Design and Applications of Differentially Private Mechanisms:
Adherence to Query Range Constraints and Obfuscation of Facial Images

by

William Lee Croft

A thesis submitted to the Faculty of Graduate and Postdoctoral
Affairs in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

Carleton University
Ottawa, Ontario

© 2020

William Lee Croft

Abstract

Collection and dissemination of data are common tasks motivated by a number of benefits attained through the analysis of rich datasets. Yet many datasets contain sensitive information about individuals which must be duly protected if the data is to be used or shared. Differential privacy is a commonly used disclosure control method for protecting sensitive information while allowing for queries to be posed on databases. The driving idea behind differential privacy is to use a randomization mechanism to add controlled noise to query responses. This is done in order to provide a guarantee on the distinguishability between potential configurations of the underlying sensitive data which gave rise to the noisy mechanism output. In this thesis, we focus on two major topics relating to the design and application of differentially private mechanisms. In the first, we focus on the design of mechanisms which employ a range of noisy responses matched to the range of the query posed on the database. Adherence to the range of the query offers the potential for improved utility in a mechanism, yet attaining improved utility in a manner which preserves the differential privacy guarantee is not straightforward. We propose two different approaches to the design of range-adherent mechanisms, one of which is based on the use of a truncated and normalized Laplace distribution, while the other employs linear programming. The second major topic of the thesis covers the application of differential privacy to the task of obfuscating facial identity in images. Often, depiction of identity in images may be seen as a breach of privacy, yet the preservation of other information in the images may be desirable. We propose a framework for the application of a distance-based generalization of differential privacy via generative models for images. We provide details on the configuration of a mechanism to achieve a differentially private guarantee in this setting and show how to achieve photo-realistic obfuscated images through the use of various generative models. Within both topics of our work, we implement our proposed approaches and analyze the results of experimental comparisons between our work and other relevant approaches from the literature. We demonstrate improvements in the utility of obfuscated data achieved via our methods.

Acknowledgments

I would like to thank my supervisors, Professor Jörg-Rüdiger Sack and Professor Wei Shi for their valuable support both within our research and in administrative matters. From guidance in the development of ideas for new directions in our work to the editing of papers detailing our results, their assistance has always been an invaluable asset. Their suggestions and feedback throughout this work have been greatly appreciated. I would also like to thank Professor Kurt Mehlhorn for a valuable discussion on differential privacy, Professor Hans-Georg Zimmermann and Professor Tony White for informative discussions on neural networks, Kyle Harvey for his time spent elucidating for us some of the intricacies of the Maple analysis software, and Professor Matthias Neufang for his advice and insights on the analysis of complex exponential expressions. I am very grateful for the time and advice that they have all given me in these consultations.

Contents

1	Introduction	4
1.1	Background	4
1.2	Motivation	5
2	Literature Review	8
2.1	Differential Privacy	8
2.1.1	Syntactic Anonymity	9
2.1.2	Overview of Differential Privacy	10
2.1.3	Mechanisms and Measures of Utility	11
2.1.4	Constraint Adherence	12
2.1.5	Specialized Topics	13
2.2	Facial Obfuscation	14
2.2.1	Formalized Approaches to Facial Obfuscation	15
2.2.2	Neural Networks	16
2.2.3	Facial Obfuscation via Generative Neural Networks	16
3	Thesis Overview and Contributions	18
4	Notation and Definitions	21
4.1	Absolute Value and Set Cardinality Notation	21
4.2	The Differential Privacy Guarantee	21
4.3	Query Sensitivity	22
4.4	The Laplace Mechanism	22
4.5	The Exponential Mechanism	22
4.6	Boundary-Snapping	23
5	Constraint-Aware Mechanisms	24
5.1	Normalized Laplace Mechanisms	25
5.1.1	Preliminaries	26
5.1.2	Single Infinite Constraint	34
5.1.3	Arbitrary Finite Constraints	38
5.1.4	Arbitrary Constraints	42
5.1.5	Summary of Results	44
5.2	Linear Programming Mechanisms	44
5.2.1	Design Considerations	44
5.2.2	Formalization of Range-Adherent Mechanisms	45
5.2.3	Boundary-Snapping	48

5.2.4	Linear Programming	50
5.3	Experimental Comparisons	53
5.3.1	Normalized Laplace Experiments	53
5.3.2	Linear Programming Experiments	57
5.4	Future Work	69
5.5	Summary	69
6	Differentially Private Obfuscation of Images	71
6.1	Notation and Definitions	72
6.1.1	Distance-Generalized Differential Privacy	73
6.1.2	AttGAN	73
6.1.3	Principal Component Analysis	75
6.2	Weaknesses in Existing Facial Obfuscation	76
6.2.1	Non-Formalized Approaches	76
6.2.2	The k-same Privacy Guarantee	77
6.2.3	Other Formalized Alternatives	79
6.3	Differential Privacy for Generative Models	80
6.3.1	Framework for Generative Models	80
6.3.2	Benefits of Differentially Private Facial Obfuscation	82
6.3.3	Implementation Details	83
6.3.4	Obfuscation of Unseen Images	83
6.4	Differentially Private Obfuscation Via a GAN	85
6.4.1	Image Encoding and Model Architecture	85
6.4.2	Loss Calculations for Noisy Encodings	88
6.4.3	Pose Preservation	88
6.4.4	Objective Function	89
6.4.5	Image Background	90
6.5	Obfuscation Within a PCA Basis	91
6.5.1	Proposed Usage of PCA	91
6.5.2	Privacy Budget Allocation	92
6.5.3	Model Training	94
6.6	Privacy Interpretation	94
6.6.1	Differential Privacy Guarantee Interpretation	94
6.6.2	Contextual Information in Images	96
6.6.3	Information Leakage via Auxiliary Model Inputs	96
6.7	Differential Privacy in Pixel-Space	97
6.8	Experimental Comparisons	98
6.8.1	Comparisons Using AAM and GGN-Based Methods of Obfuscation	99
6.8.2	Comparisons Using GAN-Based Methods of Obfuscation	108
6.8.3	Privacy-Utility Trade-offs in Pixel-Space	114
6.9	Future Work	117
6.10	Summary	117
7	Conclusions	119

Chapter 1

Introduction

1.1 Background

The collection, dissemination and usage of a wide variety of data are indispensable activities within both the public and the private sector. The availability of data has increased greatly over recent decades, coming from a broad range of sources including government programs such as surveys and censuses [34], medical records from hospitals and clinics [201], personal information from social networks [61], and usage data from smart devices [200]. Along with the rise in the availability of data has come increased attention in studying methods to exploit the potential value of data [209]. Effective usage of data is beneficial for a broad range of applications including governmental decision-making [14], public and private sector research [128], and corporate decision-making [177].

Public sector information refers to data that is collected by a government. This covers a wide range including agricultural, geographic, environmental, economic, social and scientific data [34]. Despite the large amount of data collected by governments, it has traditionally been kept solely for internal usage. However, there have been increasing demands to allow for open access to public sector information and some countries have shifted towards policies that make this data available to the general public [22, 143]. Allowing for open access to public sector information can be highly beneficial to private companies and researchers for whom this type of valuable data may otherwise be much more difficult or costly to obtain. This is desirable from an economic point of view as it promotes growth in the private sector [199]. The usage of this data by a broader audience is in turn advantageous to the the general population and the government as they benefit from the growth of local companies and the dissemination of knowledge and results from new research [63, 163, 195]. Beyond this, the sharing of data between different countries can produce similar benefits [14].

To protect privacy, released information must be devoid of directly identifiable information about individuals [104]. However, it may still be tabulated in a *microdata* format [22] (at the granularity of individuals). The increase in the technological ability to gather and compile this type of individual-level data over the years has led to richer and more accessible sources of microdata gathered by organizations related to financial services, education, health care, law enforcement, retail and number of others [177]. Even social networks are now seen as valuable source of information about individuals [61]. Some uses of this type of microdata include enabling better research in the socio-economic domain and informing policy-making decisions both for governments and private companies [128].

While the individual-level data of interest to most companies relates to information about lifestyles, purchases and social and economic status, there is also a high demand in other areas for health and medical-related data. Readily available health data can greatly impact research by

reducing the burden of initially collecting data for studies and improving the quality of subsequent work by enriching collections of data that are available [98]. Having access to larger databases with greater numbers of attributes allows for more meaningful studies to be conducted by taking into consideration greater numbers of factors [158]. Genome-based data can be particularly useful for studies involving genetic diseases [21, 52]. There also exist initiatives to combine health records from different institutions in order to create data that gives a more complete and informative view of individuals' health [201]. The creation of centralized repositories for individual-level health data is seen as an approach to provide a patient-centric model of health care referred to as personalized medicine [9, 81, 132].

Another valuable source of data comes in the form of *visual data* such as photos and video. This type of data is becoming available in staggering quantities due to the rapidly increasing presence of data capturing devices, e.g., cameras and closed-circuit television systems [26], visual sensor networks [202] and a host of other devices and methods [167]. Systems that employ these devices have many benefits including mitigation of crime [26, 202], improved care in assisted-living [160], and useful services such as Google Street View [80].

Further to this, the majority of teenagers and adults in countries with advanced economies own smartphones [191] and many of them actively share images on social media platforms [50]. Facebook now has over 2 billion users [2] with more than 350 million image uploads per day [3] and Instagram has over 1 billion users [4] with more than 100 million images and videos uploaded per day [5]. There are various social motivations, such as the desire to connect with others and share experiences, that encourage users to consistently share images on these platforms [156].

Despite the broad spectrum of benefits from the collection and sharing of data, it is essential to protect the privacy of the individuals to whom the data pertains. Many forms of information such as financial or health data may be considered highly personal and thus should be appropriately protected. Similarly, the depiction of a human face in visual data provides strong identifier of who the individual is [160, 167] and the potential for infringement on personal privacy must be taken seriously. To this regard, the privacy-related implications of data sharing has become a major field of study.

1.2 Motivation

When collecting and releasing individual-level data of a sensitive nature, it is necessary to protect the privacy of the individuals to whom the data pertains. This general idea can be stated as the goal of preventing the disclosure of *sensitive information* about specific individuals where sensitive information refers to any information that the individuals might not wish to share with the recipient(s) of the released data. To achieve this, a process of *data sanitization* is required. In such a process, any directly identifying information, such as names and identification numbers, must be stripped from any data that is to be released [51]. This alone, however, is insufficient as *re-identification* (the re-association of identities with their corresponding sensitive values) may be achieved by cross-referencing the remaining attributes with other data sets such as voter registries [189]. Yet, if any useful data is to be released, the risk of disclosure cannot be nullified [44, 204]. The concept of disclosure risk has long been studied in the context of analysis on statistical databases [40, 41, 66]. Even if a database itself is not released, simply allowing for queries to be posed on the database, presents a similar risk of disclosure [7].

With the ubiquity of computers and the increase in their processing power, re-identification of records and the extraction of sensitive information has become an easier task for a wider audience. Developments in the computational extraction of previously unknown patterns in databases, known

as *data mining*, has had a large impact on the potential for drawing new inferences from individual-level databases [181]. While this can be highly beneficial for many research applications, it also greatly increases the risk of disclosure of sensitive information [62]. As a result, a field of *privacy-preserving data mining* has emerged with the intent of developing methods of data mining that mitigate the risk of disclosure [209]. Regardless of how data is released, it is essential to provide evidence that the level of protection against disclosure is demonstrably sound in some way, generally by showing that some limitation is imposed on the risk of disclosure [203].

Issues of disclosure risk are particularly important in the context of health care data. Due to the highly sensitive nature of this type of data, special interest is often afforded. For example, the Health Insurance Portability and Accountability Act (HIPAA) provides legislation in the United States on specific requirements to be met for the sanitization and release of health care data [196]. Despite the strong requirements put in place by HIPAA, re-identification risk remains an issue on databases that it deems safe [19]. Studies have been performed to analyze the levels of risk that exist within different models of release for adverse drug event records [59] and hospital prescription records [60]. For data of such a sensitive nature, general policies remain insufficient for effective data sanitization [170].

Sensitive information also often resides within visual media such as images and video. In cases where potentially sensitive visual data is disseminated to the public, it is imperative to hide the identities of individuals depicted in the data. For example, with the large quantities of images published by Google Street View, face blurring is applied to hide identities [71]. Failure to sufficiently protect privacy may allow undesirable inferences to be drawn about individuals or enable malicious activities such as voyeurism or stalking. In the context of online image platforms and social media, users must constantly weigh the benefits of sharing images against the risks to privacy [105]. It is clear that users want to be able to control how, and with whom, their images are shared [100, 133]. Furthermore, individuals appearing in images taken by others often want some say in what happens to the images. In cases involving multiple parties, conflicts are often resolved either through unaltered release of the image or complete redaction [186]. Yet this type of all or nothing approach is not ideal for handling nuanced cases and is likely unsatisfactory for some of the parties. It is also not uncommon for individuals to be captured accidentally or without authorization in the backgrounds of images, making the issue of privacy more complex.

Beyond privacy concerns regarding images being seen by others, rapid advances in machine learning approaches to computer vision also pose threats to privacy. Detection and identification of faces in images are now tasks that can be accomplished with great accuracy. Recent neural networks have achieved over 99% accuracy in tasks of facial verification and classification [161, 174] and are increasingly able to handle difficult conditions such as low resolution, non-frontal poses and poor illumination [1, 117]. These advances enable the development of software that can readily track and profile mass populations. This may, for example, raise concerns over the ability of governments and authorities to track populations on a highly personal level [96]. In some areas, closed-circuit cameras are present in the millions, capturing vast quantities of visual data [26]. While these technologies no doubt have benefits such as improved quality of services [110, 160] and prevention of crime [160, 198], these benefits must be carefully weighed against the infringement on personal privacy. Furthermore, as with other types of sensitive information about individuals, the collection of visual data must be conducted in a manner that is compliant with privacy-centric laws such as GDPR [17, 97].

Although effective methods of data sanitization are necessary to protect the privacy of individuals, this requirement is at odds with the notion of releasing databases with greater levels of detail. Methods of data sanitization induce modifications on the data that negatively impact the utility of the data [10]. As such, there exists a trade-off between privacy protection and utility. Examples of this detrimental effect on utility can be seen in the impact that HIPAA has had on health and

medical research [111, 144, 154, 157]. There has been much attention given to the study of methods of data sanitization that aim to protect privacy while maintaining a high level of utility [72]. Similarly, it is difficult to strike a good balance between privacy and information sufficiency when applying methods of protection such as blurring and pixelization to visual data [92, 121]. With the continued evolution of data collection and dissemination, the identification of new disclosure risks, and the issue of levels of data utility that remain not fully satisfactory, privacy preserving data publishing remains a very active field of research.

Chapter 2

Literature Review

In this chapter, we provide a review of the material related to our work. All of the work throughout this thesis is centered around the study and application of differential privacy. We review this topic in Section 2.1, both at a high level and at finer granularity within the subtopics of greatest relevance to our work. In Section 2.2, we then turn to the topic of facial obfuscation. There, we cover literature on the theoretical approaches to privacy in facial images as well as the machine learning tools that are employed to realize the privacy guarantees.

2.1 Differential Privacy

Within the context of privacy-preserving data analysis, the main focus is to derive useful results and models from sensitive data without having full access to the original, unaltered data. The nature of the limitations enforced on access to the data characterizes the extent of the privacy that is achieved. This area of study has a long history as can be witnessed by the many different research directions that have been investigated. For a survey of early such works, see [7]. In the current state of the field, methods can be roughly categorized as being either syntactic or semantic in nature. *Syntactic privacy* requires the alteration of data prior to its release in order to adhere to syntactic conditions on the state of the released data such as limited distinguishability between the records of a sanitized database (see e.g., [131, 150, 190]). The syntactic conditions apply strictly to the released data and are therefore oblivious to the potential for disclosure of sensitive information when the released data is combined with other unmodeled sources of data. Due to this, a release offers only protection against specific attack models that are thwarted by the chosen syntactic properties of the data; there is no guaranteed protection against other unmodeled attacks.

Semantic privacy takes a different approach, aiming instead to enforce a guarantee limiting what can be inferred about the original data when presented with the altered release, regardless of the attack model and prior knowledge of the attacker. However, it has been recognized that it is neither practical nor possible in most cases to offer utility in the released data while protecting against all possible attack models [43, 55]. An attacker with particularly detailed background knowledge may be able to accurately infer sensitive information when presented with a seemingly harmless release. Therefore, most semantic approaches tend to rely on stochastic operations such as the controlled *perturbation* of data in order to limit the accuracy with which inferences can be drawn about the original data [29, 62]. In this context, perturbation refers to the addition of noise to the data in order to hide sensitive information. Thus, rather than making an assertion regarding protection against specific attacks, a semantic privacy guarantee tends to characterize how accurately the released

data reflects the original data. One can subsequently analyze such a guarantee in the context of specific attack models to perform more targeted forms of risk analysis if necessary. The framework of differential privacy has quickly become one of the most widely-studied approaches to semantic privacy and is often favoured over syntactic approaches by virtue of a robust and provable privacy guarantee [45].

We begin this section with a review of syntactic privacy in order to provide a broader view of the background in the field of privacy-preserving data analysis. We then turn our focus to the topic of differential privacy, giving first an overview of the general framework. We follow this with a review of mechanisms that have been proposed to achieve differential privacy and measures of utility related to these mechanisms. We then cover work on the design of constraint adherence in differential privacy. We close the section with a brief review of some specialized topics in the field of differential privacy that are relevant to our work.

2.1.1 Syntactic Anonymity

Syntactic anonymity refers to a method of publishing sensitive databases by producing a sanitized database that satisfies certain required properties to protect against specific types of attacks [33]. The required database properties, and thus the attacks against which protection is achieved, are determined by the model of syntactic anonymity that is chosen. Generally, these models deal with properties of *equivalence classes* of database records. Equivalence classes are groups of records that possess identical values across their *quasi-identifiers*. Quasi-identifiers are attributes that do not directly identify records but can still be linked to other databases to assist in re-identification [190]. These are generally demographic type attributes such as age and gender. Thus, records that share the same values across all of their quasi-identifiers are members of the same equivalence class. In order to prevent certain types of unwanted inferences from being drawn, equivalence classes are required to satisfy different properties depending on which model of syntactic anonymity is being applied. Typically, operations of *generalization* (the reduction of data precision) and *suppression* (the deletion of data) are applied to the database to achieve the desired properties. Such operations preserve the truthfulness of the data as they do not introduce any changes that are inconsistent with the original data [171].

One of the most well-known models of syntactic anonymity is *k-anonymity* [190]. In this model, each equivalence class is required to have a cardinality of at least k , where k is a user-specified parameter. The basic idea is that the indistinguishability of records within the same equivalence class induces a degree of uncertainty in any attempts made at cross-referencing the records with another database. When attempting to match an identity to a record, it becomes necessary for the attacker to guess which record from the equivalence class matches the chosen identity. When selecting at random, the chance of a correct association is thus upper-bounded by $\frac{1}{k}$.

A common assumption in syntactic models of privacy is that the attacker initially has no knowledge about any of the sensitive information. In practice, however, attackers may have already gained knowledge about sensitive information related to some of the population. Use of this additional information can lead to inferences with a probability of success that exceeds the worst-case bounds asserted by the parameters of the syntactic models [18]. Furthermore, syntactic anonymity is susceptible to both *composition attacks* which use anonymized data released from multiple sources in combination [73] and to inferences that draw on general knowledge about trends in sensitive data such as rates of incidence among demographic groups [205].

While stronger models of syntactic anonymity (e.g., *l-diversity* [131] and *t-closeness* [150]) ameliorate the situation with respect to the protection of anonymity, they greatly increase the level of information loss and remain susceptible to some forms of attacks [45]. Furthermore, as the number

of attributes to be release in a database increases, the number of equivalence classes induced by the partitioning over these attributes grows very quickly. This has the effect of greatly increasing the level of information loss required to achieve the syntactic conditions necessary for the release of an anonymized database [8]. For this reason, syntactic anonymity is seen as inappropriate for releases involving large numbers of attributes.

2.1.2 Overview of Differential Privacy

Privacy via data perturbation can be achieved either through direct perturbation of database records or through perturbation of responses to queries posed on an unaltered version of the database [7]. The later of the two options has been widely studied as a method to achieve semantic privacy. A key challenge with this approach is to determine what is an appropriate amount of perturbation to apply to the query responses. An important result in this area showed that for subset sum queries, it is possible to calculate a required magnitude of perturbation such that any perturbation beneath this level would result in a privacy violation [43]. More specifically, insufficient perturbation was shown to allow an attacker to reconstruct the original database with a high degree of accuracy using a polynomial number of queries.

The notion of calculating an appropriate amount of noise to add to query responses was later extended to be applicable for general query functions [55]. The key idea was that the standard deviation of the required amount of noise could be calculated using a measure of *sensitivity* of the query function. The sensitivity was defined in this context to be the maximum amount the unperturbed query response could differ between two *adjacent databases* (adjacency meaning that they differ by a single row). This was applied to develop the notion of ϵ -*indistinguishability* which is the requirement that the respective probabilities for two adjacent databases to produce the same perturbed query response may differ by at most a multiplicative factor of $1 + \epsilon$, where ϵ is a user-selected *privacy parameter*. The reasoning behind this requirement was that a greater degree of indistinguishability between potential configurations of the underlying database induces greater uncertainty in an attacker who is trying to determine what the true configuration is.

Further development and formalization of the concept of perturbing the responses of general queries led to the creation of the framework that is now known as *differential privacy* [53]. The basic idea in differential privacy is to add noise to query responses using a *randomization mechanism* configured according to the query sensitivity and the user-selected privacy parameter ϵ . A randomization mechanism is a stochastic function used to perturb its input value. Within the context of differential privacy, a mechanism is said to be ϵ -differentially private if, when applied to the query responses of any pair of adjacent databases, the difference between the probabilities of the two databases producing the same noisy response is bounded by a factor of e^ϵ . As such, smaller values of ϵ imply a stronger level of privacy. This condition is referred to as the *privacy guarantee*. For a formalized definition of the framework, see Section 4.2.

The noisy responses produced by a differentially private mechanism are immune to *post-processing* [56]. This means that without gaining access to the original sensitive data itself, there exists no function an attacker could apply to the noisy response to render it less differentially private. Furthermore, the privacy guarantee composes linearly across repeated uses of the mechanism. The *composition theorem* states that the combination of the responses from the independent usage of an ϵ_1 -differentially private mechanism and an ϵ_2 -differentially private mechanism is $(\epsilon_1 + \epsilon_2)$ -differentially private [56]. This shows the degradation of privacy over repeated uses of differentially private mechanisms. To guarantee privacy over multiple uses, it is therefore necessary to enforce an upper bound on the composition of ϵ . A total value of ϵ deemed acceptable to be spent over any number of queries is referred to as a *privacy budget*.

There are two main settings in which differential privacy can be applied - an *interactive setting* and a *non-interactive setting* [56]. In the interactive setting, users pose queries one at a time and receive a noisy response after each query. In the non-interactive setting, some form of task is performed using a batch of queries known in advance. Knowledge of the query workload in the non-interactive setting allows for more sophisticated mechanisms and steps of post-processing that take advantage of relationships between the queries to be developed. The non-interactive setting covers goals such as answering batches of queries [116], producing sanitized *marginals* [16] (i.e., projections of record counts over subsets of the database attributes) as well as more complex learning tasks including classification and clustering [149]. While the non-interactive setting can offer improved accuracy over interactively posing each query individually, it is also more restrictive in terms of what the user is able to do.

Since its inception, there has been a great deal of work done on differential privacy in both settings with a wide range of new developments and applications [54, 149, 210]. Relaxations of the privacy guarantee have also been explored as a means to improve utility. For example, (ϵ, δ) -differential privacy includes an additive constant δ in the allowable difference between probabilities considered in the privacy guarantee [56]. In our work, we do not deal with such relaxations of the differential privacy guarantee. Furthermore, we primarily focus this review on developments in interactive differential privacy as this is the setting within which our work falls.

2.1.3 Mechanisms and Measures of Utility

One of the most well-known differentially private mechanisms is the *Laplace mechanism* [55]. The mechanism is implemented using a Laplace distribution parameterized with the true query response as its location parameter and the query sensitivity over ϵ as its scaling parameter. Although this mechanism is widely used due to its applicability to a broad range of tasks, there has also been a great deal of research into the development of mechanisms that are better suited to specific goals. Most commonly, the intent in such research is to improve upon the utility of the noisy query responses. Utility has often been considered in terms of *distortion*, measured in terms of expected distance between true query responses and noisy query responses. Reduction of distortion has been a common goal in the design of mechanisms [30, 87] and minimization of distortion has also been used as a criterion for optimality [184].

For *counting queries* (the query of how many records in a database satisfy a particular predicate), the *geometric mechanism* has been proposed [77]. The geometric mechanism is a discrete variant of the Laplace mechanism and has been proven to be *universally optimal* for counting queries. Universal optimality refers to the fact that the mechanism is simultaneously optimal in utility for every user regardless of the prior knowledge they possess about the database. The users are considered to be *Bayesian users*, meaning when given a noisy query response, they will apply *Bayesian post-processing* (described in [77]) to map the response to another response (which may or may not be the same) in order to minimize their expected loss. The remapping is guided by user-specific loss function which is monotonically increasing as the distance between the remapped guess and the true query response increases. For any rational user who applies Bayesian post-processing to minimize their expected loss over all possible true query responses, it is shown that the geometric mechanism attains the lowest possible level of expected loss. The result of universal optimality with the geometric mechanism has also been extended to a broader class of *risk-averse users* [86]. This class of users seeks to minimize the worst-case expected loss over all noisy query responses. The authors show that, in fact, the geometric mechanism is universally optimal within the context of counting queries for the class of all users who measure expected loss using any function that monotonically increases with the distance between the guess and true response.

Another study examining the measure of expected loss for Bayesian users shows that counting queries are essentially the only type of query for which such a universally optimal mechanism can be designed [23]. Given this result, subsequent studies have focused on the design of non-universal optimality. The *staircase mechanism* [76] has been proposed to minimize the worst-case expected loss across all query responses. However, unlike the geometric mechanism, the staircase mechanism assumes the loss function is applied to noisy query responses taken at face value rather than being subject to post-processing. A family of staircase-shaped probability density functions has been applied to create a mechanism which is optimal within this context for real-valued query responses. Other goals in mechanism design have included a mechanism that is optimal for a particular subset of users [58], a mechanism that bounds worst-case accuracy on noisy responses [89] and non-optimal mechanisms intended to provide increased utility for more general usage [88, 135].

Due to the variety of utility measures available, a number of these measures have been studied within the context of axioms of information preservation introduced in [123]. A distinction is made between the utility of a mechanism in terms of *usability* (i.e., ease of use for specific tasks) and *information preservation* (i.e., how much of the original information is maintained in the noisy query responses). The results of the study show that distortion-based measures of utility do not satisfy all axioms of utility whereas measures of expected loss for Bayesian users do.

2.1.4 Constraint Adherence

Often, queries have natural *constraints* associated with them which can be exploited to improve the utility of the noisy query responses. Broadly speaking, these types of constraints can be divided into two categories: 1) constraints for consistency between the responses of a set of queries in the non-interactive setting, and 2) range constraints for responses to individual queries in the interactive setting. We briefly review the work in both of these areas, focusing on the later, within which our work falls.

In the non-interactive setting, where batches of related queries are posed, known relationships between the responses to a batch of queries should be preserved. This was first considered in the context of the queries needed to break a contingency table for a database up into marginals. Linear programming was applied after a transformation to the Fourier domain to ensure consistency such that the noisy marginals do not appear to have been derived from different databases [16]. A subsequent study has shown that for values of ϵ needed to ensure a reasonable level of privacy, the noise induced by the aforementioned method leads to inadequate statistical inferences [68]. Other studies have developed alternate methods to improve the utility in marginals through adherence to consistency constraints. Post-processing has been proposed to remap a vector of noisy query responses to a new vector that satisfies known relationships between the elements of the vector while minimizing the L_2 distance between the vectors [93]. In [113], it is shown that more accurate responses can be obtained by instead minimizing the L_1 distance as this accounts for the distribution of the noise added to each element of the vector. Constraints have also been considered in the context of constrained optimization problems that require a sanitized database to offer a solution to an optimization problem within a specified factor of the optimal solution on the clean data [69].

In the interactive setting in differential privacy, independent queries are answered on an individual basis. There, it may be desirable to ensure that noisy responses are consistent with a (known) valid range of the query. A well-known implementation of this is the *truncated geometric mechanism* [86], a variant of the geometric mechanism [77] which snaps out-of-bounds values to the nearest valid response. In many cases, this approach can lead to a high density of noisy responses on the boundaries of the valid range of the query. To avoid undesirable properties such as these types of spikes in probability density, the *explicit fair mechanism* [37] was designed to answer queries in

a range-adherent manner while providing properties such as a monotonically decaying probability density function (PDF). Both the geometric and the explicit fair mechanisms are specifically designed to answer counting queries (i.e., the number of database records satisfying a particular predicate). As such, they only produce integer-valued responses and cannot be applied to more general numeric queries. However, the concept of snapping out-of-bounds noisy responses to the nearest valid value can be equally applied to any type of randomization mechanism and has also been studied in the context of the Laplace mechanism in order to handle more general numeric queries [124].

The Laplace mechanism has also been applied to answer queries in a range-adherent manner by truncating the PDF to the valid range of query responses and then normalizing it. This avoids the issue of inflated probability mass at the boundaries of the range while allowing for general numeric queries to be answered. This has been proposed to be the equivalent of redrawing from a non-truncated Laplace distribution until a valid noisy response is obtained [124]. However, the study neglected to consider the impact of the information leakage inherent in the process of normalization. In independent work, concurrent to our own, Holohan et al. [99] have also proposed an alternate calculation for the scaling parameter used in a truncated and normalized Laplace distribution to guarantee differential privacy. By calculating the maximum possible change between the normalization factors for any pair of adjacent databases, they provide a method for calculating the lowest fixed scaling parameter that satisfies the privacy guarantee.

Other instances of truncated Laplace mechanisms have been proposed in the literature, however, their intent has not been adherence to a query range. In the context of approximate differential privacy [55], a truncated variant of the Laplace mechanism has been designed to maximize the probability mass decay rate of the distribution [74]. A truncated-uniform-Laplace mechanism has also been proposed for the purpose of uniformly most powerful tests on binomial data [15]. In both cases, truncation is based on mechanism parameters and is symmetric around the location parameter (i.e., the true query response). Thus, although these forms of truncation are related to utility, they do not enforce adherence to the query range.

