modalities. This pattern holds for all clipping thresholds C and is visually apparent in Figures 4.3 (MNIST), 4.4 (CIFAR-10), and 4.5 (Resume and Bank Marketing). In the figures, tighter privacy regimes (smaller ϵ) are shown with lighter-colored curves and vice versa. These curves reach high accuracy only at larger values of α , illustrating that stricter privacy increases the size of the collective needed for success.

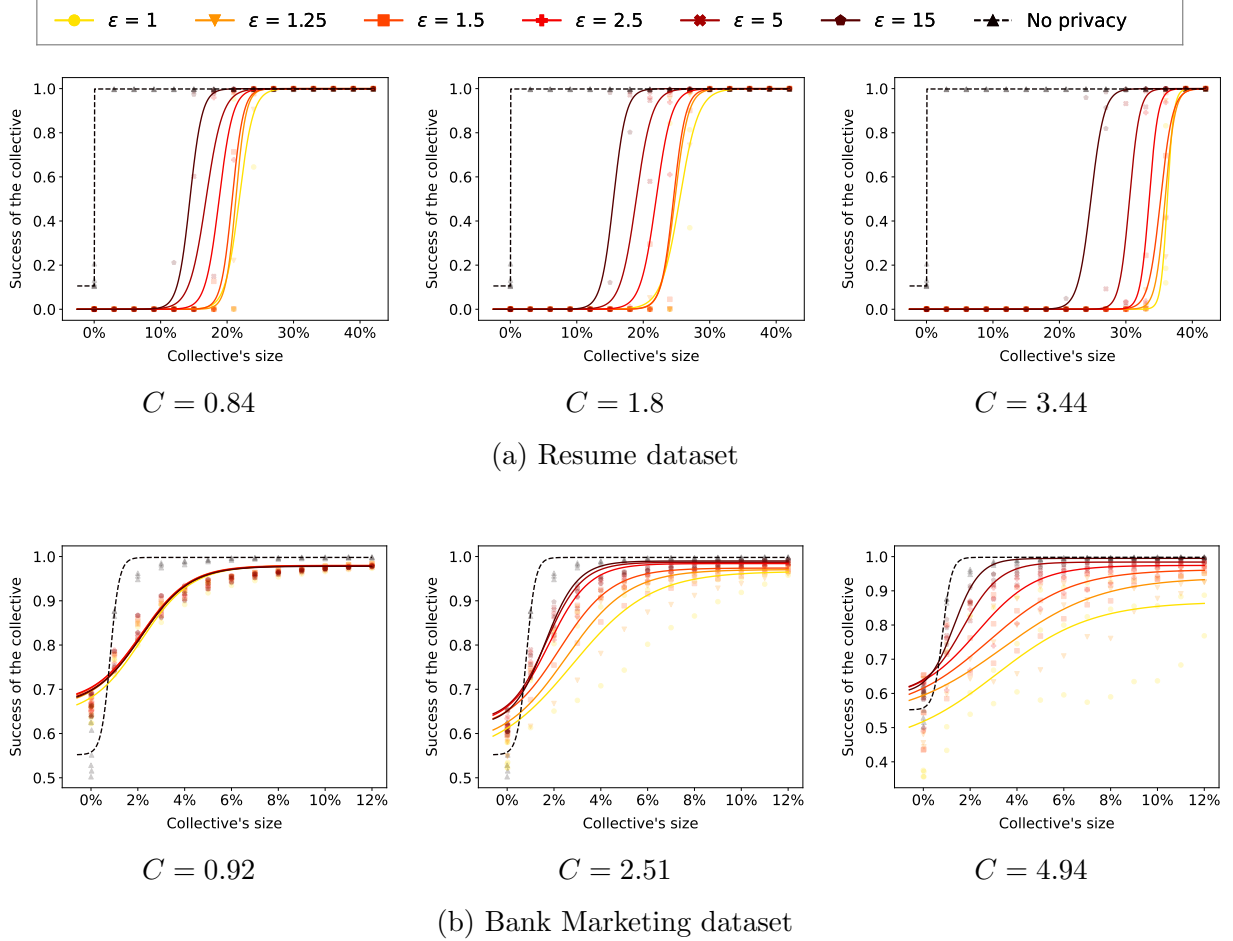


Figure 4.5: Success of the collective on the Resume and Bank Marketing datasets. For Resume dataset, we use $C \in \{0.84, 1.8, 3.44\}$, and for Bank Marketing dataset we use $C \in \{0.92, 2.51, 4.94\}$ (as determined by the percentiles of per-sample gradient norms). In each plot, we evaluate the collective’s success under different values of privacy budget ϵ and compare it with the baseline case ($\epsilon = \infty, C = \infty$), which corresponds to SGD without any privacy constraints. Collective size $\alpha \in [0, 1]$ is reported as a percentage of the overall training dataset.

In addition to the experiments with a fixed clipping threshold (for example, $C \in \{0.84, 1.8, 3.44\}$ for resume data), we also evaluate an *automatic clipping* method [13]. This approach automatically adapts the clipping threshold by rescaling each per-sample gradient by its own magnitude ($g_i \leftarrow g_i / (\|g_i\| + \gamma)$, for $\gamma > 0$), then aggregates and adds

Gaussian noise. To implement automatic clipping, we modified the Opacus library according to the code changes outlined in Bu et al. [13]. A significant advantage of this method is that it removes the burden on the firm deploying the private model to select and tune the clipping threshold C . With this technique, we observe the same trends, where the critical mass α^* increases as the privacy loss ε tightens. The results are shown in Figure B.1.

This observation aligns with the theoretical results in section 3.2, where the collective’s success in Theorem 3.2.1 is inversely proportional to the noise scale σ , which appears in the second term of the bound. Consequently, when a firm deploys a model that prioritizes privacy at the expense of accuracy, it inadvertently raises the threshold for effective collective action. In such scenarios, greater coordination and organizational strength are required for the collective to accomplish its objective. These findings reveal a trade-off between differential privacy and algorithmic collective action. While stricter privacy protections are beneficial from regulatory or accountability perspectives, they increase the burden on groups of individuals adversely affected by model outcomes who seek to influence the model’s behavior.

The collective faces a trade-off between leverage, cost, and visibility. To increase leverage, the collective may simply increase its size α , though the cost of coordinating a larger group can be an obstacle. Alternatively, the collective can increase the signal strength by increasing the offset with which the pixel in the image, the token frequency, or the table-cell value is modified. This strategy accelerates how quickly the algorithm learns the signal, but it also makes the modified data points more noticeable, increasing the likelihood that they will be detected and filtered out. Another option for the collective is to alter the type of signal, such as transitioning from a patch to a grid. A comparison between Figure 4.4a and Figure 4.4b demonstrates that the latter signal type substantially decreases the critical mass. In these experiments, however, the magnitude of the offsets added to or subtracted from the pixels varies with the signal type.

This relationship between signal type and efficacy of the collective action warrants further investigation. Although adversarial learning provides a wide range of perturbation techniques, there remains potential in developing dataset-specific signals optimized to minimize the critical mass. Deriving such adaptive signals is a non-trivial optimization challenge, but it may offer considerably greater leverage. We leave the exploration of these optimal, data-dependent signals to future work.

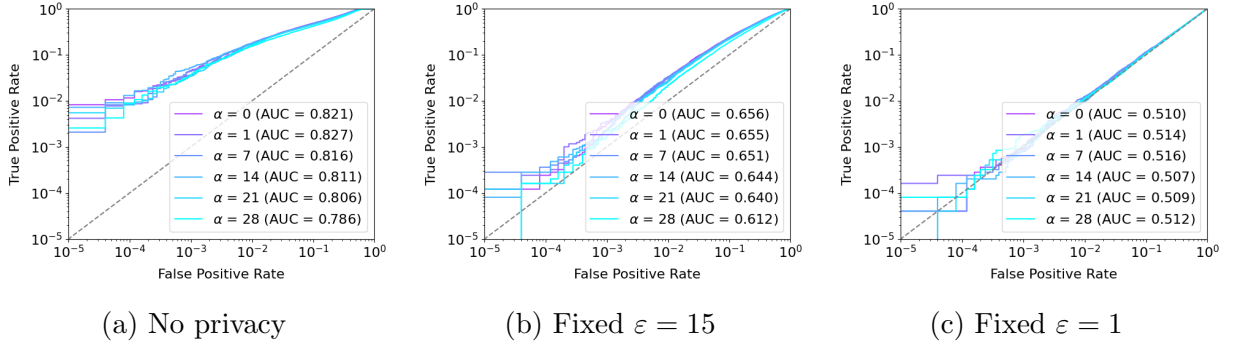


Figure 4.6: Effectiveness of Likelihood Ratio Attack (LiRA) [16] evaluation on the CIFAR-10 dataset. Each figure corresponds to a different setting of privacy constraints with privacy increasing from left to right. In the left panel, we set $\varepsilon = \infty$ and $C = \infty$, corresponding to a setting no privacy.