2.1.5 Specialized Topics

We conclude our review of differential privacy with a look at a couple of specialized topics which we draw on in our own work. Generalized differential privacy offers a useful extension to the applicability of differential privacy, allowing for the protection of sensitive data outside the context of databases. Principal component analysis (PCA) serves to produce compact representations of data which preserve much of the original information content. These tools are of particular use to us in the obfuscation of images.

Generalized Differential Privacy

One direction in which the study of differential privacy has progressed is the generalization of the privacy guarantee. One such generalization replaces the notion of adjacency between databases with an arbitrary distance metric [27]. In this way, different measures of distance can be applied to allow for different representations of distinguishability between databases. Should the protection of specific values in the database be necessary, a metric based on the distance between values in the databases can be used. Furthermore, this allows for differential privacy mechanisms to be employed for the protection of arbitrary sensitive information outside the context of statistical databases. By selecting an appropriate measure of distinguishability, a mechanism can be configured to draw noise such that the differential privacy guarantee holds for any pair of values according to the level of distinguishability between the two values as determined by the specified measure. See Section 6.1.1

for the formal representation of the generalized privacy guarantee.

Principal Component Analysis

Principal component analysis (PCA) is a machine learning approach used to calculate an alternate basis within which to represent points of data (i.e., multivariate observations) such that the dimensions of the new basis are orthogonal and are arranged in descending order of explained variance over a training dataset [178]. This technique is widely used as a method to greatly reduce dimensionality of data while preserving the most important information by dropping trailing dimensions in the new basis. The reduction in dimensionality is often useful for visualization, analysis, and machine learning tasks (e.g., facial recognition [194]).

PCA has also been recognized as a useful tool in the field of differential privacy. To reduce the heavy loss of information that is often incurred in differentially private data publishing, a sensitive database can first be transformed to a PCA basis with reduced dimensionality. The reduction in dimensionality allows for the privacy budget (which controls how much noise must be injected) to be spent in a manner better suited to the preservation of useful information. Most research on this topic has focused on how to perform the PCA learning process in a differentially private manner (e.g., [28, 57]) since the training data is private. Once the data has been transformed to the PCA basis, noise can be added and the obfuscated results can then be transformed back to the original basis and released. Although less prominently addressed in the literature, the step of adding noise in the PCA basis requires care to ensure that the magnitude of the noise is controlled based on the sensitivity of the data in its new representation. This sensitivity can be calculated using information about the PCA transformation and the queries posed on the data [103].

2.2 Facial Obfuscation

The task of *facial obfuscation* is the modification of facial images such that recognition of the depicted identity is prevented. We have provided discussion in Section 1.2 on threats to privacy in facial images which motivate this task. Perhaps the most well-known and earliest studied alterations to images for the prevention of human recognition of faces are *pixelization* [91] and *blurring* [90]. Pixelization decreases the information conveyed in an image by dividing the image into a grid of cells and setting all pixels within each cell to a common pixel intensity. Blurring involves the addition of, typically Gaussian, noise to the image. While these methods have been successful at foiling human recognition, they have been shown to be highly ineffective against machine recognition [146]. Other ad hoc methods of privacy protection involving variations on blurring [114], warping [108], morphing [107] and face swapping [20, 142] have since been studied. However, the methods that have gained the most momentum are those that offer a formal guarantee of privacy. This trend has been reinforced by the legal and legislative demands in the broader context of the release of sensitive data [22, 167, 196].

In this section, we first review the early work on formalized approaches to facial obfuscation and the subsequent developments and improvements over the years. We then change track and provide a review of the major advances in the field of neural networks applied to the generation and modification of images. These networks have become an indispensable tool for many computer vision tasks and have recently shown promising results when applied to the facial obfuscation problem. To this end, we conclude the section with a review of recent works that tackle facial obfuscation via generative neural networks.

2.2.1 Formalized Approaches to Facial Obfuscation

The most prominently studied approach to formalized privacy in facial images has been the *k-same framework* [146], an adaptation of the *k*-anonymity [190] guarantee from the field of database privacy. The first application of this concept to image obfuscation worked by aligning a set of input images on their facial features, partitioning the set into clusters of *k* or more similar images, and then averaging the pixels within each cluster to produce an obfuscated face which would replace each of the original faces in the cluster [146]. By releasing only the averaged faces, a guarantee is provided which states that neither human nor machine recognition can do better than identifying the cluster of identities that produced the image. This limits the probability of successful re-identification to an upper bound of $\frac{1}{k}$.

One issue with the original *k*-same averaging of pixels was poor visual quality due to inexact alignment of facial features, leading to superimposed features. The *k-same-m* [82] approach improved upon this by using an *active appearance model* (AAM) [35] to obfuscate faces. AAMs are generative machine learning models for the approximation of visual representations of a particular class of objects (e.g., human faces). A model is trained on a set of images in order to learn about visual patterns and minimize differences with respect to shape and texture between the original images and the generated output of the model. The *k-same-m* approach first trains an AAM and then performs the clustering and averaging process within the parameter space of the model representations of faces to be obfuscated, thus eliminating the issue of superimposed features. Subsequent variants of *k*-same obfuscation via AAMs have since been proposed such as *k-same-furthest* [137] which instead partitions images into subsets such that images within the same subset have highly dissimilar model parameters, making re-identification of the averaged identities more challenging.

Efforts have also been devoted to the preservation of utility in the obfuscated images. The *k-same-select* approach [83] proposed partitioning the input images into classes based on the information to be preserved (e.g., male and female identities). By running a separate clustering process within each partition, images within each cluster would share the same class, thus preserving this information in the averaged version. This idea has been extended to the *k-same-m* model by training a different AAM for each combination over the demographic attributes of age, gender and race [48]. By using the appropriately trained AAM for obfuscation, the attributes for which it was trained can be preserved in the output. An alternative approach to the explicit specification of classes to be preserved involves the use of multimodal discriminant analysis to allow for the representation of identity and other attributes in orthogonal subspaces [179]. This allows for *k*-anonymity to be applied within the identity subspace while preserving or modifying other attributes relevant to utility, such as age, gender and race, as desired within each of their subspaces. In the context of preservation of facial expressions, an approach has been proposed to calculate the difference between AAM parameters of original instances having a neutral expression and the target expression (e.g., happiness) and then add this difference to an obfuscated instance with a neutral expression [138]. In this way, the target expression can be transferred to the obfuscated image.

Recently, differential privacy has been applied in the context of obfuscation of images. Through the addition of sufficient noise to pixel intensities, uncertainty is induced regarding the original content of the image [65]. However, direct modification of pixels in this way leads to severe visual distortions in the images, rendering them highly unnatural. This has been improved upon through the use of an invertible transformation applied to images via singular value decomposition [64]. Noise is instead added to the feature representation of the image obtained through singular value decomposition and the noisy features are then passed through the inverse transformation to produce the obfuscated output image.

2.2.2 Neural Networks

As *neural networks* are becoming an increasingly popular tool for image processing, methods of facial obfuscation are now making use of them. These networks are able to offer impressive visual quality, often at a photo-realistic level, and provide convenient ways to manipulate parts of images while preserving desired aspects. We provide a brief review of relevant networks.

While not a particularly new area of study, *convolutional neural networks* have recently enjoyed a number of advances that have greatly improved their effectiveness in computer vision tasks such as classification of images [84]. These networks primarily consist of a layered combination of convolutional filters and non-linear activation functions which accepts an image as input and produces a compact encoding that is well-suited for the intended task. Recent variations on this concept have essentially inverted the architecture by using transpose convolutions to take encodings as input and produce images as output. This has been applied to generate novel images having specific properties, potentially in combinations not seen in the training data [47]. This architecture has also been extended by pairing the *generator network* with a *discriminator network* that aims to distinguish between real or generated images [79]. The model, referred to as a *generative adversarial network* (GAN), uses an adversarial training scheme in which the two networks compete in a minimax game, forcing the generator to learn how to produce realistic instances from a targeted distribution.

The concept of GANs has spawned a rich body of work on the generation and manipulation of highly realistic images. The DCGAN [166] architecture improved training stability, allowing for deeper network models and better image quality. DCGAN was also applied to demonstrate the ability to generate novel images by performing interpolation and arithmetic in the latent encoding space. Conditional GANs [140] extended the training process by teaching the network to generate images from a conditional distribution, allowing for the specification of desired properties to be present in the generated images.

The concept of an *encoder network* was later included, allowing GANs to take an image as input and produce an encoding which could then be *decoded* (by the generator) back to an approximation of the input image [46, 162]. This concept enabled the use of GANs for image editing tasks. Networks such as StarGAN [32] and AttGAN [95] have shown impressive results in the modification of specific features (e.g., gender or hair colour) in facial images while preserving all other information in the images. It is these types of networks that are particularly useful for approaches to facial obfuscation.

2.2.3 Facial Obfuscation via Generative Neural Networks

The advances in *generative neural networks* (i.e., neural networks that generate data from a learned distribution), particularly those applied in the context of facial images, have provided a highly convenient tool for the task of facial obfuscation. We review here a number of recent works that have explored the potential applications of such networks to produce obfuscated images.

Using a model that learns facial features that are invariant to pose and illumination, *k*-same obfuscation has been applied to generate obfuscated faces from averaged clusters of facial features [31]. A similar concept has been applied to create averaged faces via a transpose convolutional architecture while maintaining separate features for facial expressions in order to allow for control and preservation of expressions in the obfuscated images [136]. To provide a stronger level of privacy than what *k*-same offers, GANs have been applied to achieve other types of formal privacy guarantees. Using modified approaches to the *k*-same clustering process, stronger instances of syntactic privacy such as *l*-diversity or *t*-closeness can be provided [119]. Alternately, privacy can be guaranteed by completely removing faces from images and then applying a GAN to in-paint the redacted area such that it produces a novel face. Approaches of this type have employed a decoder network that

takes as input the redacted image along with either facial landmarks [102, 187] or 3D facial model parameters [188] in order to generate a replacement face.

The ability of neural networks to realistically manipulate facial features has also been exploited in contexts outside of formal privacy guarantees. Examples include the use of a training objective function that promotes changes in facial identity [206] and the application of a transformation to image encodings that induces large changes to identity-specific features while preserving certain desirable attributes [155]. Another informal approach to privacy protection in facial images is the concept of face swapping [20] which has recently been applied using neural networks to perform the transfer and blending process [122, 151].

We note that differential privacy has also been applied in the context of training GANs in a privacy-preserving fashion. However, this goal is fundamentally different from that of facial obfuscation. To protect sensitive training data, noise is added to the gradients calculated for the discriminator network during training [207, 208]. This allows for control over the influence of the sensitive information on the weights and distribution learned by the generator network in the form of a differential privacy guarantee. The generator can then produce new samples from the learned distribution while protecting the privacy of the training samples. Variations on this concept have been applied to generate synthetic datasets of facial images in a differentially private manner [193, 211]. However, none of these applications of differential privacy are suitable for the task of facial obfuscation in which privacy must be guaranteed for new images presented to the network after training is complete.

Chapter 3

Thesis Overview and Contributions

The work presented in this thesis can broadly be categorized as falling within the field of study of differential privacy. More specifically, we focus on two major topics: 1) the design of range-adherent differentially private mechanisms, and 2) the application of differential privacy to generative models for facial images. These topics are each treated in their own chapter of the thesis with further subdivision into sections covering our major contributions.

In Chapter 5, we present our work on the design of range-adherent differentially private mechanisms. We study this problem first in terms of the design of truncated and normalized Laplace mechanisms and later in terms of linear programming mechanisms. The specific contributions we make throughout the chapter are as follows:

- **Section 5.1:** We demonstrate that the process of truncation and normalization of a Laplace distribution is not sufficient on its own to preserve the differential privacy guarantee. The main reason for this results from the data-dependent nature of the normalization.
- **Section 5.1:** We show how this can be corrected by carefully calculating the scaling parameter for the Laplace distribution. For this, we generalize the differential privacy guarantee in the context of a Laplace mechanism to incorporate data-dependent normalization factors.
- **Section 5.1:** We use our privacy guarantee to guide a study of range-adherent mechanisms with respect to different classes of range constraints they are able to adhere to. For each class we show how to derive an optimal scaling parameter or an approximation thereof.
- **Section 5.2:** We provide a theoretical discussion on the desirable properties of range-adherent mechanisms and formalize a matrix representation of range-adherent mechanisms for general numeric queries.
- **Section 5.2:** We propose two variants of a range-adherent linear programming mechanism using our discretized mechanism representation. By employing a user-independent criterion for optimal utility, we produce mechanisms that are useful to a broad range of users.
- **Section 5.3:** We demonstrate the benefits of our proposed truncated and normalized Laplace mechanisms through experiments using a differentially private Naïve Bayes classifier. We show that our mechanisms are able to achieve higher classification accuracy than other range-adherent alternatives.

- **Section 5.3:** We conduct experiments on our proposed linear programming mechanisms and a variety of existing mechanisms using a number of measures of utility to assess the usability and information preservation of the mechanisms. We compare the performance of the mechanisms and demonstrate improvements in the utility of our linear programming mechanisms over the commonly used boundary-snapping method.

In Chapter 6, we present our work on the obfuscation of facial images via generative models. This work covers the development of a framework to apply differential privacy in this setting and the application of our proposed framework to different generative models. In particular, we focus on design and training details to achieve obfuscation via a generative adversarial network (GAN). Our contributions throughout the chapter are as follows:

- **Section 6.2:** We discuss deficiencies in existing methods of facial obfuscation and examine susceptibilities of k -same obfuscation to composition attacks and background knowledge. We demonstrate how the k -same privacy guarantee can be violated and discuss the implications this has on privacy in images.
- **Section 6.3:** To improve upon the deficiencies we have identified in other methods of facial obfuscation, we propose as an alternative, the formal guarantee of differential privacy. We develop the first framework to apply differential privacy for the obfuscation of facial identity in images via generative machine learning models. We then derive the configuration of a Laplace mechanism to achieve the privacy guarantee in this setting.
- **Section 6.4:** We propose architecture and training details that allow for the application of differential privacy to GAN image encodings. Our proposed approach allows for photo-realistic obfuscation of facial images from an input distribution that is largely unconstrained.
- **Section 6.5:** We propose how to apply distance-generalized differential privacy to image encodings transformed to a principal component analysis (PCA) basis in order to preserve a high degree of utility in obfuscated images. We explain how to manage the GAN training process in order to generate realistic images from noisy PCA-basis encodings.
- **Section 6.6:** We provide details on both theoretical and practical interpretations of the formal privacy guarantee that we achieve for obfuscated images. We relate our instantiation of the privacy guarantee to that of standard differential privacy (for databases) and offer some intuition on the interpretation of the privacy parameter ϵ as used in our setting.
- **Section 6.7:** We additionally propose a method to enforce differential privacy via the direct modification of pixel intensities. By giving up the high visual quality provided by generative models, this allows for a much more versatile approach that can obfuscate any image. We employ a process guided by an image quality function in order to improve the visual quality over the existing results for differential privacy in pixel-space.
- **Section 6.8:** We conduct a series of experiments to compare differential privacy to k -same obfuscation on two well-known datasets. We demonstrate the resilience of differential privacy to composition and parrot attacks. Furthermore, the results of our experiments suggest that differential privacy offers a comparable level of utility in the obfuscated images to k -same obfuscation while providing a stronger privacy guarantee.
- **Section 6.8:** We provide an experimental evaluation of our proposed GAN using our approach for obfuscation within a PCA basis. We demonstrate improved privacy-utility trade-offs over

implementations of other methods of facial obfuscation which also offer formal guarantees of privacy.

Chapter 4

Notation and Definitions

In this chapter, we provide notation and definitions that are relevant to the thesis as a whole. The differential privacy guarantee is central to any mechanism that aims to enforce differential privacy. Satisfaction of this guarantee (or variants thereof) is a requirement that spans across all of our work. We provide here the basics on the definition of the guarantee and two commonly-used randomization mechanisms which satisfy this guarantee. We then provide details on a well-known method of range-adherence, boundary-snapping. Throughout the remainder of the thesis, we provide extensions and refinements on these concepts in order to apply them to the specific settings in which we perform our work.

4.1 Absolute Value and Set Cardinality Notation

Throughout this work, we employ the notation $|\cdot|$ in different contexts. When applied to a scalar value, this implies an absolute value operation. When applied to a set, typically denoted using a capital letter variable, this implies a set cardinality operation.

4.2 The Differential Privacy Guarantee

Differential privacy [56] (Definition 1) enforces the guarantee that similar databases must have similar probabilities of producing the same noisy query responses. Similarity between databases is interpreted using an adjacency relation such that a pair of databases D_1 and D_2 are considered to be adjacent (denoted by $D_1 \sim D_2$) if they differ by a single record.

Definition 1. *Let $\mathbb{D}$ be the set of potential database configurations, $f : \mathbb{D} \rightarrow \mathbb{R}$ be a query function, $K : \mathbb{R} \rightarrow \mathbb{R}$ be a randomization mechanism and ϵ be a privacy parameter. For K to satisfy the **differential privacy guarantee** [56], the following condition must hold for all pairs $D_1, D_2 \in \mathbb{D}$ such that $D_1 \sim D_2$ and for all subsets S of the mechanism range $\mathbb{R}$:*

$$\Pr(K(f(D_1)) \in S) \leq e^\epsilon \Pr(K(f(D_2)) \in S), \forall S \subseteq \mathbb{R}. \quad (4.1)$$

This guarantees that for every pair of adjacent databases and every subset of noisy query responses, the respective probabilities of the two databases to produce a response within the given subset must be within a multiplicative factor of e^ϵ of each other, where ϵ is a privacy parameter

specified by the data custodian. The lower the value of ϵ is set, the stricter the guarantee becomes as the distributions over the noisy query responses are forced to become more similar to each other.

4.3 Query Sensitivity

In order to guide the design of differentially private mechanisms, the concept of query sensitivity [56] (Definition 2) is often used to measure the greatest impact the presence or absence of a database record can have on the query response.

Definition 2. *Given a set of potential database configurations $\mathbb{D}$ and a query function $f : \mathbb{D} \rightarrow \mathbb{R}$, the **query sensitivity** ΔF is defined to be the largest possible difference between the query responses for any pair of adjacent databases:*

$$\Delta F = \max_{D_1, D_2 \in \mathbb{D} | D_1 \sim D_2} |f(D_1) - f(D_2)|. \quad (4.2)$$

4.4 The Laplace Mechanism

For answering general numeric queries in a differentially private manner, the Laplace distribution [109] (Definition 3) is commonly used to implement the Laplace mechanism [53] (Definition 4).

Definition 3. *The univariate, continuous **Laplace distribution** [109] is a probability distribution given by the following probability density function:*

$$Lap(x|\mu, \sigma) = \frac{e^{-\frac{|\mu-x|}{\sigma}}}{2\sigma}, \quad (4.3)$$

where μ is a location parameter and σ is a scaling parameter. The variance of the distribution is $2\sigma^2$.

To satisfy the differential privacy guarantee, the scaling parameter must be set in terms of the chosen value of ϵ and the query sensitivity, ΔF .

Definition 4. *For a database $D \in \mathbb{D}$, the **Laplace mechanism** [53] draws from $Lap(x|f(D), \frac{\Delta F}{\epsilon})$ to generate noisy query responses.*

The resulting probability densities of the Laplace mechanism, interpreted through a union bound, satisfy the differential privacy guarantee [53].

4.5 The Exponential Mechanism

In some cases, it may be preferable to control the application of noise to query responses via a quality function which indicates the usefulness of a noisy response given the true query response. This may be of particular use when query responses are categorical or have their utility interpreted through more complex functions than distance between the true response and the noisy response. The exponential mechanism [135] (Definition 5) satisfies the differential privacy guarantee by drawing noisy query responses from an exponential distribution using a quality function as the exponent.

Definition 5. Given a database $D \in \mathbb{D}$, a privacy parameter ϵ , a range R of possible query responses, a quality function $q : R \times \mathbb{D} \rightarrow \mathbb{R}$ and a base distribution μ over R , the **exponential mechanism** [135] is a randomization mechanism that generates noisy responses $r \in R$ with probability proportional to:

$$e^{\epsilon q(r, D)} \mu(r). \quad (4.4)$$

4.6 Boundary-Snapping

Often, it is desirable to ensure that noisy query responses conform to a valid range of potential true query responses. Boundary-snapping [86, 124] (Definition 6) provides a simple scheme of remapping query responses to achieve such range-adherence.

Definition 6. *Boundary-snapping* [86, 124] is a post-processing operation applied to the output of a randomization mechanism in which out-of-bounds noisy query responses are remapped to the nearest valid response.

This operation can be directly incorporated into the probability distribution of a mechanism by reallocating all out-of-bounds probability mass to the nearest noisy response value. Since this alteration is equivalent to a post-processing operation and differential privacy is immune to privacy breaches induced by post-processing [56], the privacy guarantee is preserved under this modification.

Chapter 5

Constraint-Aware Mechanisms

Adherence of differentially private mechanisms to publicly known constraints can affect the utility of the mechanisms. Particularly in the non-interactive setting, where batches of queries are posed, known constraints over the set of query responses can be used to enforce consistency on the noisy responses in order to improve utility [16, 69]. For example, the sum of the responses for counting queries (i.e., the number of records that satisfy a given predicate) over disjoint subsets of a database should be equal to the response to the same counting query over the union of the same subsets. In the interactive setting, where unrelated queries are answered individually, such constraints are not present since there are no other query responses with which relationships must be preserved. However, knowledge about the range of the query can still be exploited to ensure that noisy query responses fall within the valid range [37, 124]. This typically improves the utility of the noisy responses and is beneficial for the compatibility of the mechanism with downstream software used to analyze the responses.

In many applications, queries will have natural and publicly known constraints on their range. For example, the percentage of satisfied (or dissatisfied) customers in a survey must fall in the range of 0 to 100. Inferences about the range of a query can often be drawn with ease. The sum (or mean, median, etc.) of positive-valued attributes cannot be less than zero and the number of records in a database that satisfy a particular predicate cannot be greater than the size of the database. If a valid range is known for a particular attribute, a valid range for most queries posed on that attribute can be calculated. Often, the range of attribute values can be easily determined. For example, a database may contain signal measurements from a device that has a known range on the values it can produce, or record values may be timestamps within a specified interval of time. In practice, most differentially private mechanisms add noise without consideration for such constraints. This implies that noisy responses may be generated outside of the valid range of the query.

Since range constraints are quite natural in many cases (e.g., positive-valued query responses), this information should be leveraged for the design of mechanisms better suited to the preservation of data utility. Adherence to range constraints could be achieved through post-processing of noisy query responses, either via boundary-snapping or by performing Bayesian post-processing [77]. However, these approaches may not always be appropriate in practice. Boundary-snapping leads to pooling of probability mass at the boundaries of the valid range, which may not be near the true query response, and Bayesian post-processing may not be computationally tractable for queries with a continuous range. Alternatively, mechanism design can integrate information about the valid range of the query in order to adhere to the range constraints while also taking the utility of the noisy responses into consideration. In this chapter, we study the problem of designing range-adherent, interactive

mechanisms that offer improvements in utility over boundary-snapping and other alternatives in the literature.

We propose two different approaches to the design of range-adherent mechanisms. In the first approach (Section 5.1), we truncate the probability density function (PDF) of a Laplace distribution to the valid range of the query and then normalize it. Since normalization involves the multiplication of the function by a factor greater than 1 and the Laplace distribution has higher probability density around the location parameter, this process leads to greater increases in probability density for noisy responses nearby the true response compared to those farther away. This is beneficial for the utility of the mechanism as it leads to a lower level of expected distortion when noise is drawn from the new distribution.

In the second approach (Section 5.2), we use a matrix representation of discretized probability density functions and a user-independent criterion for optimal utility to apply linear programming to produce range-adherent mechanisms. We propose two variants of this mechanism: one that offers optimal utility under the restriction of range-adherence, and one that places additional restrictions on the properties of the probability distribution of the mechanism in order to ease analysis performed on the mechanism output.

The remainder of the chapter is structured as follows: In Section 5.1, we provide our work on the design of differentially private truncated and normalized Laplace mechanisms. We follow this with our work on the design of range-adherent linear programming mechanisms in Section 5.2. We provide experimental comparisons for our proposed mechanisms in Section 5.3. Finally, we discuss future work in Section 5.4 and conclude the chapter in Section 5.5.

The content of Section 5.1 is derived from our submission to the Journal of Computer Science and Technology titled “Differential Privacy via a Truncated and Normalized Laplace Mechanism”. The content from Section 5.2 comes from our submission to Transactions on Data Privacy titled “The Impact of Range Constraints on Utility in the Design of Differentially Private Mechanisms”. The experimental comparisons in Section 5.3 are derived from both of these papers. We refer the reader back to Chapter 3 for details on these submissions.

5.1 Normalized Laplace Mechanisms

In this section, we propose an approach to generate range-adherent query responses where we first truncate the PDF of a Laplace distribution to the valid range of the query and then normalize it. This subsequent normalization is required to restore the truncated function to a valid PDF. Since the probability density in the Laplace distribution exponentially decays as distance from the location parameter increases, multiplication by the normalization factor will induce greater increases in probability density for noisy query responses that are closer to the true response than those that are farther away. This is favourable for the utility of the mechanism as it is conducive to a low level of expected distortion in the noisy query responses.

The method of truncation and normalization has been previously applied [124] in the context of the Laplace distribution. However, the study did not account for the fact that the normalization factor is a function of the true query response, which is sensitive information. Due to this, normalization in fact leaks further information. Verifying the privacy guarantee becomes more complex and requires further modification of the PDF. We propose an alternative method (to that which is used by the standard Laplace mechanism) for the calculation of the scaling parameter in the Laplace distribution. Our proposed alternative accounts for the data-dependent normalization in order to preserve the privacy guarantee of differential privacy.

In this work, we adapt the privacy guarantee in the context of the Laplace distribution to

account for data-dependent normalization factors and study this guarantee for different classes of range constraint configurations. We provide derivations of the optimal scaling parameter (i.e., the minimal value that preserves differential privacy) for each class or provide an approximation thereof. Our work thus provides a novel approach to apply the Laplace distribution to answer queries in a range-adherent and differentially private manner.

5.1.1 Preliminaries

We begin by providing the preliminaries that are needed to study the impact of truncation and normalization of a Laplace PDF with respect to the privacy guarantee. We first formalize our treatment of range constraints and lay out the calculations of normalization factors for classes of constraint configurations. We then examine the impact of adherence to range constraints and data-dependent normalization on the privacy guarantee. We define an adapted privacy guarantee to use in this setting which provides a stronger level of privacy than the standard differential privacy guarantee. With these components laid out, we provide a small example to demonstrate the violation of the privacy guarantee when truncation and normalization are applied without further modification to the PDF.

In Table 5.1, we provide a quick reference for the definitions of the variables used throughout this section.

Constraint Configurations

Given constraints on the range of a query represented as a set of ranges within which query responses cannot occur, the goal is to truncate a Laplace PDF to remove these constrained ranges. This can be achieved via boundary-snapping by shifting all out-of-bounds probability mass to the nearest boundary response. However, the pooling of probability on the boundaries is undesirable for queries in which the true response is not near a boundary. Instead, we propose to calculate an appropriate normalization factor to restore the integral over the truncated span(s) to 1 and then calculate an appropriate scaling parameter. The required scaling parameter is dependent upon a privacy parameter ϵ specified by the data custodian, the query sensitivity ΔF , and the position of the location parameter (i.e., the true query response) relative to the constraints.

The scaling parameter must be selected such that the differential privacy guarantee is satisfied. The higher the scaling parameter is set, the farther the probability mass becomes spread from the location parameter. Therefore, calculating the lowest possible scaling parameter that satisfies the privacy guarantee is one of the core aspects of our work. We consider the selection of the lowest possible scaling parameter to be the criterion for optimal utility within the class of mechanisms using a truncated and normalized Laplace PDF. While this form of optimality aims to minimize the expected noise of the mechanism, it may lead to biased responses compared to a non-optimal approach. The direction of bias-variance trade-offs is not explored in this work.

Allowing for truncated spans in a PDF provides a natural way to incorporate range constraints into a randomization mechanism. Given the wide variety of queries that can be posed on a database, many different configurations of constraints may arise in practice. To simplify the work required in designing appropriate mechanisms for the various constraint configurations, we categorize these configurations into classes such that for each class, a corresponding mechanism can be developed. Visual representations of possible constraint configurations from each class are shown in Figure 5.1.

Table 5.1: Definitions of variables

$\mathbb{R}$	The set of real-valued numbers
$\mathbb{D}$	The set of potential database configurations
K	A function $K : \mathbb{R} \rightarrow \mathbb{R}$ representing a randomization mechanism
f	A function $f : \mathbb{D} \rightarrow \mathbb{R}$ representing a query posed on a database
ΔF	Query sensitivity - The maximum possible difference between the query responses of two adjacent databases. This is a non-zero, positive, real value.
ϵ	Privacy parameter - The user-specified, desired level of privacy. This is a non-zero, positive, real value.
σ	Scaling parameter (for cases where all PDFs share the same scaling parameter) - A value calculated to determine the scale of noise required to release a database query. This is a non-zero, positive, real value.
x	The value taken on by the continuous random variable of a PDF
n_j	The normalization factor required for database D_j
σ_j	The scaling parameter required for database D_j
Δx_j	The distance between $f(D_j)$ and the value x of a continuous random variable
CL_j	The set of finite constraints to the left of $f(D_j)$
CR_j	The set of finite constraints to the right of $f(D_j)$
I_L	A boolean variable used to indicate the presence of a constraint spanning to negative infinity
I_R	A boolean variable used to indicate the presence of a constraint spanning to infinity
ΔR_j	The distance between $f(D_j)$ and the right endpoint of a constraint spanning to negative infinity
ΔL_j	The distance between $f(D_j)$ and the left endpoint of a constraint spanning to infinity
ΔR_{jk}	The distance between $f(D_j)$ and the right endpoint of a constraint C_k
ΔL_{jk}	The distance between $f(D_j)$ and the left endpoint of a constraint C_k
L_j	The summation of the integrals of all constraints to the left of $f(D_j)$
R_j	The summation of the integrals of all constraints to the right of $f(D_j)$
i	A real-valued variable used to indicate the number of spans of ΔF between two databases

Figure 5.1: An example from each class of constraint configurations is shown. The black line represents a real-valued, one-dimensional range and the shaded red blocks represent constraints. Constraints shown on the endpoint of the line span infinitely in that direction. The constraints classes are: a) a single infinite constraint, b) arbitrary finite constraints, and c) arbitrary constraints.

The first class is characterized by a single constraint that begins at any point and spans either to infinity or to negative infinity (Figure 5.1-a). An example of this is a query that has the set of natural numbers or positive real numbers as its range.

The second class may have any number of finite spans of constraints (Figure 5.1-b). These configurations would be typical of queries in which the range is subject to a periodic function. For example, a query may be posed about dates on which individuals visited a particular medical clinic where it is known that the clinic is closed during certain periods of the year.

We also consider a final class of arbitrary constraint configurations to handle any combination of the first two classes (Figure 5.1-c). Common instantiations of this class are queries that have a single finite span of valid responses. For example, a query may have a single finite range of possible age values.

Normalization Factors

Since constraints represent infeasible spans in the range of the query, the PDF must be truncated to remove these spans. This has the effect of producing a piece-wise function whose integral is no longer equal to 1. In order to restore this to a PDF, it must be scaled by a normalization factor such that the integral over the valid space is equal to 1. For a PDF with a location parameter at $f(D_j)$, let L_j be the sum of all integrals of the constraints to the left of $f(D_j)$ and let R_j be the sum of all integrals of the constraints to the right of $f(D_j)$. The normalization factor for this PDF can then be calculated as shown in Formula (5.1). An example of a truncated and normalized Laplace PDF is shown in Figure 5.2.

$$n_j = \frac{1}{1 - (L_j + R_j)}. \quad (5.1)$$

Figure 5.2: A graph of two Laplace PDFs with location parameters of 0 and scaling parameters of 1. The shaded blocks show an example of different types of constraints (infinite and finite on both the left and right of the location parameter). The red line shows the original Laplace PDF and the blue line shows one that has been truncated and then normalized.

To calculate the values of L_j and R_j , we use the Laplace cumulative distribution function (CDF).

Definition 7. The *cumulative distribution function* corresponding to a Laplace distribution of $Lap(x|\mu, \sigma)$ is given by:

$$CDF(x|\mu, \sigma) = \frac{1}{2} + \frac{1}{2} \operatorname{sgn}(x - \mu) \left(1 - e^{-\frac{|\mu - x|}{\sigma}}\right), \quad (5.2)$$

where $\operatorname{sgn} : \mathbb{R} \rightarrow [-1, 0, 1]$ is the sign function.

We define the following notation for the representation of constraints. Let CL_j be the set of finite constraints to the left of $f(D_j)$ for any database $D_j \in \mathbb{D}$. For any constraint $C_k \in CL_j$, let ΔL_{jk} and ΔR_{jk} be the distance between $f(D_j)$ and the left and right endpoints, respectively, of C_k . Let I_L be an indicator variable set to either 0 or 1 to denote the absence or presence of a constraint that spans to negative infinity and let ΔR_j be the distance between the right endpoint of such a constraint, if it exists, and $f(D_j)$. These concepts are all defined analogously for a set CR_j of finite constraints to the right of $f(D_j)$ with I_R acting as the indicator variable for a constraint spanning to infinity and ΔL_j acting as the distance between the left endpoint of such a constraint, if it exists, and $f(D_j)$. Finally, let σ_j be a data-dependent scaling parameter to be used when querying a database D_j .

The integral of a finite constraint $C_k \in CL_j$ can be calculated as $CDF(-\Delta R_{jk}|f(D_j), \sigma_j) - CDF(-\Delta L_{jk}|f(D_j), \sigma_j)$. The integral of a constraint that spans to negative infinity can be calculated as $CDF(-\Delta R_j|f(D_j), \sigma_j)$. Using this, we can now calculate the value of L_j as shown in Formula (5.3).

$$L_j = \sum_{C_k \in CL_j} \left(\frac{e^{-\frac{\Delta R_{jk}}{\sigma_j}} - e^{-\frac{\Delta L_{jk}}{\sigma_j}}}{2} \right) + I_L \frac{e^{-\frac{\Delta R_j}{\sigma_j}}}{2}. \quad (5.3)$$

Analogous calculations for the integrals of finite constraints $C_k \in CR_j$ and a constraint spanning to infinity are used in Formula (5.4) to calculate R_j .

$$R_j = \sum_{C_k \in CR_j} \left(\frac{e^{-\frac{\Delta L_{jk}}{\sigma_j}} - e^{-\frac{\Delta R_{jk}}{\sigma_j}}}{2} \right) + I_R \frac{e^{-\frac{\Delta L_j}{\sigma_j}}}{2}. \quad (5.4)$$

Generalized Differential Privacy Guarantee

When using the basic Laplace distribution to implement the randomization mechanism, the privacy guarantee of Definition 1 can be simplified to the form shown in Formula (5.5) for all pairs $D_1, D_2 \in \mathbb{D}$ such that $D_1 \sim D_2$, where Δx_j is the distance between the value x taken on by a continuous random variable and $f(D_j)$ for any database $D_j \in \mathbb{D}$. For ease of notation, we represent all subsequent forms of the privacy guarantee in terms of probability densities rather than probabilities for reporting a subset of the mechanism range. This is done with the understanding that a union bound can be applied [53] to interpret the guarantee in terms of the typical form using subsets (i.e., as it is represented in Formula (4.1)).

$$e^{\frac{\Delta x_2}{\sigma}} \leq e^\epsilon e^{\frac{\Delta x_1}{\sigma}}, \quad \forall x \in \mathbb{R}. \quad (5.5)$$

When using normalization factors induced by constraints and allowing for differing scaling parameters, we obtain a new general form of the privacy guarantee for all pairs $D_1, D_2 \in \mathbb{D}$ such that $D_1 \sim D_2$, as shown in Formula (5.6).

$$\frac{\sigma_2}{\sigma_1} \left(\frac{n_1}{n_2} \right) \left(\frac{e^{\frac{\Delta x_2}{\sigma_2}}}{e^{\frac{\Delta x_1}{\sigma_1}}} \right) \leq e^\epsilon, \quad \forall x \in \mathbb{R}. \quad (5.6)$$

To simplify the analysis, we assume that when two databases, $D_1, D_2 \in \mathbb{D}$ are being compared in a privacy guarantee, it will always be the case that $f(D_1) \geq f(D_2)$. As a result, a form for the symmetric case of the guarantee (i.e., where $f(D_1) \leq f(D_2)$) must also be considered; this is shown in Formula (5.7). To assert the differential privacy guarantee, both forms must be satisfied for all pairs of adjacent databases that apply to their respective cases.

$$\frac{\sigma_1}{\sigma_2} \left(\frac{n_2}{n_1} \right) \left(\frac{e^{\frac{\Delta x_1}{\sigma_1}}}{e^{\frac{\Delta x_2}{\sigma_2}}} \right) \leq e^\epsilon, \quad \forall x \in \mathbb{R}. \quad (5.7)$$

Adaptation for Arbitrary Distances

The differential privacy guarantee is generally interpreted in terms of pairs of adjacent databases. For such databases, the distance between their true query responses is upper bounded by the query sensitivity ΔF . Although not typically done, the privacy guarantee can also be interpreted in terms of the actual distance between query responses of arbitrary databases (i.e., pairs that are not necessarily adjacent). This is a useful interpretation for our setting since the distance between the databases can be used as a means to calculate the difference between their normalization factors as well.

For a pair of databases $D_1, D_2 \in \mathbb{D}$ having PDFs with location parameters in the same span of truncated space, the constraint endpoint distances of D_2 can be written in terms of the variables used for D_1 . This can be done by applying an offset equal to the distance between the two location parameters. We define a distance variable $i = \frac{f(D_1) - f(D_2)}{\Delta F}$, representing the number of spans of ΔF between the location parameters, which we utilize to create a new representation of the constraint integral calculations. Since we assume that $f(D_1) \geq f(D_2)$, we know that $f(D_2)$ will be closer to all constraints to the left and farther from all constraints to the right compared to $f(D_1)$. We therefore use an offset of $-i\Delta F$ for distances to the left and an offset of $i\Delta F$ for distances to the right. The distance offset versions of the constraint integral calculations are shown in Formulae (5.8) and (5.9).

$$L_2 = \sum_{C_k \in CL_1} \left(\frac{e^{-\frac{\Delta R_{1k} - i\Delta F}{\sigma_2}} - e^{-\frac{\Delta L_{1k} - i\Delta F}{\sigma_2}}}{2} \right) + I_L \frac{e^{-\frac{\Delta R_1 - i\Delta F}{\sigma_2}}}{2}. \quad (5.8)$$

$$R_2 = \sum_{C_k \in CR_1} \left(\frac{e^{-\frac{\Delta L_{1k} + i\Delta F}{\sigma_2}} - e^{-\frac{\Delta R_{1k} + i\Delta F}{\sigma_2}}}{2} \right) + I_R \frac{e^{-\frac{\Delta L_1 + i\Delta F}{\sigma_2}}}{2}. \quad (5.9)$$

To use these integral calculations in the privacy guarantee, we must consider how the required multiplicative bound between the probabilities of the databases producing the same noisy response should reflect this modification. Since e^ϵ is the multiplicative bound when $|f(D_1) - f(D_2)| \leq \Delta F$, it follows that when $((i-1)\Delta F) \leq |f(D_1) - f(D_2)| \leq i\Delta F$ for $i \in \mathbb{Z}$, the multiplicative bound must be $e^{i\epsilon}$. In fact, i does not need to be an integer; it can be real-valued. In this case, we specify the requirement that when $|f(D_1) - f(D_2)| = i\Delta F$ for $i \in \mathbb{R}$, the multiplicative bound must be $e^{i\epsilon}$.

To apply this adaptation to the privacy guarantee, we use L_1 and R_1 defined according to Formulae (5.3) and (5.4) as well as L_2 and R_2 defined according to Formulae (5.8) and (5.9) to produce the form of the privacy guarantee shown in Formula (5.10).

$$\frac{\sigma_2}{\sigma_1} \left(\frac{1 - (L_2 + R_2)}{1 - (L_1 + R_1)} \right) \left(\frac{e^{\frac{\Delta x_2}{\sigma_2}}}{e^{\frac{\Delta x_1}{\sigma_1}}} \right) \leq e^{i\epsilon}, \quad \forall x \in \mathbb{R}. \quad (5.10)$$

The guarantee we work with is in fact slightly stronger than the standard guarantee of Formula (4.1). It is clear that for a pair of adjacent databases with maximally distant query responses (i.e., when $i = 1$), we enforce the standard multiplicative bound of e^ϵ . Yet for other pairs of adjacent databases where $i < 1$, we require a stronger guarantee for a multiplicative bound of $e^{i\epsilon}$. This form helps us to calculate variable-valued scaling parameters. As we show in our experiments (Section 5.3.1), this often allows us to achieve better utility than an approach that calculates the optimal fixed-value scaling parameter for the standard form of the guarantee.

Continuous Random Variable Worst-Case Analysis

In the privacy guarantee of Formula (5.10), we can see that the variables related to the value taken on by the continuous random variable of the PDFs appear only in the third fraction of the left-hand side of the inequality. We now look for the worst case with respect to the value of the continuous random variable in order to eliminate the need for explicit quantification over this variable in the privacy guarantee. Since the left-hand side of the guarantee must always be less than or equal to the right-hand side, the worst-case occurs when the fraction is maximized.

Applying identities, the fraction containing the variables related to the value taken on by the continuous random variable can be re-written as:

$$e^{\frac{\Delta x_2}{\sigma_2} - \frac{\Delta x_1}{\sigma_1}}. \quad (5.11)$$

This expression is maximized when the exponent is maximized. If we redefine Δx_2 in terms of Δx_1 based on the distance $i\Delta F$ between the location parameters, we can write out three different cases which cover the possible values of the continuous random variable. These cases are shown in Formula (5.12). Recall that $f(D_1) \geq f(D_2)$ and $f(D_1) - f(D_2) = i\Delta F$.

$$e^{\frac{\Delta x_2}{\sigma_2} - \frac{\Delta x_1}{\sigma_1}} = \begin{cases} e^{\frac{\Delta x_1 - i\Delta F}{\sigma_2} - \frac{\Delta x_1}{\sigma_1}} & x \leq f(D_2) \\ e^{\frac{i\Delta F - \Delta x_1}{\sigma_2} - \frac{\Delta x_1}{\sigma_1}} & f(D_2) < x < f(D_1) \\ e^{\frac{\Delta x_1 + i\Delta F}{\sigma_2} - \frac{\Delta x_1}{\sigma_1}} & f(D_1) \leq x. \end{cases} \quad (5.12)$$

Of these cases, the third will produce the highest value, thus the value of the continuous random variable should be set in the interval $[f(D_1), \infty)$. In order to determine which position within this interval maximizes the value of the expression, we must consider the relationship between σ_1 and σ_2 .

If $\sigma_1 = \sigma_2$ then the value of the continuous random variable within the identified interval would have no impact on the value of the expression. However, we would ideally like to take advantage of the fact that the farther a location parameter is from a constraint, the less of the PDF integral is lost during truncation. A larger remaining integral means that a lower normalization factor is needed. As we later show, higher scaling parameters are needed to compensate for the increased normalization factors. However, higher scaling parameters lead to higher levels of expected noise. Therefore, it is desirable to allow the scaling parameters to decrease as the normalization factors decrease.

We first consider the case of a single constraint spanning to infinity (note that this can be analogously treated for a constraint spanning to negative infinity). If the constraint spans to infinity, this implies that in this configuration, $\sigma_1 \geq \sigma_2$. A greater σ_1 value means that the value of the expression goes up as Δx_1 increases. We must therefore work with the highest possible Δx_1 value. If the guarantee can be satisfied for this value then any other possible Δx_1 value will cause the value to the left-hand side to decrease, meaning that the guarantee will hold. We therefore place x on the border of the constraint going to infinity (if one exists). This means that $\Delta x_1 = \Delta L_1$. The fraction can be updated to:

$$\frac{e^{-\frac{\Delta L_1}{\sigma_1}}}{e^{-\frac{\Delta L_1 + i\Delta F}{\sigma_2}}}. \quad (5.13)$$

In constraint configurations where there are no infinite constraints, the worst-case with differing scaling parameters is undefined since the value of the continuous random variable can always be further increased. In such a configuration, it becomes necessary to use the same scaling parameter everywhere to avoid this problem. When this occurs, the fraction can be written as:

$$e^{\frac{i\Delta F}{\sigma}}. \quad (5.14)$$

We must also consider the form for the symmetric case of the privacy guarantee. This time, the worst-case occurs when we maximize:

$$e^{\frac{\Delta x_1}{\sigma_1} - \frac{\Delta x_2}{\sigma_2}}. \quad (5.15)$$

When redefining Δx_2 in terms of Δx_1 based on the distance between the location parameters, the three possible cases are as shown in Formula (5.16).

$$e^{\frac{\Delta x_1}{\sigma_1} - \frac{\Delta x_2}{\sigma_2}} = \begin{cases} e^{\frac{\Delta x_1}{\sigma_1} - \frac{\Delta x_1 - i\Delta F}{\sigma_2}} & x \leq f(D_2) \\ e^{\frac{\Delta x_1}{\sigma_1} - \frac{i\Delta F - \Delta x_1}{\sigma_2}} & f(D_2) < x < f(D_1) \\ e^{\frac{\Delta x_1}{\sigma_1} - \frac{\Delta x_1 + i\Delta F}{\sigma_2}} & f(D_1) \leq x. \end{cases} \quad (5.16)$$

The first and second cases will be larger than the third. The worst-case occurs when the second term of the exponent is minimized. The smallest possible value it can take on is 0, which occurs when $\Delta x_1 = i\Delta F$. The fraction can therefore be re-written as:

$$e^{\frac{i\Delta F}{\sigma_1}}. \quad (5.17)$$

For cases where the same σ values are used for the worst-case analysis with the symmetric form, this would be:

$$e^{\frac{i\Delta F}{\sigma}}. \quad (5.18)$$

Motivating Example

We conclude this section with an example to illustrate the need for an alternate calculation of the scaling parameter when the Laplace PDF is truncated and normalized. Consider a case with a single constraint spanning to negative infinity where a fixed value for σ is used regardless of the value of $f(D)$. The privacy guarantee would be written as shown in Formula (5.19).

$$\frac{2 - e^{-\frac{\Delta R_1 + i\Delta F}{\sigma}}}{2 - e^{-\frac{\Delta R_1}{\sigma}}} \left(e^{\frac{i\Delta F}{\sigma}} \right) \leq e^{i\epsilon}. \quad (5.19)$$

The regular calculation of the scaling parameter for a Laplace mechanism is to set $\sigma = \frac{\Delta F}{\epsilon}$. When making this substitution, we have the form shown in Formula (5.20).

$$\frac{2e^{i\epsilon} - e^{-\frac{\epsilon\Delta R_1}{\Delta F}}}{2 - e^{-\frac{\epsilon\Delta R_1}{\Delta F}}} \leq e^{i\epsilon}. \quad (5.20)$$

Now we consider the comparison of a database with a location parameter on the border of the constraint to any other database. To do so, we set ΔR_1 to 0 and simplify to produce the form shown in Formula (5.21).

$$2e^{i\epsilon} - 1 \leq e^{i\epsilon}. \quad (5.21)$$

We can see that for any $i > 0$ (i.e., any pair of databases that do not share the same true query response), the privacy guarantee is not satisfied. This shows that the regular calculation of the

scaling parameter cannot be applied when normalization has occurred. In the remainder of this section, we show how to calculate appropriate scaling parameters to preserve the differential privacy guarantee. Furthermore, since different location parameters induce different normalization factors, we allow for the scaling parameter to be calculated as a function of the position of the location parameter relative to the constraints in order to achieve lower scaling parameters and thus better utility.

5.1.2 Single Infinite Constraint

Having defined classes of range constraints and their corresponding normalization factors, we now study these classes in the context of our privacy guarantee. This form of the privacy guarantee captures the information leakage induced by data-dependent normalization, allowing us to investigate the impact that normalization has on the probabilities for any pair of databases to produce the same noisy query response. We show how it is still possible to achieve differential privacy as long as the scaling parameter of the Laplace distribution is carefully chosen to account for the additional information leakage induced by normalization. For conciseness, we provide sketches of the proofs for the lemmas and theorems used throughout this section. The full proofs related to the class of a single infinitely spanning constraint can be found in Appendix A1.

We begin with the class of single infinitely spanning constraints. In this class, there is a single span of infeasible space extending either to infinity or to negative infinity. The complement of this space is a single span of feasible space. Without loss in generality, we assume the constraint extends to infinity. The case of going to negative infinity can be seen as a horizontal reflection of this, in which all analysis is the same.

Since there is only a single constraint, the normalization factor can be calculated as the reciprocal of the integral of the truncated space. We apply this normalization factor to the privacy guarantee of Formula (5.6) along with the adaptation for arbitrary distances of Section 5.1.1 and the continuous random variable worst case in the third case of Formula (5.12). The privacy guarantee can now be written as:

$$\frac{\sigma_2}{\sigma_1} \left(\frac{2 - e^{-\frac{\Delta L_1 + i\Delta F}{\sigma_2}}}{2 - e^{-\frac{\Delta L_1}{\sigma_1}}} \right) \left(\frac{e^{-\frac{\Delta L_1 + i\Delta F}{\sigma_2}}}{e^{-\frac{\Delta L_1}{\sigma_1}}} \right) \leq e^{i\epsilon}. \quad (5.22)$$

The formulae in this section make use of the multi-valued LambertW function [36].

Definition 8. The **LambertW function** [36] is given by the inverse function of $f(x) = xe^x$ where x is a complex number.

The function is defined by multiple branches which map input values to output. An overlap in the domain of the branches gives rise to the multi-valued nature of the function. As we are interested only in real-valued output, we restrict our attention to the primary and -1 branches. We refer to these function branches as W and W_{-1} for shorthand and use W_Z for instances where either branch may be applied. The real-valued outputs of the function are determined by the primary branch of Formula (5.23), and the -1 branch of Formula (5.24). For the branches to be single-valued, the following conditions apply: $W(x) \geq -1$ and $W_{-1}(x) \leq -1$.

$$W(xe^x) = x \quad -\frac{1}{e} \leq xe^x. \quad (5.23)$$

$$W_{-1}(xe^x) = x \quad -\frac{1}{e} \leq xe^x < 0. \quad (5.24)$$

Lemma 1. *For any PDF with a location parameter at distance $i\Delta F > 0$ from the constraint, bounds on the possible values of its scaling parameter are determined by the following four inequalities:*

$$\sigma_2 \leq -\frac{i\Delta F e^{i\epsilon} \sigma_1}{W\left(-\frac{2i\Delta F e^{-i\epsilon} e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}}}{\sigma_1}\right) e^{i\epsilon} \sigma_1 + i\Delta F}, \quad (5.25)$$

$$\sigma_2 \geq -\frac{i\Delta F e^{i\epsilon} \sigma_1}{W_{-1}\left(-\frac{2i\Delta F e^{-i\epsilon} e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}}}{\sigma_1}\right) e^{i\epsilon} \sigma_1 + i\Delta F}, \quad (5.26)$$

$$\sigma_2 \geq \frac{i\Delta F}{W\left(\frac{-1}{2e}\right) i\epsilon e^{1-i\epsilon} - W\left(2W\left(\frac{-1}{2e}\right) i\epsilon e^{W\left(\frac{-1}{2e}\right) i\epsilon e^{1-i\epsilon} - i\epsilon + 1}\right)}, \quad (5.27)$$

$$\sigma_2 \leq \frac{i\Delta F}{W\left(\frac{-1}{2e}\right) i\epsilon e^{1-i\epsilon} - W_{-1}\left(2W\left(\frac{-1}{2e}\right) i\epsilon e^{W\left(\frac{-1}{2e}\right) i\epsilon e^{1-i\epsilon} - i\epsilon + 1}\right)}. \quad (5.28)$$

Proof. By isolating σ_2 in Formula (5.22), we obtain an inequality which contains the LambertW function. Since the LambertW function has two branches, this gives two restrictions on the possible values that the σ_2 values can take on. By setting $\Delta L_1 = 0$, we determine that σ_2 must fall in the intersection of the two spans specified in Formulae (5.25) and (5.26) in order for the privacy guarantee to be satisfied for a pair of PDFs with location parameters at distances of 0 and $i\Delta F > 0$ away from the constraint.

Since the LambertW function does not produce real-valued output for any input values less than $-\frac{1}{e}$, we are also able to derive restrictions on the allowable σ_2 values that produce valid input for the LambertW function. For this, we first study the modality of the input as a function of the distance variable i by taking the derivative of the input with respect to i , as shown in Formula (5.29).

$$-\frac{\frac{2i\epsilon\sigma_1 e^{\frac{\Delta L_1}{\sigma_1}} + e^{-i\epsilon} i\Delta F - i\epsilon\sigma_1 + e^{-i\epsilon} \Delta L_1}{\sigma_1 \left(2e^{\frac{\Delta L_1}{\sigma_1}} - 1\right)}}{(\sigma_1)^2 \left(2e^{\frac{\Delta L_1}{\sigma_1}} - 1\right)^2 \left(-2\sigma_1 e^{\frac{\Delta L_1}{\sigma_1}} + (i\Delta F + \Delta L_1) e^{-i\epsilon} + \sigma_1\right)^{-1}}. \quad (5.29)$$

From the factors in the derivative, we determine that zeros can only occur at the values of i shown in Formulae (5.30) and (5.31).

$$i = -\frac{\Delta L_1 \epsilon + \Delta F W_Z\left(-\frac{\epsilon\sigma_1 \left(2e^{\frac{\Delta L_1}{\sigma_1}} - 1\right) e^{-\frac{\Delta L_1 \epsilon}{\Delta F}}}{\Delta F}\right)}{\epsilon \Delta F}. \quad (5.30)$$

$$i = -\frac{\Delta L_1 \epsilon - \Delta F}{\epsilon \Delta F}. \quad (5.31)$$

The value of the input at the two zeros (determined by the two branches of the LambertW function) of Formula (5.30) both simplify to $-\frac{2}{e}$, placing them outside of the valid range of input. The zero of Formula (5.31) is valid and is a minimum of the resultant unimodal function of the input. Using this result, we are able to assert that if the function of input to the LambertW function is greater than or equal to $-\frac{1}{e}$ at this zero, then this also holds for all other values of i . We therefore begin from the restriction on the LambertW input as shown in Formula (5.32) where the input is the left-hand side of the inequality.

$$-\frac{2i\epsilon\sigma_1 e^{\frac{\Delta L_1}{\sigma_1}} + e^{-i\epsilon i\Delta F - i\epsilon\sigma_1 + e^{-i\epsilon\Delta L_1}}}{\sigma_1 \left(2e^{\frac{\Delta L_1}{\sigma_1}} - 1\right)} - \frac{2(i\Delta F + \Delta L_1)e}{\sigma_1 \left(2e^{\frac{\Delta L_1}{\sigma_1}} - 1\right)} \geq -\frac{1}{e}. \quad (5.32)$$

We substitute the instances of i with the zero of Formula (5.31) and isolate σ_1 . Since we are interested in bounds on σ_2 , we then substitute all instances of ΔL_1 with $i\Delta F$. This allows us to interpret σ_1 as the required scaling parameter at distance $i\Delta F$ from the constraint boundary. This corresponds to the location of σ_2 as interpreted in Formulae (5.25) and (5.26) and therefore acts as two further constraints on a scaling parameter at this location. The result of the substitutions is the pair of additional inequalities given in Formulae (5.27) and (5.28). $\square$

Lemma 2. *Through the selection of an appropriate σ_1 value when $\Delta L_1 = 0$, it is possible to calculate σ_2 values for any PDF with a location parameter at distance $i\Delta F > 0$ away from the constraint such that the inequalities of Lemma 1 are satisfied.*

Proof. Following the same approach as in Lemma 1, we assert that the input to the LambertW function in Formulae (5.25) and (5.26) must be greater than or equal to $-\frac{1}{e}$. Furthermore, as a function of i , the input is unimodal, with the mode being a minimum. Since we are interested in bounds on σ_1 when $\Delta L_1 = 0$, we substitute i with the value at which the minimum occurs, set ΔL_1 to 0, and then isolate for σ_1 to obtain the inequalities of Formulae (5.33) and (5.34).

$$\sigma_1 \geq -\frac{\Delta F}{W\left(-\frac{1}{2e}\right)\epsilon\epsilon}. \quad (5.33)$$

$$\sigma_1 \leq -\frac{\Delta F}{W_{-1}\left(-\frac{1}{2e}\right)\epsilon\epsilon}. \quad (5.34)$$

We then select the lower bound given by Formula (5.33) as the value to use for σ_1 when $\Delta L_1 = 0$, giving us the calculation shown in Formula (5.35). Intuitively, it is desirable to select the lowest possible values for the scaling parameters in order to minimize the amount of noise added by the resultant PDF. One might ask whether this choice for σ_1 forces the values calculated for σ_2 (namely, scaling parameters at other locations) to be higher than they could otherwise be. We later justify this choice by showing that we calculate the lowest σ_2 values that satisfy the privacy guarantee.

$$\sigma_1 = -\frac{\Delta F}{W\left(-\frac{1}{2e}\right)\epsilon\epsilon}. \quad (5.35)$$

Finally, by substituting this value for σ_1 into Formulae (5.25) and (5.26), the right-hand sides of the inequalities become equivalent to the right-hand sides of Formulae (5.27) and (5.28), respectively. As a result, in order to satisfy all four inequalities of Lemma 1, a PDF with a location parameter at distance $i\Delta F > 0$ from the constraint can be assigned one of the two possible σ_2 values (by setting Z to either 0 or -1) calculated using Formula (5.36).

$$\sigma_2 = -\frac{i\Delta F e^{i\epsilon}\sigma_1}{W_Z\left(-\frac{2i\Delta F e^{-i\epsilon}e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}}}{\sigma_1}\right)e^{i\epsilon}\sigma_1 + i\Delta F}. \quad (5.36)$$

□

In the next four lemmas, we identify properties of the σ_2 calculation of Lemma 2 as a function of the distance variable i . These properties are of use in proving the correctness and optimality of our proposed σ calculations.

Lemma 3. *The sign of the denominator in the derivative taken with respect to i of the σ_2 calculation of Lemma 2 depends on which branch is indicated by the branch index variable Z .*

Proof. We first take the derivative of the σ_2 calculation (Formula (5.36)) with respect to i to obtain Formula (5.37).

$$\frac{-\left(W_Z\left(-\frac{2i\Delta F e^{-i\epsilon}e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}}}{\sigma_1}\right) + i\epsilon\right)\Delta F e^{i\epsilon}\sigma_1}{\left(W_Z\left(-\frac{2i\Delta F e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}}}{\sigma_1 e^{i\epsilon}}\right)e^{i\epsilon}\sigma_1 + i\Delta F\right)\left(W_Z\left(-\frac{2i\Delta F e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}}}{\sigma_1 e^{i\epsilon}}\right) + 1\right)}. \quad (5.37)$$

The first parenthesized factor in the denominator is always negative due to the fact that its first term is a negative value that is always of greater magnitude than its second positive term of $i\Delta F$. The sign of the second parenthesized factor in the denominator is determined by the branch used for the LambertW function. Since the output of the principal branch of a LambertW function is always greater than or equal to -1, this factor will always be positive when Z is 0. Whereas, if Z is -1 then the output of the LambertW function is always less than or equal to -1, making the sign of the factor negative. As a result, the overall sign of the denominator will always be negative for the principal branch and will always be positive for the -1 branch.

□

Lemma 4. *As a function of i , the σ_2 calculation of Lemma 2 is unimodal for either branch index.*

Proof. Since the sign of the denominator of its derivative cannot change when restricted to the same branch index, a sign change could only be induced by the numerator. The variables outside of the parenthesized factor of the numerator will always have positive values. We therefore check for the presence of a zero in the parenthesized factor to determine if a sign change occurs in the numerator. We take the two terms of the factor and check for when they are equal in magnitude but opposite in their sign. By isolating for i in this equation, we obtain Formula (5.38).

$$i = -\frac{W_Z\left(\frac{\ln\left(\frac{\epsilon\sigma_1}{2\Delta F}\right)\epsilon\sigma_1}{\Delta F}\right)}{\epsilon}. \quad (5.38)$$

Since we are considering each branch individually, the value of Z in this equation is fixed for whichever branch we are currently using. Therefore, there is exactly one sign change in the derivative

per branch. As a result, the σ_2 calculation of Formula (5.36) is unimodal as a function of i for each branch of the LambertW function. $\square$

Lemma 5. *As a function of i , the input to W_Z in the σ_2 calculation of Lemma 2 is unimodal and is initially decreasing.*

Proof. When $i = 0$, the input is equal to 0 and is initially decreasing as i increases. We take the derivative of the input as a function of i , as shown in Formula (5.39), in order to determine its modality.

$$-\frac{2\Delta F e^{-i\epsilon} e^{-\frac{i\Delta F e^{-i\epsilon}}{\sigma_1}} (i\epsilon - 1) (i\Delta F e^{-i\epsilon} - \sigma_1)}{\sigma_1^2}. \quad (5.39)$$

Since the unparenthesized portion of the numerator and the denominator are all variables having positive values, they cannot induce a sign change. For the two parenthesized factors in the numerator, there can be a sign change in the first but not in the second. Therefore, the derivative has exactly one sign change at $i = \frac{1}{\epsilon}$, making the function unimodal. $\square$

Lemma 6. *As a function of i , the mode of the σ_2 calculation of Lemma 2 is a minimum for the principal branch and a maximum for the -1 branch.*

Proof. From Lemma 3, we know that the sign of the denominator in the derivative of Formula 5.36 with respect to i will always be negative for the principal branch and will always be positive for the -1 branch. From Lemma 4, we know that a single sign change occurs in the numerator. Based on the properties of the input to the LambertW function determined in Lemma 5, the numerator of the derivative is initially positive until its zero, from which point it is negative. This results in the mode being a minimum for the principal branch and a maximum for the -1 branch. $\square$

Theorem 1. *The σ calculation method of Lemma 2 provides a solution which optimally satisfies the differential privacy guarantee.*

Proof. To show that the worst-case analysis of Section 5.1.1 holds, we show that the σ values are monotonically decreasing as i increases. When using the σ_1 value of Formula (5.35), the mode specified in Formula (5.38) reduces to $i = \frac{1}{\epsilon}$ for both branches of Formula (5.36). By Lemma 6, the principal branch is decreasing until its mode and that the -1 branch is decreasing after its mode. Thus, by using the principal branch to calculate the values of σ_2 prior to $i = \frac{1}{\epsilon}$ and the -1 branch after this point, all σ values will be monotonically decreasing as i increases.