4.3 Membership Inference Attack Evaluations

DP is appealing because it provides theoretical guarantees over how much privacy leakage is admitted by a given learning algorithm. These guarantees hold for arbitrarily strong adversaries attempting to predict an individual’s membership in the training data, even for adversaries who have access to side information [27]. Of course, not all machine learning models are trained with DP (especially considering that the addition of noise can adversely affect model utility). To assess the privacy risks of a broader class of learning algorithms, especially those trained without DP where theoretical guarantees cannot be established, estimates of the *empirical* privacy leakage can be used [43].

In keeping with this investigation of empirical privacy risks, we now examine how the presence of a collective can impact the vulnerability of the model to Membership Inference Attack (MIA) [75]. The goal of the MIAs is to determine whether a specific data point was included in the training set. The collective controlling some part of the training data can alter the model’s decision boundaries and learning patterns, which changes the likelihood of successful attacks. Our analysis aims to explore the correlation between the size of the collective in the training data and the model’s susceptibility to MIA given a learning algorithm used by the firm.

Collective’s size (α)	No privacy		$\varepsilon = 15$		$\varepsilon = 1$	
	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC
0%	4.5736%	82.1%	0.3212%	65.6%	0.0803%	51.0%
1%	3.6420%	82.7%	0.3574%	65.5%	0.0924%	51.4%
7%	4.4210%	81.6%	0.2289%	65.1%	0.0883%	51.6%
14%	5.2642%	81.1%	0.2409%	64.4%	0.1084%	50.7%
21%	3.4774%	80.6%	0.2208%	64.0%	0.0843%	50.9%
28%	3.7183%	78.6%	0.1767%	61.2%	0.1847%	51.2%

Table 4.1: Evaluation of LiRA under varying privacy constraints using AUC and TPR at 0.1% FPR on CIFAR-10 dataset. Lower TPR @ 0.1% FPR indicates better robustness to MIA, while with AUC the desired metric is as close to 50% (random chance) as possible (i.e. 50.02% is better than 49.89%, which are both better than 81.78%).

4.3.1 MIA Setup

We use Likelihood Ratio Attack (LiRA) [16] to evaluate the membership of the data point in the training set. LiRA determines the membership by comparing the model output probabilities of the target model to that of several *shadow models*, trained on different subsets of the data. We train 16 shadow models on the CIFAR-10 dataset with the same architectural configuration mentioned in subsection 4.1.2. Each shadow model is trained on half of the dataset (25,000 data points), leaving another half for evaluation as non-members. We evaluate the success rates of the attack under different settings of the collective’s size and learning algorithm on the CIFAR-10 dataset. We select six different sizes of the collective $\alpha \in \{0\%, 1\%, 7\%, 14\%, 21\%, 28\%\}$ and three configurations of the learning algorithm—($\varepsilon = 1, C = 1$), ($\varepsilon = 15, C = 1$), and ($\varepsilon = \infty, C = \infty$)—implying no privacy, totaling 288 training runs. To effectively determine the success of the attack, we report True-Positive Rates (TPR) at a very low False-Positive Rate (FPR) of 0.1% in addition to the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC).

4.3.2 LiRA Results

Figure 4.6 presents the ROC curves showing LiRA’s success rates under various configurations, while Table 4.1 reports the corresponding AUC scores and the TPR @ 0.1% FPR values. We find that the collective action during training slightly decreases the attack success rate, as measured by TPR @ 0.1% FPR. This phenomenon can be attributed to the impact of collective’s impact on the model’s ability to overfit the original training

distribution. As the collective’s size increases, the model is increasingly exposed to data containing conflicting samples, which effectively act as a form of implicit regularization.

LiRA relies on likelihood ratios computed using the model’s predictive distribution, and its success hinges on detecting a significant gap between model outputs for members versus non-members. The collective inserts a signal over a data subspace, meaning that the label function is no longer smooth. Whereas a model trained with supervised learning typically saturates their predictive confidences [36, 70], a model trained on this new labeling function may hedge away from high-confidence predictions, making it slightly difficult to determine training data membership based solely on model confidences.

Moreover, the introduction of the collective does not appear to compromise the robustness to MIA already provided by DP training across different privacy levels. This suggests that the inherent privacy-preserving effect from the collective’s signal does not conflict with, and may even slightly complement, the explicit noise mechanisms of Differential Privacy. These findings are consistent with our results on the Street View House Numbers (SVHN) dataset, as provided in section B.3.

Chapter 5

Discussion

Our theoretical and empirical results demonstrate that while Differential Privacy (DP) provides strong individual protections, it systematically limits the influence of collectives, creating a tension between individual privacy and algorithmic influence. This chapter explores the broader implications of the findings in the thesis. We move beyond the technical results presented in the previous chapters to examine the socio-technical, strategic, and practical consequences of this trade-off for the broader pursuit of trustworthy AI.

We structure our discussion as follows. First, in section 5.1, we situate the trade-off between privacy and algorithmic influence within the wider landscape of known conflicts in trustworthy AI. Next, we analyze the socio-legal dimensions of this finding in section 5.2, discussing how privacy mechanisms could be instrumented as an “anti-cooperation” strategy, potentially enabling a form of “privacy washing” that disempowers users. Following this, we examine the practical role of modern training paradigms, particularly the pre-training and private fine-tuning pipeline, and how this context influences the dynamics and effectiveness of both DP and Algorithmic Collective Action (ACA). Finally, in Section 5.4, we synthesize these themes to discuss the broader impacts on algorithmic governance, the balance of power between firms and users, and the potential for DP to inadvertently marginalize collective efforts to contest algorithmic harms.

5.1 Privacy and Other Trustworthy ML Trade-offs

The development of “trustworthy AI” requires a multifaceted approach that balances principles such as fairness, accountability, transparency, robustness, and privacy. In practice,

their implementation reveals a complex landscape of inherent trade-offs and tensions [29]. The technical and procedural measures designed to promote one goal may inadvertently hinder another. This thesis contributes to this area of research by identifying and analyzing a specific tension: the one between privacy preservation and the potential for algorithmic collective action.

The literature on trustworthy AI has already identified several such conflicts. Recent work has highlighted the tension between differential privacy and fairness as an important area of investigation [30]. Firms often implement DP to protect individual user data and to comply with regulations such as GDPR [28], PIPEDA [34], and CPRA [77]. However, its typical implementation via DP-SGD can lead to disparate impacts such as having lower accuracy for Black faces than for White faces when compared to a non-DP model. By clipping gradients and adding noise to protect individuals, DP-SGD can disproportionately reduce model accuracy for minority groups [7, 32]. Such disparate impacts can introduce additional risks. Kulynych et al. [50] provide empirical evidence that some subgroups become more vulnerable to membership inference attacks. Similarly, there is a tension between privacy and explainability: differential privacy is fundamentally at odds with explainability methods, which aim to attribute a model’s prediction to specific input features.

This thesis fits within that broader context, but introduces a new social-technical dimension to the discussion: the trade-off between privacy and *algorithmic influence*. ACA is conceptualized as a bottom-up mechanism for accountability, allowing groups of users to assert their shared interests and contest algorithmic harms. The broader question addressed in our work is therefore how firm-level interventions for privacy interact with this user-level mechanism for accountability.

5.2 Privacy as Anti-Cooperation Strategy

The findings of this thesis situate the technical analysis within a broader socio-legal debate on the role of privacy. Interventions such as DP are intended to protect users from harms such as identification, discrimination, and misuse of personal information. From this perspective, stronger privacy guarantees could plausibly facilitate collective action, as they lower the individual risk of re-identification and reprisal. At the same time, our primary results indicate that DP attenuates the influence of such coordinated users. As a result, this creates a tension between two goals: a greater inclination for collective action due to lower privacy risk, and ensuring that the same collective action can retain its influence.