Since the privacy guarantee of Formula 5.22 is non-symmetric, we also apply the σ calculations to an alternate formulation which represents the symmetric case and we show that this formulation holds. Finally, we must consider the optimality of the σ values. Since lower σ values are preferable, we show through further examination of the four inequalities of Lemma 1 that it is not possible to calculate lower σ_2 values that could satisfy the privacy guarantee. This justifies the choice in Lemma 2 to use the smallest valid σ_1 value. $\square$

5.1.3 Arbitrary Finite Constraints

In this configuration, we have an arbitrary set of finite spans of infeasible space. The complement of these spans is the feasible space, thus it extends to both positive and negative infinity. Once again,

we provide sketches of the proofs here. The full proofs for this class of constraints can be found in Appendix A2.

As stated in Section 5.1.1, since we have no constraints that go to infinity, the continuous random variable worst-case analysis requires that we use the same scaling parameter everywhere. The normalization factors that we use here correspond to summations using finite integrals. We start by considering pairs of PDFs with location parameters in the same span of truncated space. Since the scaling parameters are all the same, we can apply the identities $L_2 = L_1 e^{\frac{i\Delta F}{\sigma}}$ and $R_2 = R_1 e^{-\frac{i\Delta F}{\sigma}}$ to produce the privacy guarantee shown in Formula (5.40). Note that since there are no infinite constraints present, the indicator variables for such constraints inside L_1 and R_1 are equal to 0.

$$\frac{1 - \left(L_1 e^{\frac{i\Delta F}{\sigma}} + R_1 e^{-\frac{i\Delta F}{\sigma}} \right)}{1 - (L_1 + R_1)} \left(e^{\frac{i\Delta F}{\sigma}} \right) \leq e^{i\epsilon}. \quad (5.40)$$

For pairs of PDFs with location parameters in different spans of truncated space, the sets of constraints to their left and right will differ. This would require each normalization factor to be written in terms of its own sets of constraints. As we will see later, there is only one specific case of PDFs with location parameters in different spans of truncated space that we must explicitly consider in order to prove the guarantee for any pair of databases.

Lemma 7. *For each span of truncated space, there exists a value of σ for which the privacy guarantee is satisfied for any pair of PDFs with location parameters within that span.*

Proof. When comparing the left and right-hand sides of the privacy guarantee for $i = 0$, both sides are equal to 1. In order for the inequality of the privacy guarantee to be satisfied, it is therefore a necessary condition that the left-hand side is increasing at a lower rate than the right-hand side. We show this in two steps. We first take the derivative with respect to i of each side of the privacy guarantee and form a new inequality which asserts that the derivative of the left-hand side must be less than or equal to the derivative of the right-hand side. By setting the value of i to 0 and rearranging the inequality, we identify a condition on the value of σ shown in Formula (5.41). This condition ensures that the left-hand side of the privacy guarantee is growing less quickly than the right-hand side at $i = 0$.

$$\sigma \geq \frac{\Delta F}{\epsilon - \frac{\Delta F(L_1 - R_1)}{\sigma(L_1 + R_1 - 1)}}. \quad (5.41)$$

Although this form does not isolate σ , it can still be interpreted as an identity for a lower bound on the value of σ . We then show that by setting σ to this lower bound, the privacy guarantee is achieved. For this, we use the lower bound to substitute the instance of σ appearing in the parenthesized factor of $\left(e^{\frac{i\Delta F}{\sigma}} \right)$ from the privacy guarantee of Formula (5.40). Despite the fact that only one of the instances of σ has been substituted, we show that it is now possible to prove the guarantee regardless of value of these other instances. We first rearrange and simplify the inequality to obtain Formula (5.42).

$$\frac{1 - \left(L_1 e^{\frac{i\Delta F}{\sigma}} + R_1 e^{-\frac{i\Delta F}{\sigma}} \right)}{1 - (L_1 + R_1)} \leq e^{\frac{i\Delta F(L_1 - R_1)}{\sigma(L_1 + R_1 - 1)}}. \quad (5.42)$$

Since the use of the lower bound causes the derivatives of the left and right-hand sides of the privacy guarantee to be equal to each other and to have the same derivative at $i = 0$, we must then show that the second derivative of the left-hand side is less than or equal to the second derivative of

the right-hand side. This second step will ensure that the right-hand side grows more quickly. The inequality of the second derivatives is shown in Formula (5.43).

$$\frac{\Delta F^2 \left(L_1 e^{\frac{i\Delta F}{\sigma}} + R_1 e^{-\frac{i\Delta F}{\sigma}} \right)}{\sigma^2 (L_1 + R_1 - 1)} \leq \frac{\Delta F^2 (L_1 - R_1)^2}{\sigma^2 (L_1 + R_1 - 1)^2} \left(e^{\frac{i\Delta F (L_1 - R_1)}{\sigma (L_1 + R_1 - 1)}} \right). \quad (5.43)$$

Since the variables L_1 and R_1 represent the summations of the integrals of the finite constraints to the left and right, respectively, of the location parameter $f(D_1)$, both of them have a range of $[0, 0.5]$. When L_1 and R_1 are restricted to these ranges, the expression on the left-hand side will always be negative and the expression on the right-hand side will always be positive. This means that the derivative of the left-hand side is decreasing and the derivative of the right-hand side is increasing. As such, the left-hand side will always be less than or equal to the right-hand side, meaning that the guarantee is satisfied when using the σ value taken as the lower bound from Formula (5.41). $\square$

Lemma 8. *Within each span of truncated space, the σ value determined from Lemma 7 acts as a lower bound for the value of σ required to satisfy the privacy guarantee.*

Proof. Starting from the value of σ obtained by using Formula (5.41) as an equality (as was done in Lemma 7), we can represent a larger value of σ by subtracting some value α from the denominator such that α is not greater than or equal to the original value of the denominator. This form is shown in Formula (5.44)

$$\sigma \geq \frac{\Delta F}{\epsilon - \frac{\Delta F (L_1 - R_1)}{\sigma (L_1 + R_1 - 1)} - \alpha}. \quad (5.44)$$

Taking this lower bound as the value to use for σ , we substitute this into the privacy guarantee, as was done in Lemma 7, to obtain the inequality in Formula (5.45).

$$\frac{1 - \left(L_1 e^{\frac{i\Delta F}{\sigma}} + R_1 e^{-\frac{i\Delta F}{\sigma}} \right)}{1 - (L_1 + R_1)} \leq e^{\frac{i\Delta F (L_1 - R_1)}{\sigma (L_1 + R_1 - 1)} + i\alpha}. \quad (5.45)$$

The only difference between Formula (5.42) and Formula (5.45) is that the exponent of the right-hand side has become larger, making the guarantee easier to satisfy. It therefore follows that any σ value larger than that which was determined in Lemma 7 will also satisfy the privacy guarantee. $\square$

Lemma 9. *For each finite constraint, there exists a value of σ that satisfies the privacy guarantee for the pair of PDFs with location parameters on the endpoints of the constraint.*

Proof. The guarantee form employed thus far has used the same sets of finite constraints to the left and right of both location parameters, implying that they must both lie within the same span of truncated space. We must also be able to show that the privacy guarantee holds for location parameters in different spans of truncated space. To show this, we first consider a pair of PDFs with location parameters that lie on opposite endpoints of a finite constraint (with $f(D_1)$ as always being the point on the right). We provide this instantiation of the privacy guarantee in Formula 5.47, using Formula 5.46 to calculate the value of the integral corresponding to the span covered by the finite constraint. Since both location parameters are adjacent to this span, the value of the integral is the same for both PDFs.

$$S = \frac{1 - e^{-\frac{i\Delta F}{\sigma}}}{2}. \quad (5.46)$$

$$\frac{1 - \left(e^{\frac{i\Delta F}{\sigma}} (L_1 - S) + R_1 e^{-\frac{i\Delta F}{\sigma}} + S \right)}{1 - (L_1 + R_1)} \left(e^{\frac{i\Delta F}{\sigma}} \right) \leq e^{i\epsilon}. \quad (5.47)$$

We know from Lemma 7 that a sufficiently high value of σ can satisfy the guarantee without the modification made here and from Lemma 8 that raising the value of σ beyond the requirement will not cause that form of the privacy guarantee to be violated. Increasing σ reduces the influence of S in Formula 5.47 by causing the value of S to asymptotically approach 0. It therefore follows that a sufficiently high value of σ will also satisfy this form of the privacy guarantee and that increasing σ beyond that value will not violate the guarantee. $\square$

Lemma 10. *All lower bounds on σ identified in Lemmas 7 and 9 are less than $\frac{2\Delta F}{\epsilon}$.*

Proof. The calculations for lower bounds on σ can be handled by treating Formulae (5.41) and (5.47) as equalities and solving for σ . We can identify bounds on the possible values of σ by studying the bounds on the variables L_1 and R_1 . Since L_1 represents the sum of the integrals of the constraints to the left of $f(D_1)$ it can be as low as 0 (if no constraints are present to the left of $f(D_1)$) and can approach but not reach 0.5 (since half of the integral exists on the left hand side). The bounds on R_1 are characterized in the same way. Based on these bounds, the calculations for σ will fall in the range of $(0, \frac{2\Delta F}{\epsilon})$. $\square$

Theorem 2. *The optimal σ value that satisfies the privacy guarantee for all pairs of databases can be found by taking the maximum out of $3n + 2$ lower bounds, where n is the number of finite constraints.*

Proof. Lemma 7 provides a lower bound on σ for pairs of PDFs with location parameters in the same span of truncated space. This applies to a form of the privacy guarantee in which $f(D_2) \leq f(D_1)$ holds. The symmetric case can be handled in the same way after an application of horizontal reflection to the configuration. Lemma 9 provides an additional lower bound on σ for PDFs with location parameters on opposite endpoints of a constraint. In Formula (5.47), if the fraction on the left-hand side is inversed, as it would be in a symmetric form, the left-hand side decreases. Thus, if the form in Formula (5.47) is satisfied, the symmetric form will be as well. For n constraints, there are $n + 1$ lower bounds on σ in Lemma 7 from the regular form of the privacy guarantee and an additional $n + 1$ lower bounds for the symmetric form. From Lemma 9, there are n lower bounds, giving a total of $3n + 2$. By selecting the largest of these, the guarantee is satisfied for all pairs of PDFs with location parameters in any same span of truncated space and for all pairs of PDFs with location parameters on opposite endpoints of a constraint. Since each of the lower bounds must be adhered to, it is not possible to select a lower value of σ than this. The chosen value of σ is therefore optimal assuming it also satisfies the privacy guarantee for arbitrary pairs of databases.

The protection of arbitrary pairs of databases follows from the transitive property of multiple applications of the guarantee forms used throughout the lemmas. For any pair of PDFs with location parameters at arbitrary points in truncated space, it is possible to represent this as a sequence of points where each adjacent pair of points in the sequence corresponds to a pair of PDF location parameters in the configuration used in either Lemma 7 or Lemma 9. The multiplicative bound for the arbitrary pair is therefore the product of the bounds of the adjacent pairs in the sequence. Since

each of the adjacent pairs satisfy the privacy guarantee, the product will also satisfy the guarantee for the arbitrary pair. $\square$

Theorem 3. *The optimal value of σ for any configuration of n arbitrary finite constraints can be calculated to a precision of d decimal places in $O\left(n^2\left(d + \log\left(\frac{\Delta F}{\epsilon}\right)\right)\right)$ time.*

Proof. As stated in Theorem 2, σ must be chosen as the maximum value out of the $O(n)$ lower bounds calculated from Formulae (5.41) and (5.47). The variables L_1 and R_1 in these inequalities represent summations of $O(n)$ exponential functions where each function contains an instance of σ in the denominator of its exponent. We know of no method to isolate σ for such configurations. It therefore takes $O(n^2)$ time to check whether a given value of σ is above all lower bounds.

From Lemmas 8 and 9, we know that any value of σ larger than the required value will also satisfy the privacy guarantee. This implies that once the inequality is satisfied, increasing σ further will never violate the inequality. Lemma 10 indicates that the value of σ will always be between 0 and $\frac{2\Delta F}{\epsilon}$, meaning that for a decimal precision of d , there are $\frac{10^d(2)\Delta F}{\epsilon}$ possible values for σ . By performing a binary search for the optimal value, the logarithm of the number of possible values of σ must be checked, leading to an overall time complexity of $O\left(n^2\left(d + \log\left(\frac{\Delta F}{\epsilon}\right)\right)\right)$. $\square$

Corollary 1. *The computational complexity for the calculation of the optimal value of σ for any configuration of n arbitrary finite constraints can be more practically interpreted as $O(n^2)$.*

Proof. From Theorem 3, the complexity is $O\left(n^2\left(d + \log\left(\frac{\Delta F}{\epsilon}\right)\right)\right)$. In many practical settings, when d and $\frac{\Delta F}{\epsilon}$ are small constant numbers, an optimal value for σ can be determined in $O(n^2)$ time. $\square$

5.1.4 Arbitrary Constraints

While the privacy guarantee forms in the configurations covered thus far have been manageable in their complexity, configurations with multiple constraints where at least one spans infinitely lead to much more complex expressions due to an increase in the number of exponential functions that appear in the normalization factors combined with variable scaling parameters. The expressions that must be worked with take the form of transcendental functions having differing polynomials as coefficients and exponents. Exact calculations necessary for determining optimal scaling parameter values now become very difficult, if not impossible, due to the necessary step of root-finding for these functions. Unlike the case of the single infinite constraint, there is no function such as the LambertW that can be easily substituted in these expressions to allow for solutions to be calculated.

When the number of constraints is small, it may still be possible to apply some form of analysis to produce very good approximate solutions which can approach the optimal values within a desired level of precision. We leave the analysis of such configurations as an open problem. For cases with greater numbers of constraints, it is likely that the analysis becomes unmanageable. For such cases, we thus assign the same scaling parameters for every PDF instead of considering variable scaling parameters. In essence, we sacrifice the better utility obtained by the calculation of variable scaling parameters for the ability to calculate the optimal value when using the same scaling parameter for all PDFs such that we are able to satisfy the privacy guarantee. The corresponding privacy guarantee is shown in Formula (5.48). This is a modified version of Formula (5.10) using same scaling parameters and the appropriate continuous random variable worst-case analysis of Section 5.1.1.

$$\frac{1 - \left(L_1 e^{\frac{i\Delta F}{\sigma}} + R_1 e^{-\frac{i\Delta F}{\sigma}}\right)}{1 - (L_1 + R_1)} \left(e^{\frac{i\Delta F}{\sigma}}\right) \leq e^{i\epsilon}. \quad (5.48)$$

Theorem 4. *When at least one infinitely spanning constraint is present, the optimal uniform value of σ for a configuration of arbitrary constraints is $\frac{2\Delta F}{\epsilon}$.*

Proof. The calculation of σ and the related proofs required here are almost identical to those of Section 5.1.3. The only difference is that L_1 and R_1 may now contain integrals of constraints that span to negative infinity and infinity, respectively. Due to this, it is now possible for L_1 and R_1 to be equal to 0.5, whereas before they could only approach this value. Note however that they cannot be simultaneously equal to 0.5 as this would imply that no feasible space exists. This implies that rather than the upper bound on σ being one that approaches $2\Delta F/\epsilon$, it is now equal to this value. The algorithm for the approximation of the optimal value of σ from Theorem 3 can thus be employed here as well.

Consider when the algorithm treats the inequality of the form stated in Formula (5.41) for the span of feasible space adjacent to an infinitely spanning constraint. Without loss of generality, assume the constraint spans to positive infinity. Since the value of R_1 will be 0.5 in this case, the inequality simplifies to the form stated in Formula (5.49).

$$\sigma \geq \frac{\Delta F}{\epsilon - \frac{\Delta F}{\sigma}}. \quad (5.49)$$

Since σ must be at least $\frac{\Delta F}{\epsilon}$, this simplified inequality indicates that we must use $\frac{2\Delta F}{\epsilon}$ as the scaling parameter. $\square$

We note here a close relationship between our result in this setting and the exponential mechanism (Definition 5, [135]).

Corollary 2. *The use of our truncated and normalized Laplace mechanism for arbitrary constraints corresponds to an instantiation of the exponential mechanism.*

Proof. One can use the exponential mechanism in a range-adherent manner by restricting the domain of the quality function (and thus the range of the mechanism) to the feasible space for noisy responses and scaling the distribution by the corresponding normalization factor for this space. When the quality function is set to the negative distance between the true response and the noisy response, the exponential mechanism becomes a Laplace distribution and provides 2ϵ differential privacy [135]. Therefore, to achieve ϵ differential privacy, the scaling parameter must be set twice as high, at $\frac{2\Delta F}{\epsilon}$. This corresponds exactly to the configuration of our truncated and normalized Laplace mechanism for arbitrary constraint configurations. $\square$

Corollary 3. *For constraint settings with an arbitrary set of finite constraints or a single infinitely spanning constraint, our proposed mechanisms require lower scaling parameters than the exponential mechanism to achieve ϵ differential privacy. Specifically:*

- *The exponential mechanism requires a scaling parameter of $\frac{2\Delta F}{\epsilon}$*
- *Our mechanism for arbitrary finite constraints requires a scaling parameter strictly less than $\frac{2\Delta F}{\epsilon}$*
- *Our mechanism for a single infinitely spanning constraint requires a scaling parameter less than or equal to $\frac{1.586\Delta F}{\epsilon}$*

Proof. These results follow directly from Corollary 2 for the exponential mechanism, Lemma 10 for the setting of arbitrary finite constraints, and Lemma 2 and Theorem 1 for the setting of a single infinitely spanning constraint. Note that $\frac{1.586\Delta F}{\epsilon}$ is simply the evaluation of Formula (5.35) using a rounded constant in place of the LambertW function to allow for a more intuitive interpretation. $\square$

5.1.5 Summary of Results

Through the study of query range constraints, we have presented a framework for the design of mechanisms that allow for the truncation and normalization of the Laplace PDF according to the configuration of constraints. We have provided a detailed analysis of the impact of constraints on the mechanisms in our framework. An important characteristic of this framework is that it maintains the ability to apply an arbitrary discretization to the truncated range of the mechanisms.

We have provided an adaptation of the differential privacy guarantee which accounts for the data-dependent operation of normalization and enforces differential privacy in this context. Using this, we have applied our method of design to various classes of constraint configurations. For each mechanism, we have proven its correctness and where applicable, we have also proven optimality with respect to the calculation of the scaling parameters. Information on the calculation of scaling parameters and their bounds is summarized in Table 5.2 for each of the constraint configuration classes that we have studied.

Table 5.2: Summary of information on σ for each constraint configuration.

Constraint Configuration	σ Calculation	Bound(s)	Notes
Single Infinite Constraint	σ_1 : Formula (5.35) σ_2 : Formula (5.36)	$\sigma_1 \approx \frac{1.586\Delta F}{\epsilon}$ $\sigma_2 < \frac{1.586\Delta F}{\epsilon}$	Optimal σ values
Arbitrary Finite Constraints	Section 5.1.3	$\sigma < \frac{2\Delta F}{\epsilon}$	Approximation of optimal σ values
Arbitrary Constraints	Section 5.1.4	$\sigma = \frac{2\Delta F}{\epsilon}$	Optimal uniform σ values

5.2 Linear Programming Mechanisms

In this section, we consider an alternate approach to the design of range-adherent, interactive mechanisms for general numeric queries. We study the design of mechanisms that use a matrix representation for a discretized probability density function. Within this setting, we apply linear programming to produce range-adherent mechanisms that optimize a user-independent criterion for utility.

We open the section with a discussion on the desirable properties of range-adherent mechanisms. We then formalize a matrix representation of range-adherent mechanisms. We redefine boundary-snapping within this representation and study some of the properties it provides. Finally, we propose a linear programming approach for the creation of range-adherent mechanisms to optimize a user-independent criterion for utility.

5.2.1 Design Considerations

We propose that the design of interactive mechanisms should incorporate publicly known range constraints of queries by restricting the noisy output to fall within the valid range of the query. Our goal is to design more useful mechanisms in terms of downstream compatibility and utility. There are a number of ways to approach this task and the design decisions will impact the resultant utility of the mechanism. It is important to first consider the intended meaning of the term “utility”. We therefore begin this section with a theoretical discussion on the design of range-adherent interactive mechanisms. We then describe the impact the design decisions have on utility.

The utility of a mechanism may refer to either information preservation or usability [123]. The former refers to how much of the original information is carried in the noisy responses, whereas the latter refers to how directly useful the noisy responses are to a particular user. The axiom of sufficiency [123] states that for any mechanism M_1 that can be composed with an arbitrary function to simulate another mechanism M_2 , the information preservation of M_1 is greater than or equal to that of M_2 . This is because any inferences that could be drawn from the behaviour of M_2 could equally be drawn from M_1 by simulating M_2 if desired. An important consequence of this axiom is that although post-processing of the noisy responses (without further access to the database) may improve usability, it cannot improve information preservation and may even be detrimental in that regard.

A commonly used method of adherence to range constraints is the post-processing step of boundary-snapping in which out-of-bounds query responses are snapped to the nearest valid value [86, 124]. By the axiom of sufficiency, incorporation of boundary-snapping into a mechanism produces a new mechanism with degraded information preservation. The non-constrained mechanism can simulate the one that uses boundary-snapping but the opposite is not true. Although adherence to range constraints is achieved, there is no other consideration made for how this affects the utility of the mechanism.

An alternative method of post-processing that achieves adherence to range constraints is Bayesian post-processing [77]. This involves remapping noisy query responses to valid values in order to optimize a user-specific measure of utility. This form of post-processing relies on the specification of the user’s prior knowledge of the sensitive data (represented as a probability distribution over the valid query responses) and the user’s utility goal (represented as a loss function). Although this improves the usability of the responses for the target user, its reliance on specific user settings causes a degradation in information preservation which can negatively impact users operating under different settings. As such, this should be kept as a separate step which is either handled independently by the user or is provided as an optional operation on a per-user basis.

Although Bayesian post-processing applied as a separate step may be an appropriate solution in some situations, there are also scenarios where it cannot be applied. If the mechanism is used as a building block in a larger system that does not allow for the specification of additional user settings, post-processing cannot be properly applied. The use of default settings can be detrimental for the utility of many users and thus should not be used. In other cases, users may simply not yet be aware of the ideal objective function to apply for their data analysis and are thus unable to apply post-processing. Furthermore, only the remapping process is optimal with respect to the user’s objective function; the underlying mechanism provides no such guarantee and may not be well-suited to a range-constrained setting. Ideally, a mechanism should be able to adhere to range constraints while optimizing a general goal for utility that is independent of user settings.

5.2.2 Formalization of Range-Adherent Mechanisms

To facilitate the incorporation of range-adherence into mechanism design, we next provide a formal representation of range-adherent mechanisms. To this end, we employ a matrix representation of mechanisms along with a set of requirements for range-adherence. We also consider the inclusion of additional, optional constraints on the matrix which may be conducive to higher levels of utility.

Mechanism PDF Discretization

In practice, the range of a query may correspond to a particular subset of $\mathbb{R}$. We refer to such a correspondence as query range constraints. We wish to define requirements for the adherence of

mechanisms to query range constraints. For this, we first consider the treatment of the underlying PDF used by the mechanism to generate noise. We discretize the range of the PDF such that it can be mapped to a set of discretized response categories. Each discretized response category represents a range of values such that reporting the category indicates a noisy query response that falls within the corresponding range. Let $P_{K(f(D))} : \mathbb{R} \rightarrow \mathbb{R}$ be the PDF used by the mechanism K for a query f posed on a database $D \in \mathbb{D}$ and let $\mathfrak{R} \in \mathbb{R}$ be the finite, continuous range that we wish to discretize. We define r_l and r_u to be the lower and upper bounds, respectively, of $\mathfrak{R}$.

Definition 9. Let R be a set of **discretized response categories** over a range $\mathfrak{R}$. Each $R_i \in R$ is a finite, continuous range, having lower and upper bounds r_i^l and r_i^u respectively such that $r_l \leq r_i^l \leq r_i^u \leq r_u$. The probability of a noisy response $R_i \in R$ is given by:

$$\Pr(K(f(D)) \in R_i) = \int_{r_i^l}^{r_i^u} P_{K(f(D))}(x) dx. \quad (5.50)$$

Matrix Representation

The discretization of a mechanism's PDF allows for the representation of the mechanism as a matrix having true query responses as columns, noisy query response categories as rows and conditional probabilities as the matrix entries. The matrix representation provides a convenient format for considering mechanism modifications for adherence to query range constraints.

Definition 10. Let $F \subset \mathbb{R}$ be the set of true responses for a query f and let R be the set of discretized noisy response categories. A **mechanism matrix** M is an $|R| \times |F|$ matrix such that its entries $M_{i,j}$ correspond to the conditional probability distribution, defined by a mechanism K as follows:

$$M_{i,j} = \Pr(R_i | F_j) = \Pr(K(F_j) \in R_i), \quad i = 1, \dots, |R|, \quad j = 1, \dots, |F|. \quad (5.51)$$

In order for a mechanism matrix to be considered a valid differentially private mechanism, it must satisfy the differential privacy guarantee and provide a column for each possible true query response such that each of these columns constitutes a valid probability distribution. To be considered range-adherent, we additionally assert that the mechanism must use a discretization of noisy response categories that is finite and that covers all of the true query responses. We formalize these requirements for a mechanism matrix as follows:

- **R.1** The conditional probability distribution must be differentially private:

$$M_{i,j} \leq e^\epsilon M_{i,k}, \quad \forall D_1, D_2 \in \mathbb{D} \text{ such that } f(D_1) = F_j, f(D_2) = F_k, D_1 \sim D_2. \quad (5.52)$$

- **R.2** The noisy response categories must be disjoint and their union must be finite and must contain all of the true query responses:

$$R_i \cap R_j = \emptyset, \quad i = 1, \dots, |R|, \quad j = 1, \dots, |R| \text{ such that } i \neq j, \quad (5.53)$$

$$r_l \neq -\infty, \quad r_u \neq \infty, \quad (5.54)$$

$$F_j \in \bigcup_{i=1}^{|R|} R_i, \quad j = 1, \dots, |F|. \quad (5.55)$$

- **R.3** Each true query response must correspond to a column representing a valid probability distribution:

$$M_{i,j} \geq 0, \quad i = 1, \dots, |R|, \quad j = 1, \dots, |F|, \quad (5.56)$$

$$\sum_{i=1}^{|R|} M_{i,j} = 1, \quad j = 1, \dots, |F|. \quad (5.57)$$

We note that in some cases, one might additionally require that the union of all the noisy query responses covers the full span of the chosen range $\mathfrak{R}$. We omit this requirement for the sake of a more general representation as it may be desirable in some cases to allow for some portions of the range to never be reported. For example, with counting queries, it is reasonable to only report integer values as the noisy query responses.

Optional Mechanism Constraints

The matrix representation also allows for the inspection of further mechanism properties which may help to improve utility in certain cases. We use the notation i_j to indicate the index of the noisy response category in which the true query response of index j falls (i.e., $F_j \in R_{i_j}$). We now consider properties that have been previously studied in the context of mechanisms designed for answering counting queries on constrained ranges [37]. We generalize these properties here in the context of mechanism matrices for general numeric queries:

- **Column Monotonicity** - For any fixed true response F_j , the probability of reporting a noisy response R_i given F_j monotonically decreases as the distance between F_j and R_i increases:

$$M_{i,j} \leq M_{i+1,j}, \quad i = 1, \dots, i_j - 1, \quad j = 1, \dots, |F|, \quad (5.58)$$

$$M_{i,j} \geq M_{i+1,j}, \quad i = i_j, \dots, |R| - 1, \quad j = 1, \dots, |F|. \quad (5.59)$$

- **Row Monotonicity** - For any fixed noisy response R_i , the probability of reporting R_i given a true query response F_j monotonically decreases as the distance between R_i and F_j increases:

$$M_{i,j} \leq M_{i,j+1}, \quad i = 1, \dots, |R|, \quad j = 1, \dots, |F| - 1 \text{ such that } F_{j+1} \leq r_i^u, \quad (5.60)$$

$$M_{i,j} \geq M_{i,j+1}, \quad i = 1, \dots, |R|, \quad j = 1, \dots, |F| - 1 \text{ such that } F_j \geq r_i^l. \quad (5.61)$$

- **Symmetry** - The matrix has 2-fold rotational symmetry (equivalent to centrosymmetry in this context):

$$M_{i,j} = M_{|R|-i+1, |F|-j+1}, \quad i = 1, \dots, |R|, \quad j = 1, \dots, |F|. \quad (5.62)$$

- **Fairness** - The probability of reporting the noisy response category containing the true query response is the same for all true query responses:

$$M_{i,j} = M_{i,j'}, \quad j = 1, \dots, |F|, \quad j' = 1, \dots, |F|. \quad (5.63)$$

Under the assumption that the range of a mechanism spans infinitely, many interactive mechanisms already satisfy these properties. Row and column monotonicity follow from the use of any PDF that monotonically decreases as distance from the location parameter increases (referred to henceforth as a monotonic PDF). Any data-independent mechanism will also satisfy the property of fairness since the true query response will not impact the probability of reporting the original value. These properties apply to well-known interactive mechanisms such as the Laplace mechanism [53], the staircase mechanism [76] and the geometric mechanism [77]. However, the adaptation of a mechanism to a finite range necessitates modification to the discretized PDF in order to achieve Requirement R.3. It is through such modification that the aforementioned properties may be lost.

These properties are considered to help avoid certain pathological behaviours that may result from blindly following optimization of an objective function in the design of a mechanism [37]. For example, although the geometric mechanism provides optimal utility for all Bayesian users [77], the well-known truncated variant (which snaps out-of-bounds responses to the nearest valid value) does not maintain column monotonicity as it induces spikes in probability mass on the noisy response categories at the query range boundaries. This behaviour has been shown to lead to poor utility for cases where the true query responses are near the center of the range, particularly in the case of small databases [37]. An alternate range-adherent mechanism for counting queries, the explicit fair mechanism, offers an improvement in utility in that context by adhering to the above four properties.

5.2.3 Boundary-Snapping

Next, we redefine the concept of boundary-snapping in the context of a mechanism matrix. Using this representation, we prove certain properties that boundary-snapping mechanisms possess.

The method of snapping out-of-bounds noisy query responses to the nearest valid value is commonly used when interactive mechanism responses must conform to a known range [37, 124]. This operation can be directly incorporated into the probability distribution of a mechanism with a finite range by reallocating all out-of-bounds probability mass to the nearest noisy response category. In the context of the discretization defined in Formula (5.50), this entails changing the lower bound of the first response category to negative infinity and the upper bound of the last response category to infinity.

Definition 11. *For any mechanism K and discretized range R , the **boundary-snapping** mechanism matrix representation of K is an $|R| \times |F|$ matrix M with its entries defined as follows:*

$$M_{i,j} = \begin{cases} \int_{-\infty}^{r_i^u} P_{K(F_j)}(x) dx & i = 1 \\ \int_{r_i^l}^{\infty} P_{K(F_j)}(x) dx & i = |R| \\ \int_{r_i^l}^{r_i^u} P_{K(F_j)}(x) dx & \text{otherwise} \end{cases} \quad i = 1, \dots, |R|, \quad j = 1, \dots, |F|. \quad (5.64)$$

Using the matrix representation of boundary-snapping mechanisms from Formula 5.64, we derive several properties possessed by such mechanisms.

Theorem 5. *The boundary-snapping mechanism matrix of Definition 11 satisfies all three requirements of a valid differentially private, range-adherent mechanism when applied to any differentially*

private mechanism K for a discretized range R that covers a finite and continuous range $\mathfrak{R} \subset \mathbb{R}$ such that $F \subset \mathfrak{R}$.

Proof. This implementation of boundary-snapping is equivalent to the post-processing step of snapping the noisy responses to the nearest valid value. Since differential privacy is immune to privacy breaches induced by post-processing [56], Requirement R.1 is satisfied, provided the original mechanism K is differentially private. Since R is a finite discretization of a range that covers all true query responses, Requirement R.2 is satisfied. Furthermore, as R is a discretization of a continuous range $\mathfrak{R}$, each matrix column j contains the full probability mass of a valid PDF $P_{K(F_j)}$ broken up over $|R|$ entries corresponding to integrals of $P_{K(F_j)}$. Since the probability density in a PDF can never drop below zero and the union of the integrals covers the full range $\mathbb{R}$ of $P_{K(F_j)}$, the column entries correspond to a valid probability distribution and satisfy Requirement R.3. $\square$

Theorem 6. *For any mechanism K based on a symmetric and monotonic PDF, range-adherence enforced through boundary-snapping on a mechanism matrix provides row monotonicity and symmetry under the assumptions that the discretized response categorizes are uniformly sized and the true query responses are uniformly spread across the range of the query. It does not guarantee the preservation of fairness or column monotonicity.*

Proof. The inflation of probability mass modifies only the boundary response categories. Prior to the inflation of probability mass, a mechanism matrix for a discretization using uniformly sized noisy response categories possesses the property of row monotonicity since each noisy response category corresponds to an integral of uniform size from a monotonic PDF. If the true query responses are uniformly spread across the range of the query, this initial matrix will also possess the property of symmetry. We must show that these properties are preserved after the inflation of probability mass to the boundary response categories.

The new probability mass associated with any boundary response category is the sum of its original value and all probability mass beyond its outer boundary. Let x be the original value of a boundary response category and x' be the shifted probability mass that is added to it. If the original mechanism K uses a monotonic PDF, the value of x is determined by a monotonic function of the distance d between the boundary response category and the location parameter. Similarly, the newly added probability mass x' is determined by a monotonic function of the same distance d as this determines how much probability mass has been shifted. Since both values that contribute to the sum monotonically decrease as the distance from the location parameter increases, it follows that the new boundary response categories maintain the property of row monotonicity.

The dependence of the inflated probability mass on distance between the location parameter and the boundary also ensures that symmetry continues to hold. For an arbitrary response category $M_{1,j}$ along the lower boundary, the true query response is $j - 1$ indices away from the lower boundary. We must show that $M_{|R|,|F|-j+1}$ is subject to the same increase in probability mass. This is a boundary response category on the upper boundary and the true query response is $j - 1$ indices away from this boundary. Since the true query responses are uniformly spread across the range of the query, it follows by the definition of boundary-snapping that the increase in probability mass will be the same. The property of symmetry therefore holds.

Column monotonicity may not hold since a boundary response category could have its probability mass increased beyond that of the adjacent, non-boundary response category despite the adjacent category being closer to the location parameter. The property of fairness also no longer holds since any boundary response category that coincides with a true response will now have a different probability mass from non-boundary response categories containing true responses. $\square$

5.2.4 Linear Programming

The adherence to range constraints achieved by boundary-snapping is an ad-hoc adaptation to mechanisms that were not designed for a range-constrained setting. As a consequence, it does not take into account the impact on the utility of the mechanism. Here, we wish to improve on this by proposing the use of linear programming to derive mechanisms that optimize a general objective function in order to provide a good level of utility for a broad range of users.

Outside of the context of counting queries, it is well-known that it is impossible to design mechanisms of universal optimality [23] (i.e., optimality for all users). We therefore cannot optimize utility for all Bayesian users. An alternative to optimality for Bayesian users is the notion of optimal risk-averse utility in which users wish to minimize the expected loss of the worst-case true query response [86, 127]. However, the measure of expected loss is taken in terms of user-specific utility goals, making the design of universally optimal mechanisms difficult for this form of utility as well. Although the staircase mechanism [76] offers optimal utility in general numeric queries for risk-averse users, it requires the assumption that users do not have any knowledge about range constraints of the query [75]. We note that there is an existing formulation of a linear program for risk-averse users posing counting queries [86]. However, this applies a user-specific objective function and is only used as an example rather than a proposed mechanism.

We therefore turn to an alternative criterion for optimality that abstracts from user-specific utility by considering the concentration of probability mass around the location parameter in the mechanism's PDF [184]. Intuitively, the concept is that for a given PDF, any rational user should prefer an alternative that can be derived from the given PDF by shifting probability mass closer to the location parameter as this would reduce the expected noise. To formalize this notion, an order was defined in [184] over PDFs based on the ability to obtain one PDF from another by shifting probability mass closer to the location parameter.

Definition 12. Let y_1 and y_2 be two PDFs having the same location parameter μ . The PDF y_1 is considered to be **smaller** than y_2 , denoted as $y_1 \leq y_2$, if y_1 can be obtained from y_2 by shifting probability mass closer to μ .

Definition 13. For any class Y of PDFs, a PDF $y \in Y$ is considered to be **optimal** if it is minimal within Y .

Under the assumption of data-independence, an optimal, staircase-shaped, differentially private mechanism was derived for this criterion [184]. The assumption of data-independence was used to assert that the PDFs used for any true query response differ only in their location parameters. As a result, the PDFs used by the mechanism are all shaped the same, allowing for the derivation of an optimal mechanism to be based on the identification of a single optimal (and differentially private) PDF. In our setting, we cannot make the same simplification as each true query response sits at a different position relative to the boundaries of the valid range of noisy query responses. As a result, the mechanism must use a different PDF for each true query response. To reflect this, we propose an adapted optimality criterion.

Definition 14. For any range-restricted query f having a set F of true query responses, let P_1 and P_2 be finite, ordered sets of PDFs of size $|F|$ used by two different mechanisms that have been designed for f . The set P_1 is considered to be **smaller** than P_2 , denoted as $P_1 \leq P_2$, if each element P_{1_i} is smaller (Under Definition 12) than its counterpart P_{2_i} as shown in Formula 5.65.

$$P_1 \leq P_2 \leftrightarrow P_{1_i} \leq P_{2_i}, \quad i = 1, \dots, |F|. \quad (5.65)$$

Definition 15. For any class $\mathbb{P}$ of finite, ordered sets of PDFs, a set $P \in \mathbb{P}$ is considered to be **optimal** if it is minimal within $\mathbb{P}$.

We must adapt this representation of optimality to a form that can be interpreted in a linear program. To make this change, we first show that the ordering over the elements of any class of finite, ordered sets of PDFs determined by the measure of expected loss using any monotonic loss function subsumes the ordering of Definition 14.

Theorem 7. Within a given class of finite, ordered sets of PDFs, the ordering of Definition 14 is subsumed by the ordering determined by the measure of expected loss over the sets of PDFs when using any non-decreasing loss function $L : \mathbb{R} \rightarrow \mathbb{R}$.

Proof. Let y_1, y_2 be two arbitrary PDFs from the same class Y . The inequality $y_1 \leq y_2$ implies that there are sets of ranges X and X' such that for each $x_i \in X$, some probability mass of y_2 can be shifted from this range to another range $x'_i \in X'$ that is nearer to the location parameter in order to produce y_1 . For each such shifted range, the decrease in the distance to the location parameter combined with the non-decreasing loss function implies a decrease in the measure of expected loss. This leads to the following inequality:

$$\int_{x \in \mathbb{R}} L(x) y_1(dx) \leq \int_{x \in \mathbb{R}} L(x) y_2(dx). \quad (5.66)$$

By Definition 14, for any pair of finite, ordered sets of PDFs P_1 and P_2 , such that $P_1 \leq P_2$, the above inequality must also hold for every pair of PDFs P_{1_i} and P_{2_i} . As a result, $P_1 \leq P_2$ implies an expected loss for P_1 that is less than or equal to that of P_2 . It follows that the ordering determined by any non-decreasing loss function subsumes that of Definition 14. $\square$

Corollary 4. For any class $\mathbb{P}$ of finite, ordered sets of PDFs, a set $P \in \mathbb{P}$ is **optimal** under Definition 15 if it minimizes expected loss for a non-decreasing loss function.

Proof. By Theorem 7, the ordering of Definition 14 is subsumed by the ordering determined by any non-decreasing loss function. As a result, any finite ordered set of PDFs that minimizes expected loss must also be minimal in the ordering of Definition 14, making it optimal under Definition 15. $\square$

With this result, we are now ready to formulate the linear programming objective function. Given that any non-decreasing loss function can be applied to achieve the required ordering, we apply L_1 distance between the true query response and the center of the noisy response category, and accordingly set the objective function as the minimization of the expected L_1 distance. Setting the conditional probabilities $M_{i,j}$ of the mechanism as the variables for the linear program, we define the objective function as follows:

$$\text{Minimize : } \sum_{i=1}^{|R|} \sum_{j=1}^{|F|} M_{i,j} \left| \frac{r_i^u + r_i^l}{2} - F_j \right|. \quad (5.67)$$

To ensure that the linear program produces a valid differentially private mechanism, we apply Requirements R.1 and R.3 of the matrix representation as constraints for the linear program. The inequalities of Formulae (5.52), (5.56) and (5.57) can be directly carried over without any changes in notation.

In some cases, users may wish to enforce additional properties in the mechanism in order to prevent certain pathological behaviours [37]. To address these concerns, we define an alternative linear program that is identical to the first but employs additional constraints to provide the properties of row and column monotonicity, symmetry and fairness. Once again, we can directly apply the inequalities and equations as specified in Formulae (5.60) - (5.63). To distinguish between these two linear program formulations, we will refer to the formulation without the additional constraints as Variant 1 and the formulation with the additional constraints as Variant 2.

We provide a small example to illustrate the resultant mechanism matrix for a counting query with a range of 0 - 5. For counting queries, it is natural to use the same set of integers from the possible true query responses as the set of noisy responses (i.e., $R = F$). The matrix for Variant 1 is shown in Table 5.3. We note that Variant 1 has assigned a probability of zero to the noisy response categories on the boundaries of the valid range. Although this means that these responses will never be reported by the mechanism, it allows for a more favourable level of expected utility overall. Given that it is expected for a randomization mechanism to frequently report noisy values that are not equal to the true query response, the impossibility of reporting a particular value may not be problematic as long as noisy responses nearby the true value have a high probability of being reported. The matrix for Variant 2 is shown in Table 5.4. The additional constraints have forced the entries with a probability of zero in Variant 1 to now take on non-zero values. The additional properties of this variant may be desirable for some data analysis tasks but they come at the cost of less flexibility in the optimization of the objective function.

	$\mathbf{f}(\mathbf{D}) = 0$	$\mathbf{f}(\mathbf{D}) = 1$	$\mathbf{f}(\mathbf{D}) = 2$	$\mathbf{f}(\mathbf{D}) = 3$	$\mathbf{f}(\mathbf{D}) = 4$	$\mathbf{f}(\mathbf{D}) = 5$
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 0$	0	0	0	0	0	0
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 1$	0.771	0.622	0.378	0.229	0.139	0.084
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 2$	0.090	0.149	0.245	0.149	0.091	0.055
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 3$	0.055	0.090	0.149	0.245	0.149	0.090
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 4$	0.084	0.139	0.229	0.378	0.622	0.771
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 5$	0	0	0	0	0	0

Table 5.3: Mechanism matrix for the linear programming Variant 1 solution of a counting query with a range of 0 - 5 and $\epsilon = 0.5$

	$\mathbf{f}(\mathbf{D}) = 0$	$\mathbf{f}(\mathbf{D}) = 1$	$\mathbf{f}(\mathbf{D}) = 2$	$\mathbf{f}(\mathbf{D}) = 3$	$\mathbf{f}(\mathbf{D}) = 4$	$\mathbf{f}(\mathbf{D}) = 5$
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 0$	0.315	0.191	0.116	0.070	0.043	0.026
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 1$	0.315	0.315	0.191	0.116	0.070	0.043
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 2$	0.231	0.265	0.315	0.191	0.116	0.070
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 3$	0.070	0.116	0.191	0.315	0.265	0.231
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 4$	0.043	0.070	0.116	0.191	0.315	0.315
$\mathbf{K}(\mathbf{f}(\mathbf{D})) = 5$	0.026	0.043	0.070	0.116	0.191	0.315

Table 5.4: Mechanism matrix for the linear programming Variant 2 solution of a counting query with a range of 0 - 5 and $\epsilon = 0.5$

5.3 Experimental Comparisons

We now present an experimental comparison between our proposed mechanisms and other range-adherent alternatives. Through these experiments, we demonstrate the improvements in utility that our mechanisms provide. We consider two different experimental settings to highlight where each of our proposed methodologies are most appropriate in practice. Our work on truncated and normalized Laplace mechanisms is appropriate for scenarios where the mechanism covers a large output range or where a suitable granularity for the discretization of mechanism responses is not directly obvious. In these situations, the use of linear programming may be impractical or computationally expensive. Contrarily, if the range is sufficiently small and discretization of responses is appropriate, our linear programming mechanisms are likely to provide better utility. We explain each of these settings in greater detail over the remainder of this section, comparing the performance of our applicable mechanisms and other range-adherent alternatives in each of them.

5.3.1 Normalized Laplace Experiments

We begin by showing a setting well-suited to our approach for truncating and normalizing the Laplace distribution. When running Naïve Bayes classification on a differentially private dataset, it is necessary to obfuscate the output values from the training process of the classifier. Among these are standard deviation values which must remain positive after obfuscation in order to be used by the classifier. Since the range of possible values is unbounded towards infinity, an appropriate discretization of the mechanism range is not obvious without access to further information about the dataset, which is sensitive and thus not directly available. This makes the normalized Laplace mechanisms an appropriate choice.

We first describe in detail this experimental setting, providing additional information about differentially private Naïve Bayes classification. We then describe the other mechanisms to which we compare ours. Finally, we present our results and provide some discussion on our findings

Experimental Setup

To compare the mechanisms, we set up a Naïve Bayes classifier [141] to measure classification accuracy through differentially private training on datasets for the task of multivariate classification. The goal of the classifier is to learn from the training records how to use a set of attribute values to predict a categorical class for other unseen records. In the standard setting (without differential privacy) for records having real-valued attributes, the training process aims to approximate the means and standard deviations of normal distributions for each attribute with respect to each potential class. Using these approximated distributions along with approximations of class probabilities, the classifier can apply Bayes' theorem to guess the most probable class for new records when only the attribute values are known. For a more thorough description of Naïve Bayes classifiers, we refer the reader to [141].

In instances where the training data is sensitive, differential privacy can be applied to protect the data. We follow the approach of [197] for the application of noise during the training process. In particular, noise must be applied to the means and standard deviations of the normal distributions as well as the class probabilities since these are the outputs of the training process. The means and class probabilities can be protected by any standard mechanism. We apply the basic Laplace mechanism here. The standard deviation of a normal distribution must be non-negative. Therefore, a mechanism that can produce negative values cannot be applied here. The original process [197] recommends redrawing noisy values from a Laplace mechanism until a non-negative output is obtained. However, as we have shown in Section 5.1.1, this approach is not differentially private; this has also been

independently discovered by Holohan et al. [99]. Here, we propose to instead use a truncated and normalized Laplace mechanism to obtain valid noisy responses in a differentially private manner.

To demonstrate the usefulness of our proposed mechanisms, we compare them against other range-adherent alternatives for the obfuscation of standard deviation values during the training process. We run experiments on the Banknote Authentication, Seeds, and Yeast datasets of the UCI Machine Learning Repository [49]. The Banknote dataset contains 1372 records with 4 real-valued attributes used for a binary classification task. The Seeds dataset contains 210 records with 7 real-valued attributes used for classification over 3 classes. The Yeast dataset contains 1484 records with 8 real-valued attributes used for classification over 10 classes. For each experiment, we perform 10000 runs in which the dataset is randomly divided into a 80% training 20% testing split. Each range-adherent mechanism is applied to the split data to obfuscate the standard deviations (while using the basic Laplace mechanism to obfuscate the other sensitive values). We split a privacy budget ϵ evenly across all values to be obfuscated. By the composition theorem [56], this guarantees ϵ differential privacy across all released values. We measure the usefulness of the mechanisms by the classification accuracy of the resultant classifier. Higher classification accuracy indicates better preservation of useful information.

Selected Comparison Mechanisms

Among the mechanisms we propose, our calculation of the scaling parameter for a single infinitely spanning constraint is the most appropriate configuration for the requirement of producing non-negative values. By setting a constraint that spans from zero to negative infinity, we are able to calculate optimal scaling parameters to ensure non-negative responses. We use this as our primary approach to test. For the purpose of comparison, we also apply our results for the setting of arbitrary constraints. As explained in Section 5.1.4, in the presence of an infinitely spanning constraint, the calculation of the scaling parameter can be simplified to $\frac{2\Delta F}{\epsilon}$. This coincides with the configuration that would be required for the exponential mechanism [135] to be applied in this setting. This mechanism can therefore be interpreted either as a comparison of our results under this setting or a comparison against the exponential mechanism.

To demonstrate the improvement our mechanisms can provide in the utility of the noisy query responses over those produced by a boundary-snapping mechanism, we include a Laplace mechanism with boundary-snapping [124] in our experiments. This mechanism simply draws from a Laplace distribution using the standard scaling parameter of $\frac{\Delta F}{\epsilon}$ and snaps all negative responses to the nearest valid value. Unlike the approach of continuing to draw new values until a non-negative value is obtained, this approach is simply post-processing of a noisy response and therefore preserves the differential privacy guarantee [56].

We additionally include for comparison another truncated and normalized Laplace mechanism which was independently developed by Holohan et al. [99]. Their work shares the same motivation as our own, however, the approaches taken are different. They focus on the setting of two infinitely spanning constraints (a single finite span of feasible query responses) and search for the optimal uniform scaling parameter that can be applied for a Laplace distribution that has been truncated and normalized for this span. Since this mechanism requires an upper bound for the range of query responses, we set a sufficiently high value, 10^8 , to simulate the absence of an upper bound in the experimental setting. This is the same approach taken in their original work, where an upper bound of 10^{10} was employed.

While we use a privacy guarantee that applies to query responses at arbitrary distances from each other, the work of Holohan et al. maintains the form of a privacy guarantee for adjacent databases. A key advantage in our work is that the use of a variable scaling parameter allows us in many cases to

calculate a lower scaling parameter than what is required by their mechanism. However, as we have explained in Section 5.1.1, the privacy guarantee we employ is slightly stronger than the standard guarantee for adjacent databases. Therefore, despite the use of a uniform scaling parameter, they are able in some cases to achieve lower scaling parameters than our own calculations (we explain conditions under which this can occur in the forthcoming section on our results). This makes their work an interesting candidate for comparison in our experiments.

Results and Analysis

The results of our experiments are presented in Figure 5.3. We show the baseline classification accuracy from the original data as well as the accuracies for each mechanism over a range of privacy budgets (i.e., values for ϵ). The advantage of using a truncated and normalized distribution over the post-processing approach of boundary-snapping is made clear by the large gain in classification accuracy for all normalized mechanisms over the boundary-snapping mechanism. The relatively high probability for boundary-snapping to report noisy responses on the boundary forces the classifier to use many standard deviations that have the same value. This greatly reduces their utility as a means to distinguish between different potential classes.

In Figures 5.3-a and 5.3-b, we observe that among the normalized mechanisms, our mechanism for single infinitely spanning constraints offers the highest accuracy, followed by the mechanism of Holohan et al. and then our mechanism for arbitrary constraints. Given that this would be the expected order of the mechanisms from lowest to highest scaling parameters used, these results line up with our expectations.

Figure 5.3: A comparison of classification accuracies for our mechanisms and other range-adherent alternatives applied to three different datasets. The x-axis indicates the total privacy budget ϵ . Higher classification values on the y-axis indicate better utility.

In Figure 5.3-c, the mechanism of Holohan et al. is able to provide higher classification accuracy than our mechanisms. This can occur when the true query responses are very close to the constraints (in terms of distance measured by spans of query sensitivity). This type of configuration requires high scaling parameters to achieve differential privacy. Due to the stronger privacy guarantee used by our mechanisms, we end up calculating higher scaling parameters than the other normalized mechanism here. However, all mechanisms suffer in this type of setting since the high degree of noise leads to responses that are far from their original values. This results in very low classification

accuracy unless the value of ϵ is made unacceptably high. In this example, ϵ had to be raised well above 100 to achieve reasonable accuracy. Therefore, we argue that differential privacy is not appropriate for use in this type of setting in the first place.

With respect to the variance of the responses produced by the mechanisms, one can compare the normalized mechanisms in terms of the scaling parameters used to configure the mechanisms. Therefore, the mechanism for arbitrary constraints has the highest variance in its noisy responses, followed by the mechanism of Holohan et al., and the mechanism for the single infinitely spanning constraint has the lowest variance for its responses. Given that the boundary-snapping mechanism requires no modification to its scaling parameter as the other mechanisms do, it offers the lowest level of variance in its noisy responses. Furthermore, the inflation of probability mass on the boundaries of the range of the query can often lead to even lower variance when the true query response is near a boundary. However, despite the lower variance intuitively pointing toward better performance from the mechanism, it is the inflation of probability mass at a fixed point that ultimately leads to worse performance. For the task of creating a classifier, a high probability for the same classes to produce the same standard deviation values on the boundary of the range detracts from the classifier's ability to distinguish between such classes. As a result, boundary-snapping performs poorly in this context.

5.3.2 Linear Programming Experiments

Next, we study scenarios in which query response ranges can be reasonably discretized, making the use of linear programming mechanisms appropriate. We apply the mechanisms to both synthetic data as well as microdata from Statistics Canada [25]. We investigate a number of factors in the the performance of range-adherent mechanisms including the size of the dataset, the granularity of the response discretization and whether or not the mechanism is configured to adhere strictly to the range of the query.

We first provide details on five other range-adherent interactive mechanisms to which we compare our linear programming mechanisms. Then, we state the utility measures that we apply. This is followed by a detailed discussion on our experiments where we give our analysis. By using both synthetic data as well as real data, we gain insight into expected behaviour and practical performance.

Mechanisms

We compare our linear programming approach to boundary-snapping variants of three well-known mechanisms: the Laplace [53], the staircase [76] and the geometric [77] mechanisms (Mechanisms 1, 2 and 3, respectively). Additionally, we include our normalized Laplace mechanism for arbitrary constraints and the explicit fair mechanism [37] (Mechanisms 4 and 5, respectively). Whereas the geometric and explicit fair mechanisms are applicable only to counting queries, all others may be employed for any numeric query.

Mechanism 1. Laplace - For a database D , noise is drawn from a Laplace distribution using $f(D)$ as the location parameter and $\frac{\Delta F}{\epsilon}$ as the scaling parameter:

$$\text{Lap}\left(x|f(D), \frac{\Delta F}{\epsilon}\right) = \frac{\epsilon e^{-\frac{\epsilon|f(D)-x|}{\Delta F}}}{2\Delta F}. \quad (5.68)$$

Mechanism 2. Staircase - For a database D , a mechanism drawing from a staircase-shaped PDF centered at $f(D)$ minimizes the expected distortion of the noisy query response:

$$Stair(x|f(D)) = \begin{cases} y & |f(D) - x| \in \left[0, \frac{\Delta F}{1+e^{\frac{\epsilon}{2}}}\right] \\ ye^{-\epsilon} & |f(D) - x| \in \left[\frac{\Delta F}{1+e^{\frac{\epsilon}{2}}}, \Delta F\right] \\ e^{-k\epsilon} Stair(x - k\Delta F) & |f(D) - x| \in [k\Delta F, (k+1)\Delta F] \text{ for } k \in \mathbb{N}, \end{cases} \quad (5.69)$$

$$y = \frac{1 - e^{-\epsilon}}{2\Delta F \left(\frac{1}{1+e^{\frac{\epsilon}{2}}} + e^{-\epsilon} \left(1 - \frac{1}{1+e^{\frac{\epsilon}{2}}} \right) \right)}. \quad (5.70)$$

Mechanism 3. Geometric - For a counting query posed on a database D , a mechanism drawing from a geometric distribution centered at $f(D)$ provides optimal expected utility for Bayesian users:

$$Geo(z|f(D)) = \frac{e^\epsilon - 1}{e^\epsilon + 1} \left(\frac{1}{e^\epsilon} \right)^{|f(D)-z|} \text{ for } z \in \mathbb{N}. \quad (5.71)$$

Mechanism 4. Normalized Laplace - For a database D , noise is drawn from a Laplace distribution truncated to the range of $[r_l, r_u]$ and normalized within this range, using $f(D)$ as the location parameter and $\frac{2\Delta F}{\epsilon}$ as the scaling parameter:

$$NormLap\left(x|f(D), \frac{2\Delta F}{\epsilon}\right) = \begin{cases} \frac{2Lap(x|f(D), \frac{2\Delta F}{\epsilon})}{2-e^{-\frac{\epsilon(f(D)-r_l)}{2\Delta F}} - e^{-\frac{\epsilon(r_u-f(D))}{2\Delta F}}} & r_l \leq x \leq r_u \\ 0 & x < r_l, x > r_u. \end{cases} \quad (5.72)$$

Mechanism 5. Explicit Fair - For a counting query posed on a database D , a mechanism drawing from a distribution centered at $f(D)$ provides the properties of row monotonicity, column monotonicity, symmetry and fairness:

$$EF(x|f(D)) = \begin{cases} \frac{y}{e^{\epsilon|f(D)-x|}} & |f(D) - x| < \min(f(D), |D| - f(D)) \\ \frac{y}{e^{\epsilon\lceil \frac{|f(D)-x| + \min(f(D), |D| - f(D))}{2} \rceil}} & |f(D) - x| \geq \min(f(D), |D| - f(D)), \end{cases} \quad (5.73)$$

$$y = \frac{1 - \frac{1}{e^\epsilon}}{1 + \frac{1}{e^\epsilon} - \frac{2}{e^{\epsilon\left(\frac{|D|}{2} + 1\right)}}}. \quad (5.74)$$

Utility Measures

To compare the utility of the mechanisms, we employ measures of both usability and information preservation. Recall that usability indicates ease of use for specific tasks while information preservation indicates how much of the original information is maintained in the noisy query responses. We therefore consider usability in terms of taking noisy query responses at face value (i.e., without applying any post-processing). See e.g., [76, 87] for instances of utility measured in this way. Specifically, we measure both the expected absolute error and the expected squared error of noisy query responses with respect to the true query response. These expected measures are calculated over all true query responses and over all possible noisy query responses for each true response. We assume a uniform distribution for the probabilities of the true responses. The probabilities of the noisy query responses are determined by the conditional probability of the mechanism given a particular true query response. The expected absolute error acts as a measure of the amount of distortion expected

to be induced by the mechanism, giving preference to mechanisms that have low distortion in the expected case. The expected squared error assigns a penalty that grows rapidly as the amount of distortion increases, giving preference to mechanisms that avoid allowing for high distortion in some cases in order to achieve low distortion in other cases. We have selected these measures as they provide a good representation of typical properties that are desirable in terms of the usability of a mechanism. The two measures are defined as follows:

$$Ex_Err = \sum_{f \in F} \frac{\sum_{r \in R} \Pr(r|f) |f - r|}{|F|}, \quad (5.75)$$

$$Ex_Sqr_Err = \sum_{f \in F} \frac{\sum_{r \in R} \Pr(r|f) (f - r)^2}{|F|}. \quad (5.76)$$

We additionally measure utility for Bayesian users [77] having a uniform distribution over the true query responses as their prior knowledge. This models a common scenario in which the user initially has no additional information about distribution of the true query responses. A Bayesian user will remap noisy query responses according to their prior knowledge about the true query responses in order to optimize an objective function (e.g., minimization of absolute error). The measure of expected loss under Bayesian post-processing is appropriate as a measure of information preservation [123] and is therefore useful as contrast against the measures of usability. For close correspondence between these measures, we employ instances of Bayesian post-processing for minimization of expected absolute error and expected squared error. Given a noisy response $r \in R$, this class of Bayesian users will guess a query response $f \in F$ that minimizes the expected loss according to their prior knowledge. We define this measure of loss as shown in Formula (5.77). Note that the function $\text{Loss}: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ may be set to any appropriate loss function and for our experiments, we use absolute error (Formula (5.78)) and squared error (Formula (5.79)).

$$Ex_Baye_Loss = \sum_{r \in R} \Pr(r) \min_{f \in F} \sum_{f' \in F} \frac{\Pr(r|f') \text{Loss}(f', f)}{|F|}. \quad (5.77)$$

$$Err(f, f') = |f - f'|. \quad (5.78)$$

$$Err_Sqr(f, f') = (f - f')^2. \quad (5.79)$$

Experimental Setup

In the following subsections, we use boundary-snapping versions of the Laplace and staircase mechanisms to query the mean of an integer-valued scalar attribute that can take on 5 possible values (0 through 4). We use a boundary-snapping version of the geometric mechanism and the explicit fair mechanisms for a counting query over a binary attribute. We employ our linear programming mechanisms for both types of queries in order to compare them to all other mechanism types. To solve the linear programs needed for our mechanisms, we used WinGLPK [175], a Windows executable version of the GNU Linear Programming Kit [164]. In all experiments, we assume that the database sizes and attribute response categories are public knowledge.