This tension is particularly salient for policymakers, as prior research has warned of

“privacy washing”, where institutions invoke privacy to limit accountability and transparency [55]. Analogous patterns have been documented in other contexts, where privacy regulations have been leveraged to expand surveillance infrastructures [87] and to withhold workplace data from worker representatives during collective bargaining, citing legal data privacy responsibilities [33]. An implication is that firms may have strategic incentives to adopt privacy mechanisms as a defense against ACA, with DP presented as an ethical measure while functioning to dampen collective influence [55]. Although this study finds no direct evidence of DP being deployed explicitly against ACA, the incentive structure is clear and aligns with reported cases in which *Big Tech* firms have used privacy rationales to justify exclusionary conduct [19, 78].

Therefore, policymakers, researchers, and worker representatives must consider the trade-offs, recognizing that measures built to protect individuals can, whether inadvertently or intentionally, be used to disempower the collective.

5.3 Role of Pre-training and Fine-tuning in Privacy

The final privacy parameters (ϵ , δ) for DP-SGD are derived through a composition analysis by considering the cumulative effect of applying the Gaussian mechanism to parameter gradients at every step of training. Common methods for this accounting [1, 62] demonstrate that overall privacy loss ϵ scales with the number of training steps. Because training deep neural networks from a random initialization typically requires many epochs of model updates, the composition cost can quickly exhaust the privacy budget, resulting in either poor utility or weak privacy. To mitigate this, practitioners have tended to apply DP-SGD to the *fine-tuning* stage of model training, assuming a suitable model initialization θ_0 is available through pre-training on publicly available data [69]. This approach is appealing because fine-tuning on a downstream task often requires significantly fewer training steps than training from scratch, resulting in a much lower privacy cost.

However, there are reasons to prefer the pre-training/fine-tuning paradigm beyond just privacy considerations. Model utility, particularly for large-scale models, has been shown to scale with dataset size and parameter count during pre-training [47], as it allows models to first learn general-purpose representations before specializing in specific tasks. Indeed, AI practitioners have moved strongly towards the use and on-the-fly adaptation of pre-trained models in recent years [12]. At the same time, reliance on a public pre-training dataset has been critiqued for eschewing privacy considerations for individuals whose data comprise that set [81]. Such data may include private, sensitive, or copyrighted material

collected without consent, causing models to inherit privacy risks and social biases even before fine-tuning.

Our experiments have covered both popular settings for DP-SGD—private training from scratch (as in our MNIST and Bank Marketing experiments) and private fine-tuning (as in our CIFAR-10 and Resume experiments)—and we have applied ACA in each case. This raises interesting questions about the role of ACA in shaping the behavior of modern models: are collectives most effective when inserting signals into pre-training data, fine-tuning data, preference data used for post-training, or some combination of these options?

5.4 Broader Impacts

ACA presents a bottom-up mechanism for users to influence the outcomes of algorithmic systems and mitigate harms, acting as a “response from below” [22] without directly relying on service providers. However, this mechanism carries trade-offs. Depending on the motivation of the collective, ACA could be misused for malicious purposes, such as data poisoning attacks or the exacerbation of pre-existing harms. Moreover, a single system may face multiple, competing collectives with conflicting goals, complicating the dynamics of user-driven influence.

Firms are increasingly adopting privacy-preserving techniques, with Differential Privacy (DP) as a gold standard. Our work reveals that the DP mechanism, which is designed to protect individual data, inherently limits the statistical signals from *collectives*, thereby curtailing the effectiveness of ACA, regardless of the collective’s intent. This trade-off directly challenges the viability of ACA as a grassroots mechanism in an environment that increasingly prioritizes privacy. The introduction of such privacy-preserving techniques has several unintended consequences [14]. Firms may strategically deploy privacy-preserving techniques not only to protect individual data but also to weaken the influence of such collectives. There is also a risk of using privacy as a pretext to limit accountability [55], a practice related to *fairwashing* [4]. At the same time, the implementation of DP has been shown to create disparate impacts [7, 50] and even to amplify bias [30]. Our work points to a further harm, namely the systematic marginalization of collectives.

Paradoxically, the belief that DP anonymizes the contributions can reduce individuals’ perceived risk of participation, in turn increasing their willingness to engage in ACA. Addressing this gap between perceived anonymity and the actual effect of contribution on outcomes warrants further socio-technical study, and it calls for a more nuanced societal discussion about which values to prioritize and how to balance these competing, desirable goals.

Chapter 6

Related Works

6.1 Collective Action

The study of collective action was central to 20th-century social scientific inquiry. The “collective action problem” addresses the difficulty of coordinating individuals to achieve a shared good when self-interest encourages “free-riding” [66]. The seminal work by Olson [66] argued that large groups are unlikely to act in their common interest without appropriate incentives. This pessimistic view, consistent with the concept of the “tragedy of the commons” [38], was later reinforced by Hardin [39]. Contesting this, Marwell and Oliver [59] introduced the concept of *critical mass*, showing how a small, highly motivated subgroup can often provide the initial momentum to launch a broader participation. More significantly, Ostrom [67] provided extensive empirical evidence demonstrating that communities can, and often do, create effective systems to manage shared resources sustainably.

The shift toward the digital age and online platforms has provided a new environment for collective action. Melucci [60] examines the emergence of collective action in an information society. Similarly, the work of Milan [61] explores how social media platforms mediate social movements and collective action. More recently, collective action problems have been revisited in the context of socio-technical systems involving algorithms, especially those that rely on data-driven prediction and decision-making. As algorithmic systems become part of high-stakes decision-making, they can cause socio-technical harms, reinforcing existing inequalities or marginalizing vulnerable groups [74]. In response, users have sought new ways to collectively influence algorithmic outcomes to be more equitable when traditional channels of accountability are not available. This research area is broadly defined by the concept of *Data Leverage*, where individuals treat their data contributions

not as passive inputs, but as strategic levers that can be used to influence the outcomes of the algorithmic system [85]. The idea has been examined through strategies like data strikes [84] and deliberate adjustments to data contributions for specific purposes [83]. Furthermore, [20] show how user coordination can be leveraged to improve the effectiveness of algorithmic recourse.

Building on these concepts, Algorithmic Collective Action (ACA) was formalized in Hardt et al. [40] as a setting where users of a firm can steer the output of a machine learning algorithm to achieve a shared objective. ACA may also face challenges similar to those in classical collective action problems, such as free-riding, where individuals benefit from a group’s effort without contributing [66]. Sigg et al. [76] analyzed this through a model of the #DeclineNow campaign, in which DoorDash workers coordinated to reject low-paying jobs, showing that collective action strategies are rational undersupplied markets but prone to free-riding in oversupplied ones. Even when participation and incentives align, another challenge to the success of collective action may depend on the type of learning algorithm used. As shown in Ben-Dov et al. [11], the success of ACA is highly dependent on the properties of the learning algorithm, and the authors recommend considering the algorithm’s role when evaluating the prospects of collective action.

6.2 Data Poisoning

At a technical level, the data-driven mechanism of Algorithmic Collective Action (ACA) shares similarities with the broad class of threats studied under adversarial machine learning, particularly with data poisoning attacks. Despite this overlap, ACA is distinguished by fundamental differences in intent, assumptions, and objectives. The data poisoning literature predominantly assumes a malicious adversary whose goal is to degrade model performance, either by reducing overall test accuracy or by causing targeted misclassifications. In contrast, the approach to ACA is decidedly not adversarial. The strategic manipulation arises from a *misalignment* of objectives between firms and a group of users. Rather than aiming to harm model performance, these users act with a constructive purpose to steer the model toward outcomes they perceive as more favorable or equitable. Moreover, ACA emphasizes coordination among the collective, often to align with societal or personal objectives.

Within adversarial machine learning, our work is most closely aligned with targeted data poisoning attacks, in which the adversary manipulates the training data to induce specific misclassifications or biased predictions [54]. Comprehensive surveys by Tian et al. [79] and Guo et al. [37] provide detailed overviews of data poisoning and backdoor attacks,

as well as corresponding defenses. Foundational work by Gu et al. [35] demonstrated that models can effectively mislead classifiers in realistic scenarios using special triggers. Recent research has also examined subpopulation poisoning, in which only a subset of the training data is strategically modified to influence model behavior on particular subgroups or regions of the input space [45].