With respect to the discretization of the noisy query responses, it is natural to use the set of integers that correspond to the valid true query responses for counting queries. For a counting query on a database D , this will be $\{0, \dots, |D|\}$. With most other types of queries, it is necessary to specify

a set of noisy response categories that cover the full valid range of the query (e.g., no gaps between noisy response categories) in order to obtain valid probability distributions from mechanisms that use continuous PDFs. When querying a mean value on a database D with records taking on integer values from 0 to k , there are $k|D| + 1$ possible true query responses, each of which are evenly spaced at a distance of $\frac{1}{k|D|}$ apart over the valid range of $[0, k]$. To handle the discretization, we set the size of each noisy response category to $\frac{1}{k|D|}$ and position them such that each noisy response category is centered on a true query response. For all distance measures, we consider the value of a noisy response category to be the midpoint of the span it covers.

In all experiments, we have measured expected absolute error and expected squared error both with and without Bayesian post-processing. Since the relationships between the plotted curves for expected squared error showed no significant differences from those of the expected absolute error, we only show the later for space considerations.

Database Size

We begin our experimental analysis by plotting expected loss as a function of the database size for fixed values of the privacy parameter ϵ . In this experiment and all others, lower values indicate better utility. Through these experiments, we demonstrate that the most notable differences in utility between the mechanisms occur in small databases. This is in line with the original results on the explicit fair mechanism [37], which found that, in the context of counting queries, improvements in utility obtained through alternate methods of adherence to range constraints are most prominent for queries posed on small databases. Results are shown in Figures 5.4 - 5.7. As the database size increases, the levels of utility from the mechanisms appear to converge. Given this observation, we focus our other experiments on small databases (10 records) in order to better observe the differences in the behaviours of the mechanisms. In practice, small datasets may arise when studying instances of rare diseases or marginal groups of a population.

Figure 5.4: Expected loss as a function of the database size for a mean query with $\epsilon = 0.2$

Figure 5.5: Expected loss as a function of the database size for a mean query with $\epsilon = 0.5$

Figure 5.6: Expected loss as a function of the database size for a counting query with $\epsilon = 0.2$

Figure 5.7: Expected loss as a function of the database size for a counting query with $\epsilon = 0.5$

Discretization

As we have chosen a specific method for determining the noisy response category size in the mechanism discretizations, we include experiments using different sizes for the noisy response categories to observe the effects that this may have on the utility of the mechanisms. Results are shown in Figure 5.8 for a mean query. Each plotted line shows a different size of discretized response categories. The numbers associated with each plot in the legends indicate a scaling factor for the noisy response category size. Here, lower numbers indicate finer granularity and higher numbers indicate coarser granularity. Surprisingly, the boundary-snapping Laplace mechanism achieves better levels of utility with a coarser granularity in high privacy settings. This is due to the fact that the high level of privacy leads to the probability mass being very spread out from the location parameter and the impact of boundary-snapping causes this mass to pool at the boundaries of the valid range. The finer the granularity of the discretization is, the closer that probability mass moves toward the extremities of the valid range, leading to greater expected distance between noisy query responses and true query responses. This suggests that boundary-snapping is a poor choice for queries posed on small databases under high privacy settings. To further illustrate this point, we include an additional mechanism that simply uses a uniform distribution over the noisy query responses in these experiments. In the high privacy setting, even this trivial mechanism performs better than the boundary-snapping variant.

For the discretization experiments on the linear programming mechanism, we observe the expected trend of coarser granularity leading to worse levels of utility. We note that for the three finest levels of granularity tested, the differences in the levels of utility are almost negligible. Intuitively, the loss in query accuracy induced by the injected noise negates much of the benefit that could otherwise be achieved through the use of higher precision. This supports our decision to use $\frac{1}{k|D|}$ as the size of the noisy response categories in all other experiments. We omit graphs for the other measures as they display similar trends.

Figure 5.8: Expected loss as a function of ϵ for a mean query. Each plot uses noisy responses category sizes equal to the minimum distance between two different true responses scaled by the postfixed number in the legend (e.g., a postfix of 5 indicates categories 5 times larger).

Expected Utility

With a fixed size for the database and the noisy query response categories, we are able to plot the levels of utility as a function of the privacy parameter ϵ to compare different types of mechanisms. This allows for a more comprehensive understanding of the behaviours of the mechanisms at varying levels of privacy. Our results are shown in Figures 5.9 and 5.10.

The linear programming Variant 1 consistently provides the best levels of utility, with the greatest improvements over other mechanisms appearing in the non-post-processing measures at high levels of privacy. As the level of privacy is reduced, the utility measures of the mechanisms converge. For the mean query, the mechanism for Variant 1 has a slight lead over the other mechanisms for the measures of post-processing while in the counting query, it is on par with the boundary-snapping geometric mechanism. For the linear programming Variant 2, although it does not reach the same levels of utility as Variant 1, it still outperforms the other mechanisms in the non-post-processing measures and is on par with the boundary-snapping variants for the measures of post-processing in the mean query. For the post-processing measures in the counting query, Variant 2 performs very marginally worse than Variant 1 and the boundary-snapping geometric mechanism.

Figure 5.9: Expected loss as a function of ϵ for a mean query

Figure 5.10: Expected loss as a function of ϵ for a counting query

Extended Range

As we have hypothesized that the incorporation of range constraints into the design of a mechanism will have a positive impact on the level of utility, we provide a comparison between range-adherent mechanisms and non-adherent variants. Since any computer implementation of a mechanism must have a finite range, we apply a greatly widened range to model a lack of adherence to range constraints. To do so, we widen the range of noisy query responses to 300% of the constrained range, keeping this span centered on the valid range. Since probability mass is always focused near the true query response, any further widening of the range beyond 300% has little impact and does not provide additional insight. Results are shown in Figures 5.11 and 5.12. For visual clarity, we omit the boundary-snapping variant of the Laplace mechanism and Variant 2 of the linear programming mechanism from the graphs. The behaviour of the extended range versions of these omitted mechanisms matched those of the boundary-snapping staircase mechanism and the linear programming Variant 1, respectively. Since the explicit fair mechanism is explicitly formulated in terms of the true range of the query, a range-extended version cannot be created, however we include the original version in the experiments as a point of reference.

The most notable difference in the range-extended mechanisms is that the boundary-snapping variants show substantially worse levels of utility in the non-post-processing measures at high levels of privacy. In the measures of post-processing, there are no notable differences between the range-adherent and range extended version except with the normalized Laplace mechanism which performs slightly better in the extended-range variant. The linear programming mechanisms show no differences in any measures of utility between their range-adherent and extended range variants.

Figure 5.11: Expected loss as a function of ϵ for a mean query. Mechanism variants using an extended range are marked by the postfix (*ext*).

Figure 5.12: Expected loss as a function of ϵ for a counting query. Mechanism variants using an extended range are marked by the postfix (*ext*).

Real Data

Finally, to gain insight on the practical performance of the mechanisms, we have run experiments on real data. We have used the Statistics Canada 2011 National Household Survey public use microdata file [25] which we have partitioned into groups of size 10, using only records from the province of Ontario. In total there are 341,253 such records. As required by Statistics Canada’s data use regulations, we state that the results or views expressed here are not those of Statistics Canada. To measure the utility of a query posed on a particular group, we employ each mechanism variant to generate a noisy response. For the measures of usability, the response is taken at face value, whereas for Bayesian users, the response is mapped to their best guess. In both cases, the resultant value is used to measure the loss (as absolute error and squared error) with respect to the true query response. Since the mechanisms are stochastic functions, this process is repeated 1000 times for each mechanism variant to obtain mean measures of utility. This is done over each of the groups and once again, mean values are computed across the groups.

We have run this experiment for both query types, mean and counting, in order to test all mechanisms in this setting. For the mean query, we posed a query for the mean family size of the respondents. The valid range of responses are integers from 1 (no family) up to 7 (interpreted as 7 or more family members). The mean value over all groups for this query is 2.02. Results are shown in Figure 5.13. For the counting query, we posed the number of respondents holding a university degree. The mean value over all groups is 2.27. Results are shown in Figure 5.14.

Figure 5.13: Expected loss as a function of ϵ for mean family size

Figure 5.14: Expected loss as a function of ϵ for total respondents with a university degree

The results for the non-post-processing measures in the mean query follow the same trends as in the experiments of expected utility on synthetic data. The linear programming Variant 1 performs best with Variant 2 next, most prominently at high levels of privacy. With the measures

of post-processing in the mean query, the linear programming Variant 1 also performs best at mid-to-low levels of privacy. Surprisingly, Variant 2 performs best at high levels of privacy for the post-processing version of mean error. Since most of the data groups are expected to have similar true query responses, we hypothesize that the mechanism is simply able to perform well on the specific query responses that were commonly encountered despite having an overall lower level of expected utility than Variant 1. For our results on the counting query, the trends match those of the synthetic data with the linear programming variants performing best in the non-post-processing measures and matching the boundary-snapping geometric mechanism in the post-processing measures.

Discussion

We conclude this section with a discussion on the practical implications of the results from our analysis. We assess the impact of adherence to range constraints on the utility of the mechanisms and discuss the performance of our linear programming mechanisms.

Our results indicate that adherence to query range constraints can indeed be beneficial for the utility of mechanisms, provided adherence is appropriately handled. As with counting queries [37], the cases where these results are most meaningful appear to be for small group sizes and high levels of privacy. For cases where queries are posed on small groups, high levels of privacy are arguably the most important setting as these individuals are likely to be more susceptible to unwanted inferences than those who are hidden among larger group sizes. From our results on experiments using an extended mechanism range, we see that adherence to range constraints greatly improves the measures of non-post-processing, indicating a significant improvement in usability. The correspondence in the measures of post-processing between the range-adherent mechanisms and the extended-range variants indicates that the mechanisms have not suffered in information preservation through the modifications required for adherence to range constraints. This suggests that the explicit adherence to range constraints offers a favourable improvement in utility with respect to a trade-off between usability and information preservation.

Overall, the linear programming Variant 1 performs the best both in terms of expected error and expected squared error, offering superior levels of utility in the non-post-processing measures and matching or slightly exceeding the levels of utility achieved by boundary-snapping variants in the measures of post-processing. While not matching the levels of utility achieved by Variant 1, the linear programming Variant 2 does match or exceed the levels of utility of all other mechanisms with the exception of the boundary-snapping geometric mechanism in measures of post-processing for which there is a very minor degradation in the level of utility. These results are also visible in the experiments on real data with the exception some minor differences in the post-processing measures of the mean query at high levels of privacy. Given that we are interested both in usability as well as information preservation, we do not consider this discrepancy to outweigh the overall benefits of the linear programming Variant 1. For queries posed on small databases at mid-to-high levels of privacy, we recommend the linear programming Variant 1 for the best levels of utility. For users interested in mechanisms that offer the additional properties of row and column monotonicity, symmetry and fairness, the linear programming Variant 2 is a good alternative.

The number of variables in the linear programs is the product of the number of true query responses and the number of noisy response categories. Due to this, the computational expense of the linear program will grow rapidly as the size of the database increases, making it impractical to derive linear programming mechanisms for large cases. Despite this apparent shortcoming, we expect that this poses little practical significance. Our results suggest that on larger databases sizes, the utility levels of the different mechanisms begins to converge. The major differences in utility occurred for input sizes that can be reasonably handled by a linear programming solver. We

therefore recommend the use of the linear programming mechanisms for small-sized databases (exact sizes depend also on the chosen discretization and the computational power of the target machine) and the use of an alternative such as boundary-snapping for larger databases.

Finally, we note that although we expressly chose an optimality criterion that abstracts from user-specific goals as the basis for the linear programming objective function, the selection of a specific loss function necessarily imposes some degree of user preference according to the selected measure of loss. The subsumed ordering of Definition 14 remains the same as long as a non-decreasing loss function is chosen, however, there are many pairs of finite, ordered sets of PDFs that are incomparable within this order but are comparable once a loss function is imposed. This implies that some user-specific preferences leak into the final objective function. As such, it may be the case that other loss functions or different objective function formulations may offer better results for general users.

5.4 Future Work

The performance of the normalized Laplace mechanisms relative to the alternatives depends largely on the trade-off between the improved utility which is achieved through normalization and the reduction in utility induced by the need for a higher scaling parameter. As we have shown, the increased scaling parameter is necessary to ensure the preservation of the differential privacy guarantee when the data-dependent operation of normalization is applied. Yet, the distance-based privacy guarantee we employ is slightly stronger than the standard guarantee. Although this has allowed us to calculate variable scaling parameters to achieve good utility in the case of a single infinitely spanning constraint, we have not identified a method to similarly calculate variable scaling parameters for fully arbitrary constraint configurations. Such a calculation would lead to a more versatile mechanism with improved utility over our current application of $\frac{2\Delta F}{\epsilon}$ for the scaling parameter.

Furthermore, we postulate that under an appropriate relaxation of the differential privacy guarantee, one may be able to strike a better balance between privacy and utility through the use of lower scaling parameters as enabled by the chosen relaxation. This has the potential to further improve the utility of normalized Laplace mechanisms. Similarly, one could design new range-adherent linear programming mechanisms that satisfy relaxed forms of differential privacy. We leave the investigation of these topics as an open problem.

Lastly, a study of the variance in the noisy responses of the mechanisms, particularly those produced by the linear programs, may be of interest. While our mechanisms provide lower levels of expected error than the alternatives, this does not guarantee better performance in all cases. For example, should true query responses often fall near the boundaries of the range of the query, boundary-snapping mechanisms are likely to provide better performance. A study of properties such as variance and expected error as a function of true query responses could help to provide further insight on the strengths and weaknesses of the mechanisms. Additionally, in appropriate contexts, incorporation of information about the expected distribution of the true query responses may help to produce more useful mechanisms.

5.5 Summary

When posing queries on a sensitive database, natural range constraints of the query may be publicly known. However, most randomization mechanisms do not adhere to range constraints when adding noise to query responses. Adherence to such publicly known constraints on queries can be an effective method to improve utility in differentially private mechanisms. We have studied this problem

from two different design perspectives for range-adherent mechanisms: 1) truncated and normalized Laplace mechanisms, and 2) linear programming mechanisms.

Truncation and normalization of a Laplace PDF offers a means to generate noisy responses within a specified range while improving the utility of the mechanism by inducing an increase in probability density that is greater for noisy responses nearer the true response than for responses farther away. We have shown that since the normalization of the PDF is a data-dependent operation, this leaks sensitive information and violates the differential privacy guarantee. We have proposed a method to correct this which involves determining a scaling parameter then used by the Laplace distribution. We have introduced an adaptation of the differential privacy guarantee which can be applied to the Laplace distribution to incorporate data-dependent normalization factors. Using our privacy guarantee, we have studied different classes of range constraint configurations and provide derivations of optimal scaling parameters or approximations thereof. Using our proposed calculations, a data custodian can apply the Laplace distribution to answer general numeric queries in a range-adherent and differentially private manner. Through experimental comparisons, we have shown that our proposed mechanisms can offer improvements in utility over other range-adherent alternatives.

In addition to this, we have combined a formalized matrix representation of range-adherent mechanisms with a user-independent criterion for optimal utility in order to provide a basis for the derivation of range-adherent linear programming mechanisms. We have proposed two range-adherent linear programming variants, one of which is subject only to the differential privacy requirement while the other is additionally subject to constraints providing row and column monotonicity, symmetry and fairness in the mechanism matrix. Through experimental comparisons, we have shown that the linear programming mechanisms are able to provide improvements in utility over boundary-snapping and other range-adherent alternatives.

Chapter 6

Differentially Private Obfuscation of Images

When privacy is a concern for individuals depicted in digital images, it is often necessary to obfuscate the images in order to remove information that could lead to identification of the individuals. Many methods of obfuscation focus on the face as it is often the most identifiable piece of information [160, 167]. Trivially, the face could be covered by a uniformly coloured rectangle. This destroys all information about the face, guaranteeing that it can no longer be exploited to reveal an identity. However, this also destroys a great deal of utility and visual appeal. In scenarios where images are shared in online platforms, users have expressed a strong aversion to the use of such rectangles with respect to the visual quality and information content of the images [120, 121]. Less severe methods such as blurring or pixelization may be employed as an alternative but it is difficult to find a good trade-off between the aspects of privacy protection, visual appeal and information sufficiency [92, 121]. Given the aversion of users towards obvious applications of obfuscation, modifications that appear visually natural would be preferable. Ideally, obfuscated images should appear photo-realistic, with the obfuscated faces imitating natural distributions over facial features that we encounter in day-to-day life. Effective yet subtle obfuscation is likely to be desirable for use in contexts where users would directly view the obfuscated results such as on social media and image sharing platforms.

It is well-known that methods of privacy protection necessitate a trade-off between privacy and utility [10, 38]. For example, a study on privacy-preserving systems used for visual awareness of co-workers in remote locations found that the level of privacy protection applied to hide identities directly impacted the value of the data to those using the systems [101]. Preservation of utility is especially important for machine learning and data mining. Visual data can be used to learn about customers in retail environments [110, 125] and to detect anomalous or illegal events [13, 198]. In such cases, preservation of demographic information may be of particular importance. For example, gender and approximate age may be particularly useful for performing retail analytics using facial images captured from CCTV footage. Since not all attributes of interest may be known ahead of time, a good method of obfuscation should ideally limit its modification of non-sensitive information as much as possible.

The balance between privacy protection and image quality and utility is further exacerbated by the fact that many ad hoc methods of privacy protection have been shown to ultimately be ineffective. The human visual system has remarkable robustness against blurring in images, allowing for recognition of faces against all but the most extreme applications of blurring [145, 182]. Fur-

thermore, super-resolution techniques can offer realistic reversals of pixelization [139] and machine learning algorithms have been shown to effectively defeat obfuscation via blurring as well as other forms effective against human recognition such as strong pixelization [134, 146]. While many approaches lack a formal privacy guarantee, the k -same [146] family of approaches has gained a great deal of traction, largely thanks to its guarantee that for a chosen privacy parameter k , obfuscated individuals are indistinguishable within a group of k potential true identities. Although this privacy guarantee is appealing, it suffers from susceptibilities (e.g., composition attacks [73]) carried over from the disclosure control method of k -anonymity on which it is based.

We propose an alternate approach to facial obfuscation through the application of differential privacy to representations of images within generative machine learning models. In this way, we achieve a strong guarantee of privacy while producing photo-realistic obfuscated images. Differential privacy provides a robust and provable privacy guarantee regarding the distinguishability of obfuscated identities. The use of generative models allows for the generation of highly realistic faces depicting novel, non-existent identities. Furthermore, such generative models allow for the preservation of desired properties in the obfuscated images such as the original pose or gender from the input image. Our proposed approach is therefore able to offer favourable privacy-utility trade-offs in the obfuscated output.

The content of this chapter is structured as follows: In Section 6.1, we open with some notation and definitions which are used throughout the chapter. In Section 6.2, we discuss deficiencies that are present in the existing methods of facial obfuscation. We then propose our framework for the application of differential privacy to generative models for the obfuscation of facial images in Section 6.3. We discuss the benefits of our framework and provide implementation details for two types of generative models. In Section 6.4, we propose architecture and training details for a generative adversarial network (GAN) to which we apply our framework. We treat a number of important considerations in the design of the GAN in order to preserve a high degree of utility in facial images from a largely unconstrained input distribution. We then show in Section 6.5, how to apply differential privacy to image encodings within a transformed principal component analysis (PCA) basis in order to improve the preservation of utility in the obfuscated output. In Section 6.6, we explore the interpretation of the privacy guarantee we provide, discussing parallels to other applications of differential privacy and providing guidance on the intuitive implications of the guarantee. To broaden the applicability of our approaches, we propose in Section 6.7 the use of an exponential mechanism to alter pixel intensities in images, allowing for the obfuscation of any image (i.e., not restricted to facial images). In Section 6.8, we give the results from experimental comparisons of our implementations against other methods of facial obfuscation that also offer formal privacy guarantees. Finally, in Section 6.9, we discuss potential extensions and continuations of our work and in Section 6.10, we conclude the chapter.

The content given in Sections 6.2 and 6.3 is largely derived from our publication in the International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE) 2019 titled “Differentially Private Obfuscation of Facial Images” [39]. The content given throughout Sections 6.4, 6.5 and 6.6 comes from our paper titled “Differentially Private Facial Obfuscation via Generative Adversarial Networks”. The experimental comparisons in Section 6.8 are combined from both of these papers. We refer the reader back to Chapter 3 for details on these submissions.

6.1 Notation and Definitions

In this section, we provide details on background material of high relevance to our work. We begin with a distance-based generalization of differential privacy which we employ in order to obtain a

differentially private guarantee within the domain of generative model encodings. Following this, we cover the formulation of the training objectives for AttGAN, a GAN used for the modification of facial attributes in images. We extend and adapt AttGAN in our work in order to handle obfuscation of facial images from largely unconstrained input distributions. Lastly, we review some details on PCA which we later draw on to better control the application of noise to image encodings. Throughout this section, we cover the relevant notation from these sources which we apply and extend throughout the chapter.

6.1.1 Distance-Generalized Differential Privacy

Often, sensitive information exists in forms other than as records of a database. However, since the differential privacy guarantee is defined in terms of adjacency, a concept specific to the domain of databases, it cannot be directly applied outside of this context. A more general conceptualization of sensitive information is as a structured representation of data, referred to as a secret, pertaining to an individual. A generalization of differential privacy [27] provides a privacy guarantee for such secrets by extending the concept of adjacency in databases to distance between secrets. The key intuition of the generalization is that an appropriate distance metric acts as a measure of distinguishability between secrets and thus takes the place of query sensitivity in the configuration of noise-adding mechanisms. By controlling noise according to an appropriate distance metric, the privacy guarantee is adapted to ensure that similar secrets are highly indistinguishable while very different secrets remain distinguishable. For a pair of databases, the distance between them is the number of records by which they differ. For other types of secrets, the distance metric must be carefully chosen in order to provide an appropriate privacy guarantee.

Let $\mathbb{X}$ be the domain of secrets, $K : \mathbb{X} \rightarrow \mathbb{X}$ be a randomization mechanism and $d : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ be a measure of distance between pairs of secrets. The distance-generalized privacy guarantee is given as follows:

$$\Pr(K(X_1) = R) \leq e^{\epsilon d(X_1, X_2)} \Pr(K(X_2) = R), \quad \forall X_1, X_2, R \in \mathbb{X}. \quad (6.1)$$

Comparing this to the standard formulation of differential privacy (Definition 1), the databases D_1 and D_2 have been replaced by secrets X_1 and X_2 and the distance function now appears in the exponent of the multiplicative factor e^ϵ . The distance between any pair of secrets acts as a coefficient to ϵ when interpreting the ratio of their probabilities. Intuitively, the meaning is that the more similar a pair of secrets are to each other, the harder is it to determine which of them led to a given obfuscated instance. This hampers the accuracy with which attempts at re-identification can be made.

6.1.2 AttGAN

Manipulation of facial images is a common task in the machine learning community and has been treated quite successfully in recent works using GANs. In particular, AttGAN [95] has shown excellent results in the modification of high level facial attributes (e.g., gender, hair colour, etc.) while preserving facial identity.

AttGAN uses an encoder/decoder pair which is trained alongside a discriminator in an adversarial manner. The encoder takes an image as input and passes it through a series of convolutional layers to generate an encoding which represents high level features detected in the image. The decoder accepts an encoding along with a vector of supplementary facial attributes (e.g., gender) and passes the concatenated input through a series of transpose convolutional layers in order to produce a new image as output. The generated image is intended to resemble the original as closely as possible

while differing only in the changes to facial attributes specified by the supplementary vector fed to the decoder.

The discriminator is also a convolutional network which accepts an image as input. It serves two purposes. After matching the architecture of the encoder, it splits into two branches of fully connected layers. The first branch acts as the discriminator while the second branch acts as a classification network for the facial attributes present in the image. The outputs from these branches act as measures of how successfully the generator (i.e., the encoder/decoder pair) is performing its task. The generator and discriminator take turns updating their parameters during the training process in order to learn from each other.

Using the notation as defined in the original paper (with minor adjustments for convenience), we provide the full formulation of the objective function used to train the model. In Section 6.4, we extend this formulation to suit our intended usage for facial obfuscation.

Let G_{enc} and G_{dec} be the encoder and decoder networks, respectively, and let x^a be an image depicting a face that has facial attributes matching the specification of a vector $a = [a_1, \dots, a_m]$ of binary attributes. Let $z = G_{enc}(x^a)$ be the encoding of x^a and let $x^{\hat{b}} = G_{dec}(z, b)$ be the image produced by the decoder when given an encoding z and a new vector of binary attributes $b = [b_1, \dots, b_m]$. Finally, let D be the discriminator network (referred to in some works as a critic) and C be the classification network.

The adversarial loss portion of the objective function employs a Wasserstein critic with a gradient penalty [85]. We denote the gradient penalty function as GP and use λ_{GP} to indicate the hyperparameter that controls the weight of the gradient penalty. For a distribution p_{data} of training images and a distribution p_{attr} of binary attribute vectors, the discriminator and generator losses defined as follows:

$$L_{adv_d} = -\mathbb{E}_{x^a \sim p_{data}} D(x^a) + \mathbb{E}_{x^a \sim p_{data}, b \sim p_{attr}} D(x^{\hat{b}}) + \lambda_{GP} \mathbb{E}_{x^a \sim p_{data}, b \sim p_{attr}} GP(x^a, x^{\hat{b}}), \quad (6.2)$$

$$L_{adv_g} = -\mathbb{E}_{x^a \sim p_{data}, b \sim p_{attr}} D(x^{\hat{b}}). \quad (6.3)$$

The attribute loss is calculated using cross entropy. The prediction of the classification network for the i^{th} binary attribute of an image x is given by $C_i(x)$. The losses for the classifier and generator are defined as follows:

$$\ell_a(x, a) = \sum_{i=1}^m -a_i \log C_i(x) - (1 - a_i) \log(1 - C_i(x)), \quad (6.4)$$

$$L_{cls_c} = \mathbb{E}_{x^a \sim p_{data}} \ell_a(x^a, a), \quad (6.5)$$

$$L_{cls_g} = \mathbb{E}_{x^a \sim p_{data}, b \sim p_{attr}} \ell_a(x^{\hat{b}}, b). \quad (6.6)$$

Finally, reconstruction loss is calculated as mean absolute error between scaled pixel intensities in the original and generated images to promote accurate reconstruction of non-altered details in the image:

$$L_{rec} = \mathbb{E}_{x^a \sim p_{data}} \|x^a - x^{\hat{a}}\|_1. \quad (6.7)$$

These three objectives are then combined into the full objective function. The λ coefficients denote hyperparameters used to configure the relative importance of each objective.

$$\min_{G_{enc}, G_{dec}} \lambda_{rec} L_{rec} + \lambda_{cls_g} L_{cls_g} + L_{adv_g}. \quad (6.8)$$

$$\min_{D, C} \lambda_{cls_c} L_{cls_c} + L_{adv_d}. \quad (6.9)$$

6.1.3 Principal Component Analysis

PCA is a method of analyzing a set of observations (e.g., a dataset) in order to identify patterns and transform the data representation into a format in which it is easier to observe the patterns. This is typically done to assist in machine learning and data visualization tasks. We provide here a brief overview of how to extract and apply a PCA transformation matrix. For greater detail, we refer the reader to [178].

Many patterns in a dataset arise due to correlations between variables and therefore can be determined through analysis of the covariance matrix of the dataset. PCA involves computing a decomposition of the covariance matrix, either via eigendecomposition or singular value decomposition. The decomposition is used to calculate a transformation matrix which maps the original dataset into a new basis defined by the eigenvectors of the covariance matrix. Since the covariance matrix is a square matrix with dimensions that correspond to the number of variables in the dataset, the transformed basis will have a number of dimensions that matches the original dimensionality of the data. The dimensions of the new basis are referred to as principal components. Let D be a matrix representing the original data in row format and let Z be a matrix made from the eigenvectors of the covariance matrix arranged in descending order of their eigenvalues. The transformed data is given by $D' = Z^T \times D$.

The benefits of this transformation are twofold. First, the eigenvectors taken from the covariance matrix are orthogonal, meaning the principal components are linearly uncorrelated. This can help to account for redundancy in the data due to highly correlated features. Second, each eigenvector has an eigenvalue associated with it which reflects the ratio of the variance from the original dataset that is expressed in the corresponding principal component. The total variance of the dataset is preserved in the PCA transformation. By comparing the eigenvalues, it is therefore possible to compare the information content of each of the principal components. Since these are ranked from largest to smallest, the information content of each transformed data record monotonically decreases from the first principal component to the last. This allows methods of data analysis to focus on the leading principal components where the highest information content exists.

Since the ratios of explained variance typically decrease rapidly over the principal components, many of the later principal components (based on their order) can be dropped to reduce dimensionality while incurring very little information loss. The effect of dropping principal components is that data points in the reduced dimensional representation assume the mean value of the dataset in the omitted dimensions. For low variance principal components, this alteration has little impact on the data.

Data that has been transformed to the principal component basis can be transformed back to its original basis by applying an inverse transformation. This is given as $D = Z^T \times D'$. This is possible to do even when dimensionality reduction has been applied. If all dimensions have been preserved, the inverse transformation is lossless. If trailing principal components have been dropped, information loss is minimized in the sense that the omitted principal components were those that expressed the least amount of variance in the original data.

PCA and Differential Privacy

When applying differential privacy to multi-variate data that has strong correlations between the variables, the independent application of noise to each variable can destroy a great deal of useful information [103]. As the magnitude of noise is increased, the correlations become less pronounced in the noisy data, leading to perturbed data points that are increasingly likely to lie outside of the original distribution. This, in turn, is likely to produce obfuscated releases that hold little value from a data analysis perspective.