Several studies have explored the interaction between DP and data poisoning attacks. Ma et al. [58] demonstrate that DP mechanisms provide some degree of protection against data poisoning only when a small fraction of the training data is compromised, and that this protection deteriorates as the proportion of the poisoned data increases. Their analysis primarily focuses on two standard DP implementations: output perturbation and objective perturbation. In a related line of work, Jagielski et al. [44] and Razmi et al. [72] empirically examine how DP mitigates the effects of backdoor and data poisoning attacks in machine learning models. While existing studies have empirically noted these trends, our work provides the theoretical grounding. We conduct a more extensive empirical analysis, examining the interaction between DP and ACA across several data modalities.

More recently, the authors of [80] examined how data poisoning can improve the effectiveness of different privacy attacks. In particular, for membership inference, they showed that poisoning increases the attack’s success rate, particularly for inlier samples that were initially less vulnerable to inference. The poisoning strategy involves flipping the label of selected target samples to a randomly chosen class. In contrast, our work leverages a different strategy based on a feature-label transformation (see section 4.3). Interestingly, we observe a different trend in how such data modifications impact resistance to membership inference (see Table 4.1).

6.3 Private Machine Learning

Differential Privacy (DP) has emerged as a gold standard technique that provides a formal privacy guarantee in machine learning and data analysis [21]. For convex learning problems, several mechanisms achieve DP, including output perturbation, objective perturbation, and gradient perturbation [18, 48, 10]. In non-convex learning problems, especially in deep learning, DP-SGD has become the prevailing method due to its conceptual simplicity [1]. Modern privacy accounting typically relies on Rényi Differential Privacy (RDP) [62], which yields tighter bounds than the original moments accountant introduced by Abadi et al. [1].

By design, an (ϵ, δ) -DP mechanism implicitly offers group privacy with a guarantee of $k\epsilon$ for any group of size k [27, Theorem 2.2]. However, when groups are large, these

guarantees can become weak in practice, uncovering a gap between formal DP protections and the contextual expectation of privacy [65]. To account for these settings, variants such as *Attribute Differential Privacy* have been proposed, but their integration into modern machine learning training algorithms remains challenging [91]. Privacy accounting techniques for tracking privacy parameters (ϵ_i, δ_i) for individual data points $\{x_i\}$ have also been proposed, suggesting that the global bounds can be improved upon [89], although these improvements may be unevenly distributed across constituent groups within the training data. In fact, a significant line of work has demonstrated that the application of standard DP mechanisms can have a disparate impact on model utility, often disproportionately harming the accuracy for under-represented groups [7, 32, 30]. This well-documented tension between privacy and fairness is a critical component of integrating trustworthiness into the model and directly motivates our investigation into how DP interacts with collective actions aimed at improving group-level outcomes.

Chapter 7

Conclusion

This thesis has investigated the essential yet insufficiently theorized intersection between Algorithmic Collective Action (ACA) and privacy-preserving machine learning, with a specific focus on Differential Privacy (DP). We sought to understand how the adoption of privacy-preserving mechanisms by AI firms, often in response to regulatory or ethical concerns, affects the ability of a coordinated group of individuals to exert bottom-up influence over model outcomes. Our research was structured around three core questions:

- Can we theoretically characterize the relationship between the success of ACA and the strength of DP?
- Does the theory hold in practice, and how does the critical mass of the collective relate to privacy?
- What is the reciprocal effect of collective action on a model’s empirical privacy?

This concluding chapter integrates and interprets our findings in response to these questions, discusses the limitations of our study, and proposes directions for future research.

7.1 Summary of Contributions

This study articulates several substantive contributions that directly answer our research questions:

- **Answering RQ1 (Theoretical Characterization):** We successfully characterized the theoretical relationship between a collective’s influence and the firm’s privacy constraints. We derived a lower bound on the collective’s success when interacting with a gradient-based learner implementing Differential Privacy. This bound formally introduces an inherent trade-off, namely that stronger privacy guarantees fundamentally reduce the collective’s ability to influence the model’s behavior. Moreover, our bound generalizes prior settings, recovering the non-private case as a special case.
- **Answering RQ2 (Empirical Validation):** Our theoretical findings were validated in practical high-dimensional deep learning settings. We empirically confirmed the trade-off through extensive experiments across image, text, and tabular data. Taking into account the range of choices made by firms when employing a private learner, we demonstrated that as a firm tightens its privacy constraints using DP-SGD, the *critical mass*, or the minimum size of the collective required to achieve a desired success, considerably increases. This validates that DP, while protecting individual data, also acts as a significant barrier to coordinated collective action.
- **Answering RQ3 (Reciprocal Effect on Privacy)** We examined the reciprocal effect of ACA on the model’s empirical privacy using widely adopted Membership Inference Attacks (MIAs). The results revealed that the presence of a collective appeared to slightly enhance the model’s robustness to MIAs in lower privacy regimes. Moreover, as privacy constraints tighten, the collective’s presence does not increase exposure to membership inference risk.

In summary, our key contributions are a theoretical characterization and an empirical validation of the limitations that differential privacy imposes on collective action, highlighting how the collective’s success depends on the model’s privacy parameters.

7.2 Limitations

While this study offers meaningful contributions, it is important to acknowledge its limitations, which in turn open up directions for future research. The collective action strategy implemented in our experiments involved modifying both features and labels. In real-world scenarios, modifying labels is much costlier for users than altering features. Our analysis assumes the collective has the power to manipulate both, which may not always hold. Our investigation into MIAs was limited in scope. We did not fully explore how different types and magnitudes of the collective’s signal (the feature $g(\cdot)$) might quantitatively affect the

success rate of MIAs. A deeper analysis is needed to better understand and generalize this finding.

Finally, while our empirical results broadly confirm the theoretical trends, certain deviations were observed. These deviations likely resulted from the complex interactions among high-dimensional data, gradient clipping, and the inherent stochasticity of DP-SGD. Such factors introduce noise that may obscure the theoretical relationships we have derived. Addressing this gap between theoretical intuition and empirical behavior will require deeper analytical tools and more controlled experimental designs.

7.3 Future Research Directions

This work opens several promising directions for future inquiry at the intersection of privacy, ACA, and responsible AI. The lower bounds derived in this thesis provide an initial characterization of the DP-ACA trade-off. Future theoretical work could focus on developing tighter bounds that more precisely capture the dynamics observed in empirical settings. DP-SGD’s behavior is heavily influenced by the per-sample gradient clipping threshold. Although recent research has moved toward automatic and adaptive clipping mechanisms, a strong theoretical and empirical characterization of how this clipping affects the success of collective action would be a valuable contribution.

Future research should extend our analysis to more diverse and realistic machine learning paradigms. Our experiments focused on classification tasks, yet many contemporary AI systems with large user bases, such as recommendation engines, ranking algorithms, and large language models (LLMs), operate in very different settings. Investigating how the presence of collective affects these systems would be essential for understanding their implications in real-world contexts that increasingly shape digital life. This is particularly relevant to modern development pipelines that rely on large-scale pre-training followed by task-specific fine-tuning. Future work should examine whether collective action is more effective during pre-training, when model representations are still forming, or during fine-tuning, when privacy constraints and data granularity differ. Understanding how privacy interacts with collective strategies at each stage could yield important insights into where and when collective action can meaningfully occur, or conversely, be constrained.

Finally, future research should seek to examine the interdependence among trustworthy AI principles in the presence of a collective. While fairness and collective action have been studied separately, it remains unclear how privacy, fairness, and transparency interact when collectives actively shape the learning process. For example, mechanisms that enhance

privacy may inadvertently limit fairness interventions, while fairness constraints can alter a system’s privacy dynamics by redistributing influence across subgroups. Understanding these links is crucial for designing AI systems that protect individuals while accounting for influence through collective action.

7.4 Concluding Remarks

This thesis has demonstrated that the technical choices in the design process of trustworthy AI systems are not purely objective. They carry social and ethical consequences that reshape the balance of power within data-driven systems. We have shown that Differential Privacy, while a gold standard for protecting individual data, can simultaneously function as an anti-cooperative mechanism, centralizing power within the firm by insulating its models from collective action. As organizations increasingly adopt privacy-preserving technologies to meet regulatory needs and build user trust, our findings make it evident that the goals of trustworthy AI can sometimes be in tension. The push for individual privacy, as implemented through DP, may inadvertently impede the viability of collective action, an essential tool for users and workers to assert their rights and contest algorithmic power. This work underscores the need for a more holistic approach to AI governance, one that carefully navigates the complex trade-offs between protecting individuals and empowering collectives in the design of fair and accountable systems.

References

- [1] 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*. ACM, Oct. 2016. doi: 10.1145/2976749.2978318. URL <http://dx.doi.org/10.1145/2976749.2978318>.
- [2] J. M. Abowd. The u.s. census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3226070. URL <https://doi.org/10.1145/3219819.3226070>.
- [3] A. F. Agarap. Deep learning using rectified linear units (relu), 2019. URL <https://arxiv.org/abs/1803.08375>.
- [4] U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.
- [5] A. Aktay, S. Bavadekar, G. Cossoul, J. Davis, D. Desfontaines, A. Fabrikant, E. Gabrilovich, K. Gadepalli, B. Gipson, M. Guevara, C. Kamath, M. Kansal, A. Lange, C. Mandayam, A. Oplinger, C. Pluntke, T. Roessler, A. Schlosberg, T. Shekel, S. Vispute, M. Vu, G. Wellenius, B. Williams, and R. J. Wilson. Google covid-19 community mobility reports: Anonymization process description (version 1.1). *arXiv preprint arXiv:2004.04145*, 2020. URL <https://arxiv.org/abs/2004.04145>. v4, submitted April 8, 2020; revised November 3, 2020.
- [6] F. Bach. *Learning Theory from First Principles*. MIT Press, Cambridge, MA, 2021. URL <https://mitpress.mit.edu/9780262046434/learning-theory-from-first-principles/>.

- [7] E. Bagdasaryan and V. Shmatikov. Differential privacy has disparate impact on model accuracy, 2019. URL <https://arxiv.org/abs/1905.12101>.
- [8] B. Balle and Y.-X. Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising, 2018. URL <https://arxiv.org/abs/1805.06530>.
- [9] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- [10] R. Bassily, A. Smith, and A. Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds, 2014. URL <https://arxiv.org/abs/1405.7085>.
- [11] O. Ben-Dov, J. Fawkes, S. Samadi, and A. Sanyal. The role of learning algorithms in collective action. In *International Conference on Machine Learning*, pages 3443–3461. PMLR, 2024.
- [12] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, and et al. On the opportunities and risks of foundation models, 2022. URL <https://arxiv.org/abs/2108.07258>.
- [13] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis. Automatic clipping: Differentially private deep learning made easier and stronger, 2023. URL <https://arxiv.org/abs/2206.07136>.
- [14] A. Calvi, G. Malgieri, and D. Kotzinos. The unfair side of privacy enhancing technologies: Addressing the trade-offs between pets and fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2047–2059, 2024.
- [15] N. Carlini, C. Liu, Úlfar Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks, 2019. URL <https://arxiv.org/abs/1802.08232>.
- [16] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.
- [17] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr. Poisoning web-scale training datasets is practical, 2024. URL <https://arxiv.org/abs/2302.10149>.

- [18] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.
- [19] S. Chen. The latest interface: Using data privacy as a sword and shield in antitrust litigation. *Hastings Law Journal*, 74(2):551–582, 2023. URL <https://www.hastingslawjournal.org/the-latest-interface-using-data-privacy-as-a-sword-and-shield-in-antitrust-litigation>. Published February 16, 2023.
- [20] E. Creager and R. Zemel. Online algorithmic recourse by collective action, 2023. URL <https://arxiv.org/abs/2401.00055>.
- [21] R. Cummings, D. Desfontaines, D. Evans, R. Geambasu, Y. Huang, M. Jagielski, P. Kairouz, G. Kamath, S. Oh, O. Ohrimenko, N. Papernot, R. Rogers, M. Shen, S. Song, W. Su, A. Terzis, A. Thakurta, S. Vassilvitskii, Y.-X. Wang, L. Xiong, S. Yekhanin, D. Yu, H. Zhang, and W. Zhang. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review*, 6(1), jan 16 2024. <https://hdsr.mitpress.mit.edu/pub/sl9we8gh>.
- [22] A. DeVrio, M. Eslami, and K. Holstein. Building, shifting, & employing power: A taxonomy of responses from below to algorithmic harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1093–1106, 2024.
- [23] A. Differential Privacy Team. Learning with privacy at scale. Technical report, Apple Inc., 2017. URL <https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf>. Accessed: 2025-10-24.
- [24] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-35908-1.
- [25] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay, editor, *Advances in Cryptology - EUROCRYPT 2006*, pages 486–503, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-34547-3.
- [26] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

- [27] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.
- [28] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL <https://data.europa.eu/eli/reg/2016/679/oj>.
- [29] J. Ferry, U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala. Sok: Taming the triangle – on the interplays between fairness, interpretability and privacy in machine learning, 2023. URL <https://arxiv.org/abs/2312.16191>.
- [30] F. Fioretto, C. Tran, P. Van Hentenryck, and K. Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. In L. D. Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5470–5477. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/766. URL <https://doi.org/10.24963/ijcai.2022/766>. Survey Track.
- [31] F. Fioretto, P. V. Hentenryck, and J. Ziani. Differential privacy overview and fundamental techniques, 2024. URL <https://arxiv.org/abs/2411.04710>.
- [32] F. Fioretto, C. Tran, and P. V. Hentenryck. Decision making with differential privacy under a fairness lens, 2024. URL <https://arxiv.org/abs/2105.07513>.
- [33] S. J. Gould. Differential privacy and collective bargaining over workplace data. *Italian Labour Law e-Journal*, 17(2):133–144, Jan. 2024. doi: 10.6092/issn.1561-8048/20838. URL <https://illeg.unibo.it/article/view/20838>.
- [34] Government of Canada. Personal information protection and electronic documents act. <https://laws-lois.justice.gc.ca/eng/acts/P-8.6/>, 2000. URL <https://laws-lois.justice.gc.ca/eng/acts/P-8.6/>. S.C. 2000, c. 5.
- [35] T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2019. URL <https://arxiv.org/abs/1708.06733>.
- [36] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

- [37] W. Guo, B. Tondi, and M. Barni. An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing*, 3: 261–287, 2022.
- [38] G. Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, Dec. 1968. doi: 10.1126/science.162.3859.1243. URL <http://www.jstor.org/stable/1724745>.
- [39] R. Hardin. *Collective action*. Rff Press, 1982.
- [40] M. Hardt, E. Mazumdar, C. Mendler-Dünnér, and T. Zrnic. Algorithmic collective action in machine learning. In *International Conference on Machine Learning*, pages 12570–12586. PMLR, 2023.
- [41] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [42] D. D. Heckathorn. The dynamics and dilemmas of collective action. *American Sociological Review*, 61(2):250–277, 1996. ISSN 00031224. URL <http://www.jstor.org/stable/2096334>.
- [43] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s): 1–37, 2022.
- [44] M. Jagielski, J. Ullman, and A. Oprea. Auditing differentially private machine learning: How private is private sgd?, 2020. URL <https://arxiv.org/abs/2006.07709>.
- [45] M. Jagielski, G. Severi, N. P. Harger, and A. Oprea. Subpopulation data poisoning attacks, 2021. URL <https://arxiv.org/abs/2006.14026>.
- [46] K. F. F. Jiechieu and N. Tsopze. Skills prediction based on multi-label resume classification using CNN with model predictions explanation. *Neural Computing and Applications*, 33(10):5069–5087, May 2021. ISSN 1433-3058. doi: 10.1007/s00521-020-05302-x. URL <https://doi.org/10.1007/s00521-020-05302-x>.
- [47] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.