The application of noise within a transformed PCA basis of reduced dimensionality can help to better preserve correlations in the data while simultaneously spending the privacy budget more efficiently [57, 103]. The use of principal components that maximize explained variance within the training data ensures that the addition of noise along the axes of the transformed basis will produce noisy data that more closely resembles the original distribution. Furthermore, the ranking of the principal components by their explained variance ratios allows for a convenient means to improve the allocation of the privacy budget. Trailing axes with low explained variance ratios can be dropped without great loss of information, allowing the privacy budget to be focused on the axes with greater information content. The property of linearly uncorrelated principal components helps to avoid spending the budget in a redundant manner (as would be the case when it is distributed over highly correlated variables) and thus simultaneously reduces the potential for degradation of correlations in the obfuscated data. We refer the reader to [103] for a visualization of these benefits realized on a two-dimensional example.

6.2 Weaknesses in Existing Facial Obfuscation

To motivate our decision to apply differential privacy to generative models for images, we first provide some discussion on weaknesses in other approaches to facial obfuscation. We argue that the existing approaches to facial obfuscation do not offer a satisfactory trade-off between the aspects of privacy and utility. We first briefly discuss non-formalized approaches to facial obfuscation in order to emphasize the need for a formalized privacy guarantee. We then discuss in detail the k -same privacy guarantee and highlight a number of deficiencies in the guarantee when it is interpreted in practical scenarios. Lastly, we discuss other formalized approaches to facial obfuscation which offer strong guarantees of privacy at the cost of utility in the obfuscated images.

6.2.1 Non-Formalized Approaches

While ad hoc approaches to facial obfuscation such as blurring and pixelization may be shown to be empirically effective in some cases [160], they do not formalize an attack model and offer no guarantee regarding their effectiveness in general. Similarly, methods of face swapping (e.g., [122, 151]) provide no guarantee regarding potential attacks that might be launched against them. Furthermore, face swapping raises privacy concerns about the usage of real faces or components of real faces that are swapped into the released images.

Given the importance of preserving privacy in images, a good method of obfuscation must assert a meaningful guarantee about the level of privacy it provides. Empirical results may help to gain intuition on which approaches appear promising. However, without a formal guarantee to back up the results, it is impossible to assert that privacy will remain protected in untested scenarios or against unknown attacks. For this reason, we restrict our attention to focus only on methods of obfuscation that offer a formal privacy guarantee.

The necessity of this restriction is underscored by the concept of parrot attacks [146]. A parrot attack uses a machine learning model to classify identities in images using labeled instances of obfuscated images as the training set. Having learned about patterns in the obfuscation during training, the network is made much more effective at defeating the obfuscation. Pixelization, for example, despite being reasonably effective against human recognition and even naive machine recognition, can be completely defeated by a parrot attack [134, 146]. This formed a strong basis for the need of a formal privacy guarantee such as that provided by the k -same family.

6.2.2 The k -same Privacy Guarantee

The k -same [146] family of approaches employs a privacy guarantee derived from k -anonymity [172] which asserts that the original identity for any obfuscated image is indistinguishable from at least $k - 1$ others. This guarantee is a result of the obfuscation process which draws upon clusters of k or more images to produce averaged instances as replacements for all images in the cluster. This provides a theoretical guarantee in which the probability of re-identification is upper-bounded by $\frac{1}{k}$. However, the k -same guarantee relies on assumptions about the nature of the attack. We discuss these assumptions and show why they are often unrealistic in practice, making the guarantee weaker than it appears to be.

Background Knowledge

A well-known deficiency of k -anonymity is its susceptibility to attacks that employ background knowledge [18]. This refers to cases where the attacker uses prior knowledge about the sensitive information to draw inferences that violate the privacy guarantee. For example, in the context of a database of hospital records, k -anonymity would typically be applied to create groups of database records that are indistinguishable on their demographic attributes but with the original medical condition (i.e., the sensitive information) of each record preserved for statistical analysis. If an attacker already knows the medical conditions of one or more individuals (e.g., friends or family members), they can eliminate the corresponding records from the anonymized groups by finding the matching demographic information and medical condition. If the removal of records reduces the size of an anonymized group to less than k , this violates the privacy guarantee.

This concept carries directly over to the k -same privacy guarantee. If, via prior knowledge, the attacker knows with certainty that some of the k individuals could not be in the obfuscated image, they can discount them from the set of k identities. An attacker could come by this knowledge in a number of ways: personal knowledge about friends and family, information scrapped from other data sources such as social media, etc. The simple combination of knowledge about the time at which a photo was taken and the approximate locations of some of the k individuals at that time can be enough to derive a proper subset of the k individuals which violates the privacy guarantee.

Contextual information in an image can often enable these types of inferences. Using signs, architecture or landscapes in an image, an attacker might recognize the location or employ software to determine it. Knowledge about locations that individuals frequent may greatly increase the probability of some possibilities over others. Similarly, if some of the k identities are known to live in different cities than where the photo was taken or worse yet, different continents, these identities become much less probable. Other cues such as accessories or clothing on obfuscated individuals may also greatly impact the probabilities accorded to the k possible identities. Since the privacy guarantee asserts that each of the k identities are equally probable, this is also in violation of the guarantee.

We note that the original k -same paper does acknowledge this vulnerability to contextual information and asserts that the privacy guarantee applies strictly to the information contained within the face, not to the image as a whole [146]. While this important distinction allows for the privacy guarantee to be upheld, it is a major restriction on the practical applicability of the k -same guarantee. Most contexts in which facial obfuscation is applied will be rich with contextual information, making the privacy guarantee much less meaningful.

Composition Attacks

Another deficiency of k -anonymity is a susceptibility to composition attacks [73]. This is a class of attacks that exploit information from multiple, potentially uncoordinated, obfuscated releases of data to violate the privacy guarantee. A simple instance of this is the intersection attack. An attacker first identifies the clusters in which a particular individual exists from two different releases. If the releases were uncoordinated, the clusters likely differ, allowing the attacker to take their intersection to achieve a new set with a cardinality less than k .

This attack again carries directly over to the k -same approach. Consider a scenario where an individual takes a photo that they wish to upload to social media. Privacy protection might be applied to the individual or perhaps to bystanders who were captured in the background of the photo. Should the individual decide to upload the same photo to two or more social media platforms, the issue of uncoordinated obfuscation immediately arises. An attacker needs only scrape these platforms for similar photos to apply an intersection attack.

Intersection attacks may even be effective for multiple releases from the same organization if care is not taken. For example, an individual may take consecutive photos and then upload all of them. Algorithms for k -same determine clusters based on the similarity of faces but many factors beyond facial identity (e.g., pose, angle and lighting) could impact similarity. It is therefore not unlikely that multiple images of the same individual will result in different clusters. Sequences of images uploaded in this way would be an ideal target for intersection attacks.

Most k -same approaches require each individual to appear only once in the gallery of images to be obfuscated. This prevents intersection attacks for releases from the same organization but does not protect against uncoordinated releases across multiple organizations. Furthermore, enforcing this restriction may be very challenging in practice. While the primary subject in a photo might be determined based on the account used to upload the photo, other individuals in the photo cannot be correctly identified 100% of the time. Face recognition software has not yet reached this level of accuracy. Without manual labeling, such a policy cannot be enforced. Beyond this, the restriction of one image per identity is very severe and does not match typical use cases for image sharing.

Other Difficulties

We discuss here two other difficulties that arise when using k -same obfuscation in practice. Although these difficulties do not violate the privacy guarantee, they hinder meaningful applications for k -same obfuscation in some contexts.

The first problem arises from the requirement of an input gallery of images. This may be appropriate for scenarios where batches of images are obfuscated but it is awkward to apply to cases where images are sporadically uploaded (e.g., in social media platforms). One might consider the use of a preloaded static gallery or even a dynamic gallery that gets updated as new images are uploaded. This, however, is not a good solution since identities can then participate in more than one cluster. Furthermore, if an attacker records information about identities known to be in the gallery, those identities can be discounted when an image is uploaded for a new identity. An alternative solution could rely on buffering uploaded images to form a gallery that can eventually be

used to release a batch of obfuscated images. However, this necessitates a trade-off between the size of the gallery (and thus the quality of the output) and ability to deliver a timely service. In an era where users expect images to be uploaded instantly, this is not likely to be a manageable trade-off. The release of multiple batches also increases the chances of enabling composition attacks.

The second problem relates to the preservation of utility in the obfuscated output. Approaches that partition the gallery according to classes to be preserved (e.g., combinations of age and gender) place an even greater strain on the input gallery requirement. Working separately with the subset of images from each class greatly reduces the number of images available for clustering. Such an approach is not scalable for large numbers of classes that would be needed for finely grained attention to utility. In the worst case, some classes may be outliers in the overall distribution and could lack sufficient images to form a cluster. These classes would have to be merged with others in order to achieve the k -same guarantee, thus failing to achieve the desired granularity of classes.

6.2.3 Other Formalized Alternatives

We conclude this section with a brief discussion on some of the more recent works in facial obfuscation which demonstrate the potential for a stronger guarantee of privacy but which have been unable to realize a favourable balance between privacy and utility. We re-iterate that perfect privacy can be achieved through complete redaction of information, yet this is undesirable as it simultaneously destroys all utility in the image. A good trade-off between these two aspects is of key importance for facial obfuscation to find practical usage.

Methods of redaction followed by in-painting (e.g., [102, 188]) offer a guarantee of perfect privacy protection within a redacted area by removing all information and allowing a machine learning approach to rebuild an appropriate image from contextual information. A well-trained model can produce visually convincing results that resemble realistic novel faces. Since the process has no access to the original information from the redacted area, there can be no leakage of this information. Yet the contextual information around the redacted area (often the contour of the head including forehead, hair and ears) is likely to reveal too much sensitive information. As such, redaction that is tightly cropped to the face is insufficient to offer a meaningful level of privacy. The area of redaction could be increased to account for this, however, this places a greater strain on the ability of the model to in-paint the area in a manner that is both visually clean and able to preserve utility. Methods of in-painting, in their current form, do lend themselves well to a balanced trade-off between privacy and utility.

Some previous works have explored the applicability of differential privacy to the obfuscation of images. While this gives the potential for strong levels of privacy, the values of ϵ required to realize this potential result in poor levels of utility. The application of differential privacy to pixel intensities [65] leads to highly distorted and pixelized images at the values of ϵ empirically shown to achieve low levels of re-identification risk. The application of differential privacy to the singular value decomposition of a matrix of pixel intensities [64] similarly leads to highly distorted images that no longer resemble realistic human faces. Furthermore, this approach applies noise only to the singular values while leaving the other matrices of the decomposition untouched. Due to this, there is potential for the leakage of sensitive information in a manner that is not captured by the formal privacy guarantee.

Lastly, we re-iterate a point made in Section 2.2.3. Differential privacy has been applied in the context of training GANs in a privacy-preserving fashion and has been used in this way for the generation of synthetic datasets of facial images (e.g., [193, 211]). However, this is fundamentally different from the goal of facial obfuscation in which privacy must be guaranteed for new images presented to the network after training is complete.

6.3 Differential Privacy for Generative Models

Due to the deficiencies of the k -same privacy guarantee in practical applications of facial obfuscation, we argue that a more robust privacy guarantee is required. Further to this, we argue that the privacy-utility trade-offs in existing formalized approaches to facial obfuscation are not satisfactory. Following the advances in the field of database privacy, we consider the potential of differential privacy to provide a strong privacy guarantee in facial obfuscation. Differential privacy has real world applications in organizations including the US Bureau of Census, Google and Apple [152] and has been studied as a means to comply to legal definitions of privacy such as FERPA [153] and GDPR [165]. In this section, we propose how to apply a distance-based generalization of differential privacy in order to obfuscate representations of images within generative machine learning models. The use of appropriate generative models allows us to preserve visual realism and selected types of non-sensitive information in the obfuscated images. We formalize a framework to achieve differentially private obfuscation of facial images and discuss how the derived privacy guarantee addresses the issues identified with the k -same approach. We conclude the section by applying our framework to implement differentially private facial obfuscation using an active appearance model (AAM) [35] and a generative neural network (GNN) [47].

6.3.1 Framework for Generative Models

We propose that a strong guarantee of privacy paired with a high level of utility (i.e., visual quality and preservation of information) can be achieved through the application of differential privacy to the encodings of generative models for images. To this end, we begin by considering how differential privacy can be applied in this context. A generative model can represent images of instances from specific classes (e.g., human faces) using a numeric representation that abstracts from pixel intensities. Our goal is to protect the privacy of individuals in images by modifying these numeric representations to prevent facial identification while maintaining utility and visual quality. Differential privacy is ideal for this purpose as it provides a robust guarantee against the accuracy of the inferences an attacker can make about the original data. To suit our setting, in which sensitive data is represented as a numeric encoding for a generative model, we apply the distance-based generalization of differential privacy (Section 6.1.1) which extends the framework of differential privacy to domains of sensitive data beyond the context of databases. The application of noise to the model encoding allows for the generation of photo-realistic instances of novel human faces. This avoids the significant degradation in visual quality that results from the addition of noise to pixel intensities.

The notion of distance between secrets, as defined for distance-based differential privacy, is appropriate for the representation of images within a generative model. Any model that employs a numeric representation of images allows for the calculation of distance between images. While the exact representation of an image differs from model to model, they can generally be mapped to a vector of fixed length with little difficulty. We provide details on how this concept can be applied to both AAMs and GNNs in Section 6.3.3. To develop a general framework here, we consider the representation of an image to be a vector $X \in \mathbb{R}^n$, where $n \in \mathbb{Z}^+$, and the randomization mechanism to be a function $K : \mathbb{R}^n \rightarrow \mathbb{R}^n$ used to produce an obfuscated instance of the image. Thus, we can apply the distance-generalized privacy guarantee of Formula 6.1 by interpreting the domain of secrets $\mathbb{X}$ as being $\mathbb{R}^n$ and by using a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$.

To enforce a meaningful privacy guarantee, we must determine an appropriate distance metric to measure the distinguishability of the numeric representations of images. A natural choice for the distance metric is L_1 distance, however, we must be wary of the meaning of each element in the vectors. Should certain elements have differently sized ranges, they should be obfuscated using

different magnitudes of noise. If one element has a much larger range than the others, the addition of noise configured to the smaller range would do little to prevent an inference of high accuracy on the original value of the element. We therefore apply normalization such that the distance between any pair of elements in the i^{th} position of a pair of vectors falls within the range $[0, 1]$. Letting $R_i = [i_{min}, i_{max}]$ be the range of elements in the i^{th} position of a model representation vector, we define a normalized, element-wise distance metric as follows:

$$d_e(x_1, x_2) = \frac{|x_1 - x_2|}{i_{max} - i_{min}}, \quad \forall x_1, x_2 \in R_i. \quad (6.10)$$

A distance metric for vectors defined as the sum of the element-wise distances for each position would be appropriate for images represented by the same model. However, a more useful framework would allow for reasoning about the level of privacy across different models. Ideally, the meaning of a privacy parameter ϵ applied to one model should have a similar meaning for a different model. For this, we require another normalization to account for models having vectors of different lengths. We therefore define the distance metric for vectors as follows:

$$d(X_1, X_2) = \frac{\sum_{i=1}^n d_e(X_{1i}, X_{2i})}{n}, \quad \forall X_1, X_2 \in \mathbb{R}^n. \quad (6.11)$$

By using this distance metric in combination with Formula 6.1, we obtain a meaningful privacy guarantee for the model representations of images. Although this type of metric is, in general, not novel, its use in this context is. Existing work employing the distance-based generalization of differential privacy only deals explicitly with one and two-dimensional secrets [27] whereas the domain of secrets we deal with consists of n -dimensional vectors. We must therefore address how to configure a mechanism to satisfy this instantiation of the privacy guarantee. This leads to our main result in the development of a framework for the application of differential privacy to generative models for images.

Theorem 8. *Any image $X \in \mathbb{R}^n$ can be protected by ϵ -differential privacy through the addition of a vector $(Y_1, \dots, Y_n) \in \mathbb{R}^n$ where each Y_i is a random variable independently drawn from a Laplace distribution using a scaling parameter $\sigma_i = \frac{n(i_{max} - i_{min})}{\epsilon}$.*

Proof. We must satisfy the privacy guarantee (Formula 6.1) using our proposed distance metric (Formula 6.11). The form this privacy guarantee takes is our starting point in Formula 6.12. Through manipulation of this inequality and the substitution of mechanism probabilities with a Laplace distribution, we prove that the selection of an appropriate scaling parameter for each instance of the Laplace distribution allows for the privacy guarantee to be satisfied.

$$\prod_{i=1}^n \Pr(K(X_{1i}) = R_i) \leq e^{\frac{\epsilon \sum_{i=1}^n d_e(X_{1i}, X_{2i})}{n}} \prod_{i=1}^n \Pr(K(X_{2i}) = R_i), \quad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (6.12)$$

$$\prod_{i=1}^n \Pr(K(X_{1i}) = R_i) \leq \prod_{i=1}^n e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}} \prod_{i=1}^n \Pr(K(X_{2i}) = R_i), \quad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (6.13)$$

$$\prod_{i=1}^n \frac{e^{-\frac{|X_{1i}, R_i|}{\sigma}}}{2\sigma} \leq \prod_{i=1}^n e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}} \prod_{i=1}^n \frac{e^{-\frac{|X_{2i}, R_i|}{\sigma}}}{2\sigma}, \quad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (6.14)$$

$$\prod_{i=1}^n e^{\frac{|X_{2i}, R_i| - |X_{1i}, R_i|}{\sigma}} \leq \prod_{i=1}^n e^{\frac{|X_{2i} - X_{1i}|}{\sigma}} \leq \prod_{i=1}^n e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}}, \quad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (6.15)$$

$$\prod_{i=1}^n e^{\frac{\epsilon de(X_{1i}, X_{2i})}{n}} = \prod_{i=1}^n e^{\frac{\epsilon |X_{2i} - X_{1i}|}{n(i_{max} - i_{min})}}, \quad \forall X_1, X_2. \quad (6.16)$$

From Formula 6.16, it becomes clear that the inequality holds when using an independent Laplace distribution for each pair of elements X_{1i}, X_{2i} , substituting the scaling parameter σ with a corresponding value $\sigma_i = \frac{n(i_{max} - i_{min})}{\epsilon}$. □

Using the generalization of differential privacy, the notion of query sensitivity is implicitly captured in the distance metric. Since the distance metric of Formula (6.11) has a range of $[0, 1]$, the ratio of probabilities for a pair of maximally dissimilar images to produce the same obfuscated output is e^ϵ . This is akin to the meaning of the privacy guarantee for a pair of databases that differ on every record. In order to select an appropriate value of ϵ , a data custodian must keep in mind that similar images will have a very small distance between them, requiring much larger values of ϵ to provide a reasonable ratio. In Section 6.8.1, we demonstrate the implications of the choice of ϵ on the levels of privacy and utility.

6.3.2 Benefits of Differentially Private Facial Obfuscation

We now describe the improvements we obtain from the use of differential privacy for each of the problems identified in Section 6.2.

Background Knowledge

By removing dependence of the attack model on an absolute level of re-identification risk, we are able to reason about the level of privacy in the presence of attackers with background knowledge. If the location in a photo is identified as a particular city, no facial obfuscation can prevent the inference that individuals living in the identified city have a higher probability of being the obfuscated identity than individuals living elsewhere. Yet, the differential privacy guarantee continues to hold as the background knowledge does not impact the conditional probability distribution used by the randomization mechanism. Since the privacy guarantee concerns only the change in the attacker's knowledge when presented with the obfuscated data (e.g., the face), it is unaffected by other sources of information the attacker may gain access to.

Composition Attacks

Another very important property of differential privacy is its resilience to composition attacks. The composition theorem [56] states that for two differentially private releases using privacy parameters ϵ_1 and ϵ_2 respectively, the privacy guarantee holds for a privacy parameter $\epsilon = \epsilon_1 + \epsilon_2$. Thus, even in the case of uncoordinated releases, we still have a valid privacy guarantee. Furthermore, this removes the restriction on the same individual appearing only once in the release of obfuscated images.

Input Image Gallery

Differentially private image obfuscation has no need for a gallery of images in order to perform obfuscation. Since noise is added on a per-image basis, there is no computation of clusters required. Given a trained model, obfuscation of a single image or a batch of images can be performed with ease. This makes the obfuscation process much more versatile.

6.3.3 Implementation Details

AAMs and GNNs have the very useful property of producing photo-realistic images. We now describe how our framework can be applied to these models. Provided that the addition of noise is properly controlled, the output will be a photo-realistic image of any newly created identity.

The AAM representation of an image consists of a shape vector and a texture vector. It is important to note that this gives a representation of the identity that is strictly contained within the contour of the face, leaving features such as hair and ears as contextual information that is untouched by the obfuscation. This is not ideal for the goal of hiding an identity since this contextual information can greatly facilitate inferences about the identity. Although in theory an AAM could be designed to incorporate the hair and ears, we are unaware of any research in which this has been done. Thus, although we include AAMs in our experimental comparisons, we recommend the use of GNNs instead. To include AAMs in our experiments, we take the concatenation of the shape and texture vectors as the overall model vector.

For GNNs, we consider architectures that take one or more class vectors as input and employ up-convolution to transform the input into a visual representation in pixel space [47]. By considering each identity to be a different class, an input vector can specify the individual to be generated. The identity class vector is an obvious choice as the model vector to be obfuscated. However, this leads to some form of interpolation between the identities. To apply a finer degree of modification to the identity, we propose the application of obfuscation at the second layer of the network. Typically, the second layer applies convolution to the class vector and transforms it into a vector of numeric representations of high level facial features. By applying obfuscation to these features instead, we can achieve a richer variety in the potential modifications to the face. We therefore apply obfuscation to the output of the first convolutional layer of the network and pass the obfuscated feature vector on as the input to the next layer of the network. A sample architecture is shown in Figure 6.1.

Information about the range of each model vector element can be used as a means to preserve the visual quality of the obfuscated output. Noisy elements that have gone too far beyond the valid range may lead to visual artifacts or distortions in the output image. To prevent this, we snap any out-of-bounds noisy value back to the nearest valid value. Since differential privacy is resistant to any form of post-processing [56] and the ranges of the elements are non-sensitive information, this step cannot violate the privacy guarantee.

6.3.4 Obfuscation of Unseen Images

When designing a system for the obfuscation of identities in images, an important consideration is the ability to obfuscate any identity. If the chosen generative model is an AAM, this is easily achieved through the ability of the AAM to approximate the model parameters for the representation of any unseen image. However, the GNN architecture does not directly allow for the representation of classes that were not learned during the training process. Since each identity is a different class, this means that the model cannot directly represent unseen identities and thus cannot obfuscate them.

To solve this problem, we propose to formulate the approximation of an appropriate input vector for an unseen identity as an optimization problem. Since the GNN has learned a representation of each training identity, a new identity can be approximated as a weighted sum of the known identities. When provided with these weights as the input identity vector, the GNN would produce an interpolation between the identities which can act as an approximate visual representation of the unseen identity. In order to formalize this concept as an optimization problem, we must select an appropriate representation for the identities. This cannot be done using the GNN feature vectors since the representation of the target identity is unknown. We therefore employ a secondary neural

Figure 6.1: Visualization of the layer architecture in an up-convolutional neural network using differential privacy. Noise is applied to the output of the second identity layer. The numbers and shapes of the convolutional layers shown here are not exact and represent only the general structure of such a network.

network that has been trained for classification of facial identity. By removing the final layer of the classification network, its output becomes a high-level vector of facial features which we can use as the representation of an identity. The significance of using a classification network for this purpose is that it need not have seen any of the target identities in its training data in order to produce feature vectors for them. Such a network is therefore ideal to provide numeric representations of the identities from the GNN training data as well as the target unseen identity.

We now formalize the optimization problem using real-valued feature vector representations of identities from a classification network. Let $\mathbb{X}$ be a set of n -dimensional feature vectors representing m identities on which the GNN was trained (i.e., $|\mathbb{X}| = m$) and let Y be an n -dimensional feature vector representing an identity that is unknown to the GNN. Our goal is to determine a set W of weights that minimize the distance between a weighted sum of the vectors in $\mathbb{X}$ and the vector Y .

$$\min_W \left\| \sum_{i=1}^n W_i \mathbb{X}_i - Y \right\|. \quad (6.17)$$

While this optimization problem is similar to the format of an objective function for a linear program, the necessity of absolute value calculations for the L_1 distance between the vectors prevents this from being written as a linear function. However, this problem, known as least absolute deviations, can be rewritten in an alternate but equivalent formulation that avoids the need for absolute value functions through the introduction of additional variables [67]. This formulation is as follows:

$$\min \sum_{i=1}^n u_i, \quad (6.18)$$

subject to the following constraints:

$$u_i \geq y_i + \sum_{j=1}^n W_j \mathbb{X}_j, \quad i = 1, \dots, m, \quad (6.19)$$

$$u_i \geq - \left(y_i + \sum_{j=1}^n W_j \mathbb{X}_j \right), \quad i = 1, \dots, m. \quad (6.20)$$

The added constraints ensure that the values assigned to the new variables respect the absolute value functions from the original problem. This formulation can be given to any linear programming solver in order to find the optimal weights for the approximation of the unseen identity represented.

6.4 Differentially Private Obfuscation Via a GAN

To truly take advantage of the ability of generative models to realistically capture varied distributions of images, we focus on applying some of the recent successes from the study of GANs. These networks have demonstrated excellent results for the task of manipulating facial images (e.g., [32, 95]). However, the model must be carefully designed and trained in order to effectively execute its intended task. Throughout this section, we cover important considerations for the structure of the image encoding, model architecture, training process and image pre/post-processing. We describe these details in the context of an extension and adaptation of AttGAN, however, the concepts can equally be applied to other models. Samples of obfuscated images are shown in Figure 6.2 to demonstrate the impact each of our major design considerations have on the obfuscated output.

We define the following additional notation to be used throughout this section. Let $x^{a,p}$ be an image depicting a face that has facial attributes matching the specification of a vector $a = [a_1, \dots, a_m]$ of binary attributes and a pose matching the specification of a vector $p = [p_1, \dots, p_k]$ of real-valued parameters. Let N be a noise-adding function and let $\tilde{x}^{\hat{b},\hat{q}} = G_{dec}(N(z), \hat{b}, \hat{q})$ be the image produced by the decoder when given a noisy encoding $N(z)$, a new vector $\hat{b} = [\hat{b}_1, \dots, \hat{b}_m]$ of binary attributes and a new vector $\hat{q} = [\hat{q}_1, \dots, \hat{q}_k]$ of real-valued pose parameters.

6.4.1 Image Encoding and Model Architecture

We begin by considering the design of the image encoding used by the model. Using our proposed framework from Section 6.3.1, we will be adding noise to the encodings of images in order to induce changes to high level facial features as a means of obfuscation. The encoding must therefore be appropriately receptive to such noise. Regardless of the shape of the encoding, it is ultimately a multi-dimensional representation of numeric features to which noise can be added. However, the shape plays an important role in determining how visual quality is impacted by the addition of noise. Consider the $4 \times 4 \times 1024$ encoding used in AttGAN. This can be interpreted as 1024 many 4×4 grids where each grid is a spatially compressed representation of the presence of some type of feature in the image. The 4×4 spatial dimensions present two problems. The first is that each cell of the grid relates to a local part of the image, meaning noise applied to a grid cell will have a localized effect rather than modifying the image at a global level. The second issue is that when convolutional filters

Figure 6.2: An example of obfuscated instances of an image with each row showing results from a different architecture and training configuration. Row a) uses the original $4 \times 4 \times 1024$ encoding with no changes to the model training. Each subsequent row incorporates one cumulative modification to the model. Row b) uses a $1 \times 1 \times 2048$ encoding with no changes to the model training. Row c) adds noise to encoding samples during training. Row d) uses additional inputs to preserve pose. Row e) applies background masking.

are applied to overlapping areas of the output from the previous network layer, the resulting features (grid cells) describe overlapping areas and are thus highly correlated. Even when applications of a filter do not overlap (i.e., the convolutional stride matches the filter size), correlation between features remains likely due to the close spatial proximity of the filter applications. When noise is added to the encoding, the independent modification of features destroys their correlations with each other. Since this is done to features representing local areas in the image, the result is a highly distorted image which no longer resembles a human face. An example of this distortion can be seen in Figure 6.2-a.

To avoid distortions induced by inconsistencies between previously correlated features representing local areas in the image, we propose that noise should instead be added to global facial features. Further compression of the encoding down to 1×1 spatial dimensions guarantees that each numeric value in the resultant encoding is a representation of a feature at the global level. This eliminates the potential for inconsistencies at the local level since the correlations are now compressed into a single value. Our proposed encoder/decoder architecture which implements this approach is given in Table 6.1. The improvement in visual quality from this change to the architecture can be seen in the difference between rows a and b of Figure 6.2.

Encoder	Decoder
Input: $128 \times 128 \times 3$ Image	Input $1 \times 1 \times 2048$ Encoding
Conv ($4 \times 4 \times 64$, Stride=2), BN, LReLU	Concat Binary Attributes, Concat Pose Parameters
Conv ($4 \times 4 \times 128$, Stride=2), BN, Leaky ReLU	Transpose Conv ($2 \times 2 \times 1024$, Stride=1), BN, LReLU
Conv ($4 \times 4 \times 256$, Stride=2), BN, Leaky ReLU	Transpose Conv ($4 \times 4 \times 1024$, Stride=2), BN, LReLU
Conv ($4 \times 4 \times 512$, Stride=2), BN, Leaky ReLU	Transpose Conv ($4 \times 4 \times 1024$, Stride=2), BN, LReLU
Conv ($4 \times 4 \times 1024$, Stride=2), BN, Leaky ReLU	Transpose Conv ($4 \times 4 \times 512$, Stride=2), BN, LReLU
Conv ($4 \times 4 \times 1024$, Stride=2), BN, Leaky ReLU	Transpose Conv ($4 \times 4 \times 256$, Stride=2), BN, LReLU
Conv ($2 \times 2 \times 2048$, Stride=1), BN, Leaky ReLU	Transpose Conv ($4 \times 4 \times 128$, Stride=2), BN, LReLU
	Transpose Conv ($4 \times 4 \times 3$, Stride=2), Tanh

Table 6.1: Encoder and decoder architectures. BN refers to batch normalization while LReLU refers to the leaky rectified linear units activation function.

We note also that while the use of shortcut connections [168] is a common strategy in encoder/decoder pairs to help retain finer details in the image by bypassing the encoding bottleneck, we do not employ this strategy. Since our intent is to obfuscate images through the addition of noise to the encoding, any information that bypasses this bottleneck leads to leakage of sensitive information regarding the identity depicted in the image. This would render the differential privacy guarantee meaningless. Although one could potentially apply noise to the information passed through the shortcut connections, this would present a number of additional challenges including the degradation of newly introduced correlations, obfuscation applied at a local level, and the requirement for a more complex privacy budget management scheme. We expect that this would not be a worthwhile venture but we leave the possibility open as a potential extension of our work.