- [48] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In S. Mannor, N. Srebro, and R. C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL <https://proceedings.mlr.press/v23/kifer12.html>.
- [49] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- [50] B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, and C. Troncoso. Disparate vulnerability to membership inference attacks, 2021. URL <https://arxiv.org/abs/1906.00389>.
- [51] A. Kurakin, S. Song, S. Chien, R. Geambasu, A. Terzis, and A. Thakurta. Toward training at imagenet scale with differential privacy, 2022. URL <https://arxiv.org/abs/2201.12328>.
- [52] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302 – 1338, 2000. doi: 10.1214/aos/1015957395. URL <https://doi.org/10.1214/aos/1015957395>.
- [53] Y. LeCun and C. Cortes. MNIST handwritten digit database. <http://yann.lecun.com/exdb/mnist/>, 2010. URL <http://yann.lecun.com/exdb/mnist/>.
- [54] J. Lin, L. Dang, M. Rahouti, and K. Xiong. Ml attack models: Adversarial attacks and data poisoning attacks, 2021. URL <https://arxiv.org/abs/2112.02797>.
- [55] R. V. Loo. Privacy pretexts. *Cornell Law Review*, 108(1):1–66, 2022. ISSN 0010-8847. URL https://scholarship.law.bu.edu/faculty_scholarship/2898/. Published online December 27, 2022.
- [56] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.
- [57] Z. Luo, D. J. Wu, E. Adeli, and L. Fei-Fei. Scalable differential privacy with sparse network finetuning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5057–5066, 2021. doi: 10.1109/CVPR46437.2021.00502.

- [58] Y. Ma, X. Zhu, and J. Hsu. Data poisoning against differentially-private learners: Attacks and defenses, 2019. URL <https://arxiv.org/abs/1903.09860>.
- [59] G. Marwell and P. Oliver. *The critical mass in collective action*. Cambridge University Press, 1993.
- [60] A. Melucci. *Challenging Codes: Collective Action in the Information Age*. Cambridge Cultural Social Studies. Cambridge University Press, 1996.
- [61] S. Milan. When algorithms shape collective action: Social media and the dynamics of cloud protesting. *Social Media + Society*, 1(2):2056305115622481, 2015. doi: 10.1177/2056305115622481. URL <https://doi.org/10.1177/2056305115622481>.
- [62] I. Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, page 263–275. IEEE, Aug. 2017. doi: 10.1109/csf.2017.11. URL <http://dx.doi.org/10.1109/CSF.2017.11>.
- [63] S. Moro, P. Cortez, and P. Rita. Bank Marketing. UCI Machine Learning Repository, 2014. DOI: <https://doi.org/10.24432/C5K306>.
- [64] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). NIST AI 100-1 NIST AI 100-1, NIST, Gaithersburg, MD, 2023. URL <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>.
- [65] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books. Stanford University Press, Stanford, CA, Nov. 2009. ISBN 9780804752374.
- [66] M. Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups, Second Printing with a New Preface and Appendix*. Harvard University Press, 1971. ISBN 9780674537507. URL <http://www.jstor.org/stable/j.ctvjjsf3ts>.
- [67] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge University Press, 1990.
- [68] E. Ostrom, J. Burger, C. B. Field, R. B. Norgaard, and D. Policansky. Revisiting the commons: Local lessons, global challenges. *Science*, 284(5412):278–282, 1999. doi: 10.1126/science.284.5412.278. URL <https://www.science.org/doi/abs/10.1126/science.284.5412.278>.

- [69] N. Papernot, S. Chien, S. Song, A. Thakurta, and U. Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy, 2020.
- [70] V. Papyan, X. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.
- [71] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 33–44, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372873. URL <https://doi.org/10.1145/3351095.3372873>.
- [72] F. Razmi, J. Lou, and L. Xiong. Does differential privacy prevent backdoor attacks in practice?, 2023. URL <https://arxiv.org/abs/2311.06227>.
- [73] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL <https://arxiv.org/abs/1910.01108>.
- [74] R. M. Shelby, S. Rismeni, K. Henne, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, J. Gallegos, A. Smart, E. Garcia, and G. Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.
- [75] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [76] D. Sigg, M. Hardt, and C. Mendler-Dünner. Decline now: A combinatorial model for algorithmic collective action, 2024. URL <https://arxiv.org/abs/2410.12633>.
- [77] State of California. California privacy rights act of 2020 (cpra). <https://oag.ca.gov/privacy/ccpa>, 2020. URL <https://oag.ca.gov/privacy/ccpa>. Proposition 24, approved November 3, 2020.
- [78] N. Tůmová. Data privacy: A handy shield against anti-competitive behaviour in eu digital markets? Research Paper 2024/II/2, Charles University in Prague Faculty of Law, July 2024. URL <https://ssrn.com/abstract=4910478>. Prague Law Working Papers Series.

- [79] Z. Tian, L. Cui, J. Liang, and S. Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.
- [80] F. Tramèr, R. Shokri, A. San Joaquin, H. Le, M. Jagielski, S. Hong, and N. Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2779–2792, 2022.
- [81] F. Tramèr, G. Kamath, and N. Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. In *International Conference on Machine Learning*, pages 48453–48467. PMLR, 2024.
- [82] U.S. Census Bureau. Census bureau sets key parameters to protect privacy in 2020 census results. Press release, June 2021. URL <https://www.census.gov/newsroom/press-releases/2021/2020-census-key-parameters.html>. Accessed: YYYY-MM-DD.
- [83] N. Vincent and B. Hecht. Can ”conscious data contribution” help users to exert ”data leverage” against technology companies? *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), Apr. 2021. doi: 10.1145/3449177. URL <https://doi.org/10.1145/3449177>.
- [84] N. Vincent, B. Hecht, and S. Sen. ”data strikes”: Evaluating the effectiveness of a new form of collective action against technology companies. In *The World Wide Web Conference, WWW ’19*, page 1931–1943, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313742. URL <https://doi.org/10.1145/3308558.3313742>.
- [85] N. Vincent, H. Li, N. Tilly, S. Chancellor, and B. J. Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2020.
- [86] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
- [87] R.-J. Yew, L. Qin, and S. Venkatasubramanian. You still see me: How data protection supports the architecture of ai surveillance, 2024. URL <https://arxiv.org/abs/2402.06609>.
- [88] S. Youn. Uber, lyft drivers coordinate to manipulate surge pricing at virginia airport over pay concerns, May 2019. URL <https://abc7news.com/>

[post/uber-lyft-drivers-coordinate-to-manipulate-surge-pricing-report/5306900/](https://www.abcnews.com/news/5306900/post/uber-lyft-drivers-coordinate-to-manipulate-surge-pricing-report/). ABC News report on coordinated driver log-offs at DCA.

- [89] D. Yu, G. Kamath, J. Kulkarni, T.-Y. Liu, J. Yin, and H. Zhang. Individual privacy accounting for differentially private stochastic gradient descent. *arXiv preprint arXiv:2206.02617*, 2022.
- [90] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization, 2017. URL <https://arxiv.org/abs/1611.03530>.
- [91] W. Zhang, O. Ohrimenko, and R. Cummings. Attribute privacy: Framework and mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 757–766, 2022.

APPENDICES

Appendix A

Additional proofs

A.1 Proof of Hardt et al. [40, Theorem 10]

Since Theorem 2.1.4 restates Hardt et al. [40, Theorem 10], we reproduce their proof here for completeness.

Proof. Fix a time step t and a model θ_t . Denote by $\mathcal{P}_t$ the gradient-redirecting distribution found at step t and let $\xi(\theta_t) = \frac{\|g_{\mathcal{P}'}(\theta_t) + \frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}(\theta_t)\|}{\|\theta_t - \theta^*\|}$. Then, the gradient-redirecting strategy induces the following gradient evaluated on $\mathcal{P}_t$:

$$\begin{aligned} g_{\mathcal{P}_t}(\theta_t) &= \alpha g_{\mathcal{P}'}(\theta_t) + (1 - \alpha) g_{\mathcal{P}_0}(\theta_t) \\ &= -\alpha \frac{1 - \alpha}{\alpha} g_{\mathcal{P}_0}(\theta_t) + \alpha \xi(\theta_t)(\theta_t - \theta^*) + (1 - \alpha) g_{\mathcal{P}_0}(\theta_t) \\ &= \alpha \xi(\theta_t)(\theta_t - \theta^*). \end{aligned}$$

Now let $c = \min_{\lambda \in [0,1]} \xi(\lambda \theta_0 + (1 - \lambda) \theta^*)$. Applying the strategy repeatedly across time steps yields

$$\begin{aligned} \|\theta_T - \theta^*\| &\leq \|\theta_{T-1} - \eta \alpha \xi(\theta_{T-1})(\theta_{T-1} - \theta^*) - \theta^*\| \\ &\leq (1 - \eta \alpha \xi(\theta_{T-1})) \|\theta_{T-1} - \theta^*\| \\ &\leq (1 - \eta \alpha c) \|\theta_{T-1} - \theta^*\| \\ &\leq (1 - \eta \alpha c)^T \|\theta_0 - \theta^*\|, \end{aligned}$$

which yields $S_T(\alpha) = -\|\theta_T - \theta^*\| \geq -(1 - \eta \alpha c)^T \|\theta_0 - \theta^*\|$. Setting $C(\alpha) = \alpha c$ concludes the proof. $\square$

A.2 Proof of Lemma 3.2.2

Proof. Since each $Y_i \sim \mathcal{N}(0, \sigma^2)$, we can write $Y_i = \sigma Z_i$, where $Z_i \sim \mathcal{N}(0, 1)$. Then,

$$S = \sqrt{\sum_{i=1}^d Y_i^2} = \sigma \sqrt{\sum_{i=1}^d Z_i^2} = \sigma \sqrt{U}$$

Let $U = \sum_{i=1}^d Z_i^2$. A standard tail bound for the chi-squared distribution (refer to the Corollary 1 in [52]) gives, for any $t > 0$,

$$\mathbb{P}\left(U \geq d + 2\sqrt{dt} + 2t\right) \leq e^{-t}$$

Substituting $t = \log(1/\delta)$ we can obtain, with probability at least $1 - \delta$,

$$\begin{aligned} U &\leq d + 2\sqrt{d \log(1/\delta)} + 2 \log(1/\delta) \\ &\leq d + 2\sqrt{2d \log(1/\delta)} + 2 \log(1/\delta) \\ &= \left(\sqrt{d} + \sqrt{2 \log(1/\delta)}\right)^2 \end{aligned}$$

Taking square root and multiplying by σ on both sides, we get the required bounds. $\square$

A.3 Theorem 3.2.1 in the No-Privacy Case

In this section, we show that, in the *no-privacy* regime, our main bound reduces to the bound stated by Hardt et al. [40, Theorem 10], which is restated in Theorem 2.1.4. We begin with the bound from our main Theorem 3.2.1:

$$S_T(\alpha, \sigma, C) \geq -(1 - \eta B_{\alpha, C})^T \|\theta_0 - \theta^*\| - \frac{\sigma C}{|\mathcal{B}|_{\min}^2} \Gamma_T \Delta_{d, \delta}. \quad (\text{A.1})$$

The term $B_{\alpha, C}$ is defined as follows from Equation (3.10):

$$B_{\alpha, C} = \alpha \xi_{\min}^c$$

where $\xi_{\min}^c = \min_{\lambda \in [0, 1]} \xi^c(\lambda \theta_0 + (1 - \lambda) \theta^*)$. To better understand the behavior of ξ^c , we focus on its value at a particular point θ_t along the trajectory, as defined in Equation (3.9):

$$\xi^c(\theta_t) = \frac{\left\| \bar{g}_{\mathcal{B}'_t}^{\text{clip}}(\theta_t) + \frac{1-\alpha}{\alpha} \bar{g}_{\mathcal{B}_0}^{\text{clip}}(\theta_t) \right\|}{\|\theta_t - \theta^*\|}. \quad (\text{A.2})$$

No-privacy specialization ($\sigma = 0$ and $C \rightarrow \infty$). Setting $\sigma = 0$ eliminates the second term from Equation (A.1), giving

$$S_T(\alpha, C) \geq -(1 - \eta B_{\alpha, C})^T \|\theta_0 - \theta^*\| \quad (\text{A.3})$$

By the definition of clipping operator, as the threshold $C \rightarrow \infty$, the clipping becomes inactive. Using this fact and considering expectation of unclipped gradient, Equation (A.2) can be written as:

$$\xi(\theta_t) = \frac{\|g_{\mathcal{P}'}(\theta_t) + \frac{1-\alpha}{\alpha} g_{\mathcal{P}_0}(\theta_t)\|}{\|\theta_t - \theta^*\|}. \quad (\text{A.4})$$

This definition of ξ is identical to original Equation (2.9). Hence, when $\sigma = 0$ and $C \rightarrow \infty$,

$$\begin{aligned} B_{\alpha, C} &= \alpha \xi_{\min}^c \\ &= \alpha \xi_{\min} \\ &= C(\alpha) \quad (\text{from Theorem 2.1.4}) \end{aligned} \quad (\text{A.5})$$

Finally, the Equation (A.3) reduces to:

$$S_T(\alpha) \geq -(1 - \eta C(\alpha))^T \|\theta_0 - \theta^*\| \quad (\text{A.6})$$

which matches the original bound of Hardt et al. [40].

Appendix B

Additional Results

B.1 Baseline accuracies

Table B.1 presents the baseline predictive accuracies of DP-trained classifiers across all datasets.

Dataset	CIFAR-10			MNIST			Resume			Bank Marketing		
C	0.44	3.97	21.83	0.08	0.63	3.39	0.84	1.80	3.44	0.92	2.51	4.94
$\varepsilon = 15.0$	73%	61%	48%	91%	97%	97%	86%	86%	86%	83%	82%	82%
$\varepsilon = 5.0$	71%	57%	44%	90%	97%	97%	86%	86%	85%	83%	82%	82%
$\varepsilon = 2.5$	68%	49%	40%	88%	96%	96%	86%	86%	85%	83%	83%	81%
$\varepsilon = 1.5$	63%	47%	34%	85%	95%	96%	86%	86%	84%	83%	82%	80%
$\varepsilon = 1.25$	60%	46%	33%	83%	95%	95%	86%	86%	84%	83%	81%	79%
$\varepsilon = 1.0$	57%	44%	29%	81%	95%	95%	86%	86%	84%	83%	80%	74%

Table B.1: Test set accuracies under different privacy losses (ε) for models trained on four datasets: CIFAR-10, MNIST, Resume, and Bank Marketing. The clipping thresholds were selected based on the 25th, 50th, and 75th percentiles of per-sample gradient norms computed over the first few epochs of non-private training [1]. For CIFAR-10, MNIST, and Bank Marketing, the reported metric is classification accuracy, while for Resume, the reported accuracy corresponds to $1 - \text{Hamming loss}$ for the multi-label classification setting. Non-private baselines ($\varepsilon = \infty$, $\sigma = 0$, $C = \infty$) achieve accuracies of 86%, 100%, 96%, and 82%, respectively.

B.2 Critical mass

Figure B.1 shows the results for automatic clipping on the MNIST and CIFAR-10 datasets.

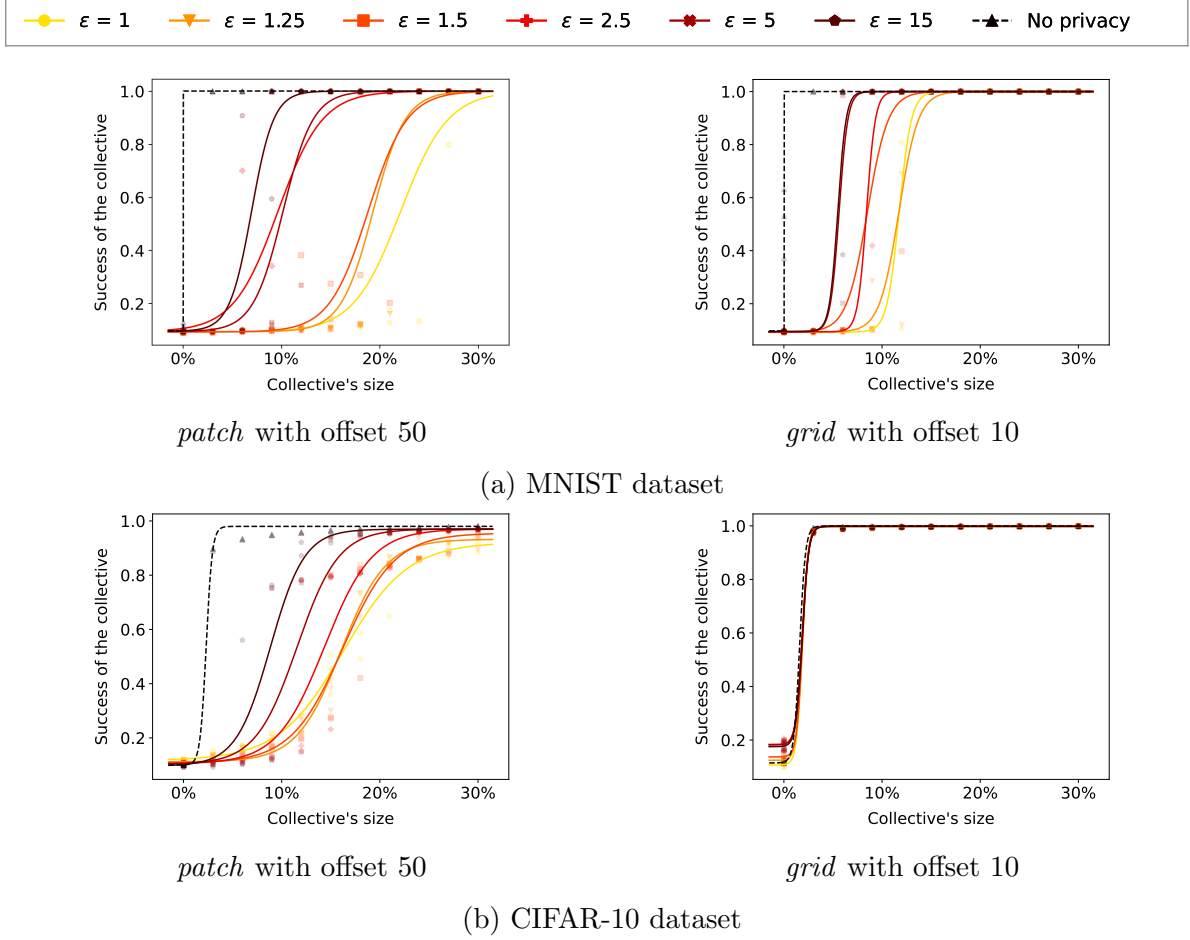


Figure B.1: Success of the collective with automatic clipping [13]. Each row corresponds to a different datasets (MNIST and CIFAR-10), and each column to a different collective's strategy.

B.3 LiRA results

Figure B.2 and Table B.2 extend our results from section 4.3 to include MIA results on the SVHN dataset. The SVHN dataset consists of real-world images of house numbers obtained from Google Street View, commonly used for evaluating visual recognition models on digit classification tasks.

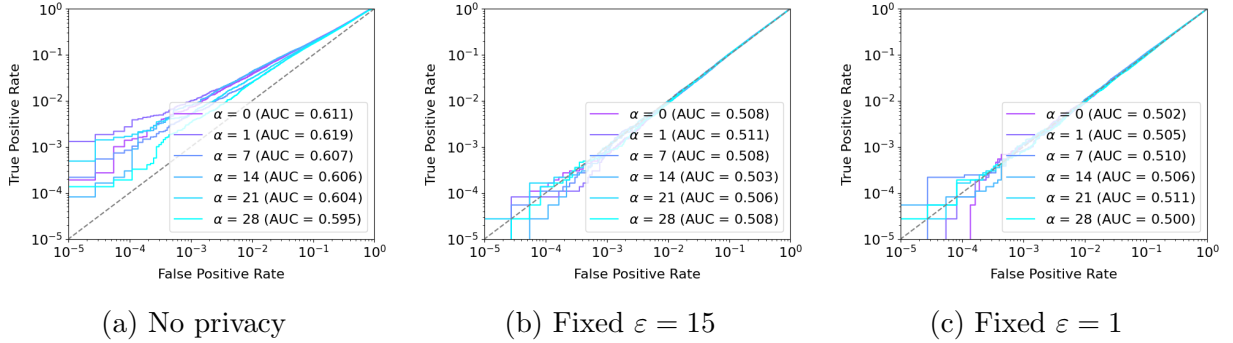


Figure B.2: Effectiveness of Likelihood Ratio Attack (LiRA) [16] evaluation on the SVHN dataset. Each figure corresponds to a different setting of privacy constraints with privacy increasing from left to right. In the left panel, we set $\varepsilon = \infty$ and $C = \infty$, corresponding to a setting with no privacy.

Collective's size α	No privacy		$\varepsilon = 15$		$\varepsilon = 1$	
	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC	TPR @ 0.1% FPR	AUC
0%	0.73%	61.1%	0.068%	50.8%	0.114%	50.2%
1%	1.00%	61.9%	0.076%	51.1%	0.114%	50.5%
7%	0.50%	60.7%	0.087%	50.8%	0.084%	51.0%
14%	0.79%	60.6%	0.092%	50.3%	0.111%	50.6%
21%	0.52%	60.4%	0.100%	50.6%	0.122%	51.1%
28%	0.33%	59.5%	0.076%	50.8%	0.079%	50.0%

Table B.2: Evaluation of LiRA under varying privacy constraints using AUC and TPR at 0.1% FPR on SVHN dataset. Lower TPR @ 0.1% FPR indicates better robustness to MIA, while with AUC the desired metric is as close to 50% as possible.

B.4 Visualization

Figure B.3 shows the empirical distributions of per-sample gradient norms used to select the clipping thresholds for each dataset.

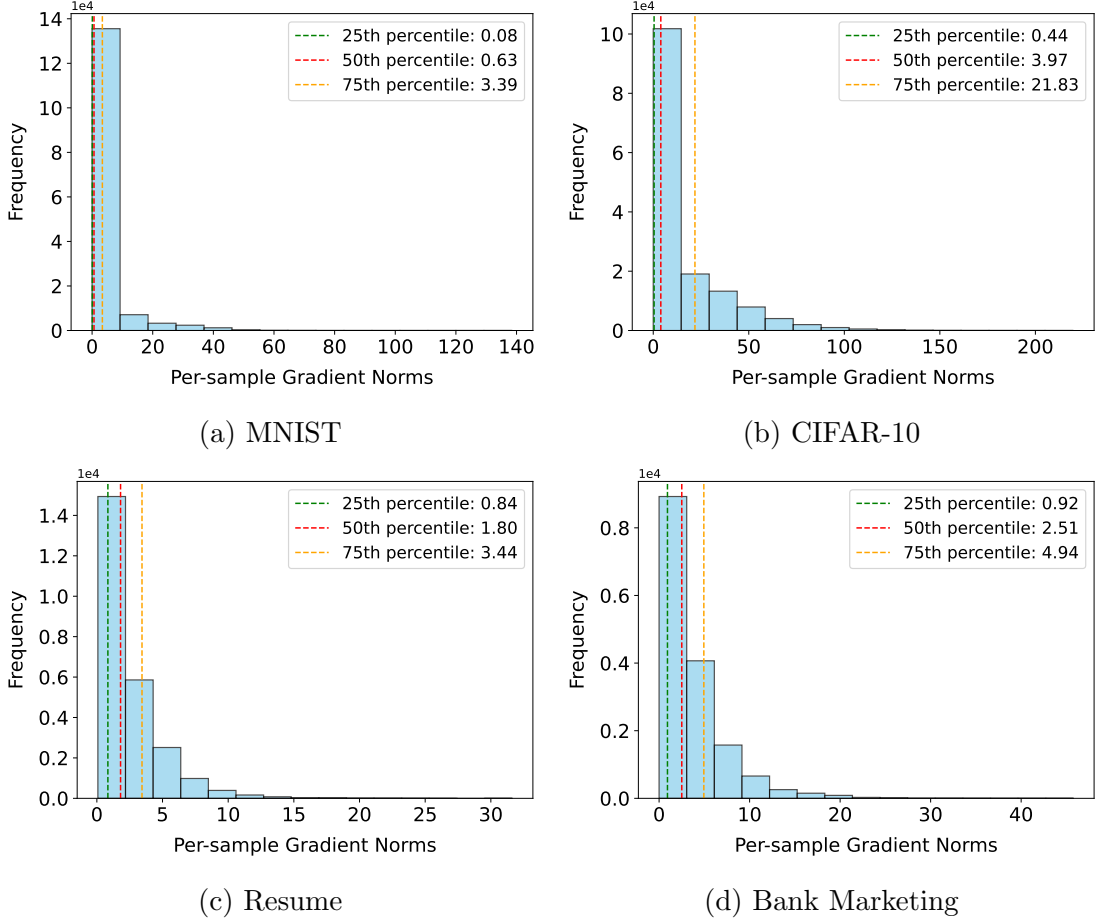


Figure B.3: Histograms of per-sample gradient norms across all datasets: MNIST, CIFAR-10, Resume, and Bank Marketing dataset. Sample gradients were collected during the first few training epochs without applying DP constraints.