6.4.2 Loss Calculations for Noisy Encodings

In addition to using an encoding amenable to the application of noise, the decoder network must also be trained on how to handle noisy encodings. Networks such as DCGAN [166] have demonstrated that appropriate manipulation of encodings can lead to novel, realistic images within the targeted data distribution. However, since our modification of the encoding is stochastic by design, this makes the task of producing realistic images more difficult. Furthermore, despite the compression of features to a global level, each encoding feature is independently perturbed, leaving the possibility for unrealistic output images due to resulting combinations of noisy features that fall outside of the data distribution.

Specifically, the decoder must be trained to accept noisy encodings and produce realistic faces depicted in the output images. The adversarial nature of the training process can be leveraged to achieve this. To this end, we modify the training process by injecting noise into the encodings of training samples before passing the encoding to the decoder. We modify the adversarial losses for the training process as follows:

$$\begin{aligned} L_{adv_d} = & -\mathbb{E}_{x^a, p \sim p_{data}} D(x^{a,p}) + \\ & \mathbb{E}_{x^a, p \sim p_{data}, b \sim p_{attr}, q \sim p_{pose}} D(\tilde{x}^{\hat{b}, \hat{q}}) + \\ & \lambda_{GP} \mathbb{E}_{x^a, p \sim p_{data}, b \sim p_{attr}, q \sim p_{pose}} GP(x^a, \tilde{x}^{\hat{b}, \hat{q}}), \end{aligned} \quad (6.21)$$

$$L_{adv_g} = -\mathbb{E}_{x^a, p \sim p_{data}, b \sim p_{attr}, q \sim p_{pose}} D(\tilde{x}^{\hat{b}, \hat{q}}). \quad (6.22)$$

The preservation of specific facial attributes in obfuscated images is also one of the goals for our model. To ensure that the injection of noise does not alter the attributes we wish to preserve, we modify the classification loss of the generator such that it is also calculated for training samples to which noise has been injected (note that ℓ_a remains as defined in Formula 6.4):

$$L_{cls_g} = \mathbb{E}_{x^a, p \sim p_{data}, b \sim p_{attr}, q \sim p_{pose}} \ell_a(\tilde{x}^{\hat{b}, \hat{q}}, b). \quad (6.23)$$

The loss for the classification network remains as it originally was with the exception that the pose parameters are now added to the notation:

$$L_{cls_c} = \mathbb{E}_{x^a, p \sim p_{data}} \ell_a(x^{a,p}, a). \quad (6.24)$$

The change from row b to row c in Figure 6.2 demonstrates the improvements in the visual quality and realism of the obfuscated faces using noisy samples during the training process.

6.4.3 Pose Preservation

Given that images in unconstrained settings capture individuals with a wide variety of poses, it is desirable to preserve this information in the obfuscated images. Unless care is taken in the design and training of the model to manage specific types of information, it will remain entangled in the encoding of the image. As a result, it is also subject to perturbation when noise is applied to the encoding. For example, if an image depicts an individual with their head turned to the side, the addition of noise to the encoding may alter the pose of the head, likely changing it to a forward-facing pose as this is typically the predominant orientation in training data. This produces an output that may look unnatural and destroys important contextual information in the image. To address this, we extend the model architecture and training process to accommodate for additional decoder inputs that capture pose information.

In order to extract pose information from images, we employ RingNet [173], a neural network which is able to regress 3D facial model parameters from 2D input (images). An important aspect of these parameters is a separation of identity-specific information, captured in shape parameters, from the non-identity-specific information of camera position, pose and facial expression. This separation allows us to target the non-sensitive information that is of interest to us. The RingNet model uses 3 parameters that capture axis-angle representations of the global rotation for the depicted head. Since we align all input images to our model (as later described in our training details, Section 6.8.2), we require only the yaw and pitch angles. We use RingNet to annotate these two parameters for both training images as well as the images to be obfuscated.

We modify the decoder architecture to concatenate these parameters along with the vector of facial attributes. We modify the discriminator by adding an extra branch (parallel to the branches of the discriminator/critic and the classifier) with two fully connected layers to approximate the parameters. Details on the architecture of these networks can be found in Tables 6.1 and 6.2.

Discriminator	Classifier	Pose Regressor
Input: 128×128×3 Image		
Conv (4×4×64, Stride=2), IN, LReLU		
Conv (4×4×128, Stride=2), IN, LReLU		
Conv (4×4×256, Stride=2), IN, LReLU		
Conv (4×4×512, Stride=2), IN, LReLU		
Conv (4×4×1024, Stride=2), IN, LReLU		
Fully Connected (1024), LReLU	Fully Connected (1024), LReLU	Fully Connected (1024), LReLU
Fully Connected (1)	Fully Connected (1)	Fully Connected (2)

Table 6.2: Discriminator, classifier and pose regressor architectures. IN refers to instance normalization while LReLU refers to the leaky rectified linear units activation function.

Let P denote the pose regression network embodied by the added branch and let $P_i(x)$ denote the output of P for the i^{th} parameter of an image x . In order to train the model on these new parameters, we introduce an additional loss function to minimize of the mean squared error in the approximated values:

$$\ell_p(x, p) = \sum_{i=1}^k \frac{(P_i(x) - p_i)^2}{k}, \quad (6.25)$$

$$L_{pose_r} = \mathbb{E}_{x^{a,p} \sim p_{data}} \ell_p(x^{a,p}, p), \quad (6.26)$$

$$L_{pose_g} = \mathbb{E}_{x^{a,p} \sim p_{data}, b \sim p_{attr}, q \sim p_{pose}} \ell_p(\tilde{x}^{\hat{b}, \hat{q}}, q). \quad (6.27)$$

The change from row c to row d in Figure 6.2 demonstrates the impact of pose preservation in the obfuscated images.

6.4.4 Objective Function

With the new loss calculations laid out, we are now able to put together the full objective function for the model training process. We leave the reconstruction loss essentially untouched, updating

only the notation to reflect the use of unmodified pose parameters as input for the decoder:

$$L_{rec} = \mathbb{E}_{x^{a,p} \sim p_{data}} \|x^{\hat{a},\hat{p}} - x^{a,p}\|_1. \quad (6.28)$$

While high fidelity reconstruction of images is not explicitly a goal in our setting, it is desirable for the model to learn that noiseless images should resemble the original image for the sake of preserving utility in features that are not captured in the other loss calculations. In this way, under a low-privacy setting, the injected noise will only make minor adjustments to the depicted identity, which may be desirable in many cases.

The full objective function is given below. The λ coefficients denote hyperparameters used to configure the relative importance of each objective.

$$\min_{G_{enc}, G_{dec}} \lambda_{rec} L_{rec} + \lambda_{cls_g} L_{cls_g} + L_{adv_g} + \lambda_{pose_g} L_{pose_g}, \quad (6.29)$$

$$\min_{D, C, P} \lambda_{cls_c} L_{cls_c} + L_{adv_d} + \lambda_{pose_r} L_{pose_r}. \quad (6.30)$$

6.4.5 Image Background

While the focus of our model is on the face depicted in an input image, it is inevitable that portions of the image will also capture the background behind the individual. The image could be tightly cropped to the face, however this is undesirable since it excludes major portions of the head, such as hair and ears, which would leak identifiable features. Although accurate representation of the background can be handled by many GANs, allowing the background to appear in the input image leads to an entanglement of background-specific features in the image encoding. The primary concern when this happens is the undesirable distortion of the background upon injection of noise, which negatively impacts visual quality. Additionally, the inclusion of background details in the image encoding wastes encoding capacity on features that are irrelevant to the goal of facial obfuscation. This, in turn, leads to a wasteful use of the privacy budget which is intended only to be spent on the facial features.

To address these issues, we use Mask R-CNN [94] to mask the area covered by the head in each image, allowing us to subtract all background content. By pre-processing all training images in this way, the model learns a distribution in which all images depict a head in a space with a white background. This also avoids the need for features to capture complex patterns in the background of the image. When an image is to be obfuscated, we similarly apply masking and background subtraction prior to passing the image as input to the trained model. When noise is added to the encoding, the obfuscated image is likely to remain in the learned distribution, depicting a new identity over a white background.

Once an image has been masked and obfuscated, the new head must be recombined with the original image background. During the pre-processing step, we store the image background so that it can be later recombined. Some care must be taken to do this in a clean way. A simple overlay of the obfuscated head onto the original background leaves the possibility for gaps to occur in the image where the new head no longer covers areas previously occupied by the original head. To fill in these gaps in a visually realistic manner, we apply Pluralistic Image Completion [212], a GAN-based approach to facial image in-painting.

Row e of Figure 6.2 shows obfuscated images using the background masking and merging process. Unlike the obfuscated images in all other rows, the background remains untouched; only the depicted identity is changed. The pipeline for the process of masking, obfuscation and recombination is illustrated in Figure 6.3.

Figure 6.3: Pipeline for the obfuscation process using masking. **Step 1:** the input image is masked to separate the head and the background. **Step 2:** Obfuscation is applied to the masked head. **Step 3:** The obfuscated head and masked background are recombined using an in-painting GAN.

6.5 Obfuscation Within a PCA Basis

While image encodings can be directly treated as the sensitive information to be obfuscated, better utility, can be achieved by instead performing the obfuscation within a PCA basis. Since we employ distance-generalized differential privacy, we must carefully examine the measure of distance used in the new basis to ensure that a meaningful level of privacy is achieved. In this section, we explain how to manage the obfuscation process within a PCA basis and how to efficiently allocate the privacy budget. We then discuss how to update the training process for our proposed GAN to accommodate these changes to the manipulation of image encodings such that we retain the ability to generate realistic obfuscated images.

6.5.1 Proposed Usage of PCA

As discussed in Section 6.1.3, the application of differential privacy within a transformed PCA basis offers a number of benefits related to the preservation of utility and the management of the privacy budget. To obtain these benefits, we use the image encodings of our training data to learn a PCA transformation that is well-suited to the data we aim to obfuscate. When presented with a new image that requires obfuscation, we then encode it, transform the encoding to the PCA basis, apply noise, and transform the noisy data back to its original basis. This provides an obfuscated encoding that has been restored to a representation in its original basis, allowing for it to then be passed through the decoder as usual to generate the output image. This process is illustrated in Figure 6.4.

Our usage of PCA with differential privacy differs from existing works on two counts. First, our training data is not private, therefore the task of learning the PCA transformation need not be done in a differentially private manner in our setting. Second, since we employ a distance-based generalization of differential privacy, we no longer deal with the concept of query sensitivity and cannot apply approaches such as [103] to determine the required magnitude of noise within the PCA basis.

In order to properly control the addition of noise to encodings represented in the PCA basis, we employ our obfuscation framework (Section 6.3.1) to set up a distance measure and a mechanism appropriate for the protection of secrets in the form of encodings of facial identities. Let $X = [x_1, \dots, x_n]$ be the representation of an image encoding in the PCA basis. We use the distance measure defined in Formula 6.11 to interpret distinguishability between the PCA basis encodings. Recall

Figure 6.4: Pipeline for the PCA training and image obfuscation process. The image obfuscation process can be seen as an expansion of Step 2 in Figure 6.3.

from Section 6.3.1 that an encoding can be made differentially private by independently adding Laplace noise to each element of the encoding vector such that the element at the i^{th} position uses a distribution with a scaling parameter of $\sigma_i = \frac{n(max_i - min_i)}{\epsilon}$, where $(max_i - min_i)$ is a scaling factor used to ensure that the range of element is $[0, 1]$. ^{ϵ} The element-wise range scaling prevents features with large ranges from dominating the distance measurement, ensuring each element receives noise configured to the appropriate magnitude. Furthermore, the use of an L_1 distance measure is well-suited to representations in a PCA basis given that the principal components are linearly uncorrelated. We provide practical details for the interpretation of the resultant privacy guarantee in Section 6.6.1.

6.5.2 Privacy Budget Allocation

Beyond determining an appropriate distance measure, allocation of the privacy budget should also be considered. In a setting where each element of the secret carries roughly equal information, it is logical to evenly distribute the privacy budget across the elements. However, in the case of an encoding represented in a PCA basis, the information content of the elements decreases monotonically, and often quite significantly, from the first element to the last. We can take advantage of this property by selectively dropping trailing principal components in order to focus the privacy budget on the remaining components with higher information content.

Dropping components has the effect of reducing the dimensionality of the PCA basis representation. When data is transformed from the PCA basis back to its original basis, dropped components are treated as if they had assumed their mean value from the training dataset. For low variance components, this can have the effect of introducing relatively little error in the data while attaining perfect privacy protection since the original value of an encoding within the component is entirely removed from the released data. This is particularly valuable for dealing with components that have a low variance in relation to their overall range since the range-based scaling used in the configuration of the privacy mechanism would likely lead to a scaling parameter much higher than the

standard deviation of the component. Given that such a high scaling parameter would induce heavy perturbation of the data, use of the mean value (calculated from a set of low variance points) is likely to induce a much smaller error on average. As a result, carefully choosing which components to drop can achieve lower expected error on these components while simultaneously providing them with perfect privacy. Further to this, the privacy budget that would have been spent on these components is freed up to instead be spent on components with greater information content.

To guide the process of determining which components to drop, we propose the use of a binary search. Given a user's privacy budget ϵ , the algorithm will search the range of $[1, n]$, where n is the number of features in the original encodings, to find the best choice for how many of the leading components to retain. The search is guided by a user-specified threshold on the acceptable ratio of each scaling parameter used by the mechanism to the standard deviation (within the training data) of the corresponding component. For example, the user may specify that the scaling parameter used for each retained element may be at most 1.5 times the standard deviation of its component. Since the dimensionality of the PCA basis appears in the numerator of the scaling parameters, a reduction in dimensionality leads to a reduction in the magnitude of required noise. If the ratio for any of the leading components exceeds the threshold, the privacy budget is spread too thin and more components must be dropped. If none of the ratios exceed the threshold, the budget can allow for a greater number of components to be retained. The search terminates when the boundary between these two cases is found. Recall that the principal components are ordered according to their explained variance in the training data. This approach therefore determines the selection of components that retains the maximum explained variance in the data subject to the constraint on the scaling parameter threshold. We provide the steps in Algorithm 1.

Algorithm 1: Budget Allocation

Input: Encoding size n , Privacy budget ϵ , Ratio threshold α , Component standard deviations $s_1, \dots, s_n$, Component ranges $(min_1, max_1), \dots, (min_n, max_n)$

```

 $c_{min} \leftarrow 1$ 
 $c_{max} \leftarrow n$ 
 $c_{mid} \leftarrow \lceil \frac{c_{min} + c_{max}}{2} \rceil$ 
while  $c_{mid} \neq c_{max}$  do
   $reduce \leftarrow \text{False}$ 
  for  $i \leftarrow 1$  to  $c_{mid}$  do
    if  $\frac{c_{mid}(max_i - min_i)}{\epsilon} \geq \alpha s_i$  then
       $reduce \leftarrow \text{True}$ 
      break
    end
  end
  if  $reduce$  then
     $c_{max} \leftarrow c_{mid}$ 
  else
     $c_{min} \leftarrow c_{mid}$ 
  end
   $c_{mid} \leftarrow \lceil \frac{c_{min} + c_{max}}{2} \rceil$ 
end
return  $c_{mid}$ 

```

6.5.3 Model Training

With the details for the mechanism configuration and budget allocation laid out, we must next ensure that the GAN is trained to handle encodings that have been accordingly perturbed. As per the loss functions given in Section 6.4, our training process requires the addition of noise to training samples in order to teach the decoder how to produce realistic images from noisy encodings. Yet the addition of noise requires the use of the PCA transformation matrix while the calculation of this matrix (refer back to Figure 6.4) requires a training set of encodings from the target model. In short, there exists a circular dependence between the GAN and the PCA transformation. Although one could first train a model without noise and use this to generate the PCA training encodings, the model later trained with noise will differ in its learned parameters, rendering the PCA training data inaccurate with respect to the target distribution of encodings.

To handle this, we propose to compute the PCA transformation matrix at the end of each epoch of training using a snapshot of the current encoder. The transformation matrix can then be used during the next epoch of training. During the first epoch of training, we simply forgo the addition of noise and during each subsequent epoch, we add noise using the most recent transformation matrix. In this way, the noise-adding process is periodically updated throughout the training period to reflect the changes in the encoder. This allows for the model to always have access to a reasonable approximation of how it would currently be used to add noise in practice.

Since the mechanism is intended to allow for a configurable level of privacy based on the user selected parameter ϵ , we train the model for a range of different privacy budgets. Each time training data is sampled, we draw a privacy budget uniformly at random from a range of typical budgets. This helps to train the decoder for the preservation of utility at different levels of privacy. We provide more specific details on the training configuration in Section 6.8.2.

6.6 Privacy Interpretation

Although differential privacy offers a formalized guarantee of privacy, the implications of the guarantee are not always directly obvious. In particular, the interpretation of the distance-generalized guarantee we employ, and thus the selection of an appropriate privacy parameter, may be difficult for most users. In this section, we first provide information on how our privacy guarantee relates to other applications of differential privacy, one of which provides a bound on the change in knowledge (regarding the sensitive information) of attackers. We use this to offer some intuition on the meaning of the privacy guarantee in relation to the choice of ϵ . We then discuss, in terms agnostic to any particular method of obfuscation, how sensitive information can leak from portions of images that are left unobfuscated (e.g., the background of the image) and how an attacker can exploit this.

6.6.1 Differential Privacy Guarantee Interpretation

The differential privacy guarantee in Formula 6.1 provides an upper bound on the allowable difference between the probabilities of different inputs to the mechanism producing the same obfuscated output. This allows us to make an assertion that similar inputs have similar probabilities of producing the same output and are thus difficult to distinguish between. Since this is strictly a property of the mechanism, it holds regardless of the attacker's background knowledge. This is a useful property since it allows for reasoning about the level of privacy in a manner that is largely agnostic to the attack model. However, this degree of abstraction prevents the calculation of an actual level of re-identification risk for an obfuscated release. As a result, the privacy guarantee may be difficult for most users to interpret, making the task of selecting a value for ϵ rather challenging.

To help interpret the privacy guarantee, we first highlight a strong parallel between the distance measure we use and the concept of query sensitivity in standard differential privacy. Recall that we configure the randomization mechanism using a scaling parameter of $\sigma_i = \frac{n(\max_i - \min_i)}{\epsilon}$ for the i^{th} element of the image representation. The numerator of the scaling parameter acts as a measure of sensitivity for each element of the image representation, capturing its potential for influence in the overall distance measure. This serves as a surrogate for the concept of query sensitivity used in standard differential privacy for databases. Contrast this scaling parameter against that of the standard Laplace mechanism for databases, $\sigma = \frac{\Delta F}{\epsilon}$, where ΔF is the query sensitivity. Both scaling parameters are linear functions of a measure of sensitivity that captures the maximum amount by which a value can change. They are both, therefore, similarly sensitive to a worst-case interpretation of how much information could be revealed in the released data.

An important distinction in the interpretation of the distance-based guarantee is that without any notion of adjacency between secrets, we are considering a worst-case over all possible image encodings. This is akin interpreting the standard differential privacy guarantee in terms of maximally distant databases as opposed to adjacent databases. As such, it is important to recognize that the values of ϵ applied to our mechanism will have an entirely different meaning than those commonly seen in differential privacy for databases. It is in fact possible to capture the standard differential privacy guarantee in the distance-generalized framework by 1), considering the domain of secrets to be the set of all possible database configurations, and 2) using the Hamming distance between pairs of databases as the distance measure [27]. Within this setting, the range of the distance measure is $[1, n]$, where n is the number of records in the database. The privacy guarantee for adjacent databases corresponds to pairs of secrets at distance 1, a small fraction of the total range of the distance measure. Although there is no precise correspondence between Hamming distance and our distance measure, it is clear that the guarantee would be most meaningfully interpreted in terms of a small fraction of the range of the distance measure.

Moving away from the concept of differential privacy for databases, we can also contrast our privacy guarantee against that of geo-indistinguishability [12]. Geo-indistinguishability is an application of distance-generalized differential privacy to protect location data, using a measure of physical distance to configure the privacy mechanism. A key intuition in geo-indistinguishability is that location data is desirable to reveal at an approximate level in order to allow for utility in the released information. Only pairs of locations that are nearby should remain highly indistinguishable. The same notion applies to our setting of facial obfuscation in the sense that distinguishability between “distant” pairs of facial identities is acceptable and desirable for the sake of utility. To help interpret the practical implication of this, we highlight a bound used in geo-indistinguishability that reveals a limitation on the greatest possible change in the ratio of probabilities for a pair of candidate locations. We present the bound here, adapted to our context of facial identity in image encodings. Let X be a random variable that represents the unknown (to the attacker) image encoding to be protected. Let $\Pr(X')$ be the probability with which an attacker believes $X = X'$ prior to seeing the obfuscated release and let $\Pr(X'|R)$ be the probability with which they believe $X = X'$ after observing an obfuscated release R . The distance-based privacy guarantee enforces the following condition:

$$\frac{\Pr(X_1|R)}{\Pr(X_2|R)} \leq e^{\epsilon d(X_1, X_2)} \frac{\Pr(X_1)}{\Pr(X_2)}, \quad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (6.31)$$

The usage of this bound is necessarily dependent on assumptions regarding an attack model due to the use of probabilities representing the attacker’s prior beliefs about the likelihood of candidate identities. However, we provide an example of a possible scenario. Consider an attacker with no prior knowledge regarding the candidate identities, thus all prior probability ratios are equal to

1. Since the guarantee is intended to offer strong protection for similar secrets while allowing for some distinguishability between distant secrets, we may be interested in representations of image encodings that are roughly at a distance of 0.01 from each other (recall that the total range of the distance measure is $[0,1]$). From our experiments (later shown in Section 6.8), we have found empirically that $\epsilon < 300$ offers a strong level of privacy. If we want a formalized interpretation of these settings, we can apply the bound of Formula 6.31 using a budget such as $\epsilon = 200$ to find that the ratio of probabilities representing an attacker’s belief after observation of the obfuscated information is upper bounded by $e^{200 \cdot 0.01} = e^2$. Similarly, the multiplicative bound obtained for the guarantee in Formula 6.1, which is agnostic to the assumption made about the attack model, would also be e^2 . This is comparable to the implication of the privacy guarantee in standard differential privacy when ϵ is set to 2, a common choice in the context of databases.

6.6.2 Contextual Information in Images

Beyond information that is discerned from the obfuscated data, other information will typically be revealed when an image depicting an obfuscated face is released. It is important to be aware that the privacy guarantee we provide applies strictly to the obfuscated portion of the image. Since the face is one of the most identifiable parts of a person, this is often the most crucial part of an image to be obfuscated. However, when obfuscation is focused only on the face or head in an image, a great deal of other information is left untouched and can be exploited by an attacker.

The remainder of the body is a prime source of such information. Details such as skin tone, tattoos, body size and shape, clothing, accessories and more can all be used as queues to help identify an individual. Even if the head is fully redacted, these pieces of information alone may be enough to achieve a high degree of re-identification accuracy [159]. The background of the image can also give away many revealing queues. Information such as natural or man-made landmarks and signs may help to determine the location of the image. Clocks, the position of the sun or even the weather may give away the time at which the image was taken. Furthermore, some images may be released with metadata that directly provides some of these pieces of information. An attacker possessing background knowledge about cities in which individuals live, or worse yet rough ideas of individuals’ itineraries, may use this information to narrow down the candidates for an obfuscated identity.

When an obfuscated image is to be released, the user or data custodian should carefully consider all potential sources of information that an attacker can exploit. Facial obfuscation is an essential component for a system of privacy protection but may not always be sufficient by itself. Our work provides a method of achieving a provably strong level of privacy with respect to the facial information released but we do not necessarily advocate that this is a comprehensive solution to privacy protection in sensitive images. Depending on the other information present in the image, facial obfuscation may be used as one component in a larger system. For example, a method of obfuscation for bodies [24] could be applied in conjunction with our facial obfuscation.

6.6.3 Information Leakage via Auxiliary Model Inputs

In some generative models of facial obfuscation, such as ours, auxiliary inputs beyond the facial identity are provided to the model for the purpose of utility. These are inputs such as gender and facial expression which are considered useful pieces of information that are intentionally conveyed in the released image. Although some information about the identity is revealed via personal attributes such as gender, this is considered acceptable and the inputs are otherwise seen as being benign. However, there exists potential for unintended leakage of information if the auxiliary inputs are

generated from other machine learning models. Recent work [183] has shown that representations learned by a neural network model can be re-purposed to perform unintended tasks that reveal sensitive information about the input data. For example, a binary classifier for gender in facial images will learn high level facial features that are used to perform the gender classification task. The representation of these features in the penultimate layer of the network may also reveal information about other facial attributes such as ethnicity.

The attacks studied in [183] rely on access to an intermediate representation of the model input prior to its final classification output (i.e., hidden layer outputs). Such intermediate representations carry significantly more information about the input image than the final output of the model, which is highly specialized for a particular task. In the context of facial obfuscation, it is strictly the outputs of the final layer that an attacker could access. Yet, since a typical model enforces no formalized guarantee on the absence of information leakage in its outputs, it stands to reason that some minor degree of sensitive information may yet be present. This may occur due to imperfect disentanglement of features by the model or known bias in the model outputs. We do not study this aspect of information leakage but leave it as an open problem as to how much of a concern it may be for facial obfuscation.

We emphasize that this is not something unique to our application of differential privacy for facial obfuscation. Any method of obfuscation that relies on the use of automated generation of auxiliary inputs for utility-related objectives has the same potential for information leakage. Although we do not investigate these, we propose three different approaches to address this concern, should it be deemed necessary in practice:

- **Require manual entry of auxiliary inputs wherever possible.** This avoids the potential for inferences based on knowledge about models that generate the inputs. A practical example is the use of a user profile that requires entry of desired facial attributes only once and then applies the attributes each time an image of the user is to be obfuscated.
- **Apply differential privacy to the auxiliary inputs.** The degree of perturbation can be configured based on the expected risk associated with the inputs. Our approach to generalized differential privacy in a PCA basis can be applied to continuous values such as the RingNet parameters. For categorical values such as facial attributes, the exponential mechanism [135] could be applied.
- **Train a model to only use the image encoding without auxiliary inputs.** This ensures that all sensitive information is subject to the formalized privacy guarantee. In the context of differential privacy, the degree of utility preserved then becomes entirely dependent on the configuration of the mechanism.

6.7 Differential Privacy in Pixel-Space

While we have thus far considered the application of differential privacy to the numeric representations of generative models, it can also be applied directly to the pixel intensities of an image. Such an approach suffers in the visual quality of the obfuscated images. Noise added in this way is no longer guided to obfuscate only specific aspects such as identity. Yet, by operating within the pixel-space, the randomization mechanism can be applied to any image, regardless of what is depicted. This versatility can be beneficial when obfuscation is required for images that are not readily captured by available models. For example, it may be desirable to obfuscate signs, license plates or complete vehicles. A pixel-space randomization mechanism provides a means to directly achieve privacy

protection on any image. Furthermore, since this approach obfuscates all of the visual information captured in the image, this addresses the concerns of contextual information raised in Section 6.6.2.

Differential privacy via a Laplace mechanism has previously been applied in combination with pixelization to achieve pixel-space obfuscation [65]. Using an adjacency relation for images based on a threshold of differing pixels, the privacy guarantee as used for databases was directly applied for the obfuscation of images. The Laplace mechanism was configured with respect to the maximum possible difference in pixelized grid intensities between adjacent images. We propose to instead apply the exponential mechanism (Definition 5) in order to achieve higher utility. We measure utility with respect to the reduction of a visual quality loss function in the obfuscated output. Since we wish to obfuscate a sequence of pixels, we represent both the input and the output of the randomization mechanism as vectors of pixel intensities. Let $\mathbb{P}$ be the set of valid pixel intensities (e.g., integers from 0 to 255) and $\mathbb{P}^n$ be the set of potential configurations a sequence of n pixels can take on. Formally, for any function $q : \mathbb{P}^n \times \mathbb{P}^n \rightarrow \mathbb{R}$ that measures the quality of a given obfuscated output for a given original instance, the probability distribution over all obfuscated instances of an original image $P \in \mathbb{P}^n$ is defined as:

$$\Pr(P') \propto e^{\epsilon q(P, P')}, \quad \forall P' \in \mathbb{P}^n. \quad (6.32)$$

In practice, we would not use the exponential mechanism to obfuscate a full image in a single application as this would require the definition of a prohibitively large probability distribution. Consider, for example, a 100×100 pixel image in a grayscale colour-space having 256 pixel intensities. This would give 256^{10000} possible obfuscated instances to be measured with the quality function. Instead, we propose to apply the mechanism to either small groups of pixels or individual pixels in an iterative manner until the full image has been obfuscated. For a quality function such as mean squared error (MSE), there is no downside to obfuscation at the individual pixel level since the overall measure of loss is simply an average over per-pixel measures. Since MSE is a loss function rather than a quality function, we take the negative MSE to obtain the quality function to use for the mechanism implementation.

Another natural choice for a quality measure is the structural similarity index measure (SSIM) [213]. This provides a human-centric measure of image quality which is designed to match up with how the human visual system perceives information. SSIM incorporates aspects of luminance, contrast and structure into a measure that is averaged over a sliding window intended to mimic how human eyes scan a large area but focus only on local areas. We hypothesize that the use of SSIM as the quality function could lead to improvements in visual quality in the obfuscated output over the MSE variant. However, computational complexity is a roadblock for the implementation of an SSIM variant. The measure is typically taken as an average over applications within an 11×11 pixel sliding window. A single application of the window for a colour-space with 256 pixel intensities requires a probability distribution over 256^{121} possible obfuscated instances of the window. We have explored reductions of the obfuscated instances through the application of pixelization to the image, coarsening of pixel intensity granularities, and modification of the window size. However, the reductions we explored that ran efficiently did not offer superior image quality to the MSE variant. We leave the development of an efficient SSIM variant as an open problem.

6.8 Experimental Comparisons

In this section, we present the results of experimental evaluations of our proposed methods of obfuscation and comparisons against other alternatives which also offer a formal guarantee of privacy in order to gain insight on the performance of our methods in practice. We begin with our proposed

implementations of differentially private obfuscation using an AAM and a GNN. We have implemented both differential privacy as well as k -same obfuscation using the same generative models and compare their performance with respect to privacy and utility over a range of different experiments. Following this, we turn to a comparison of GAN-based models in which we employ our proposed GAN coupled with our approach for differentially private obfuscation within a PCA basis. We compare this model against GAN-based implementations of k -same obfuscation and in-painting facial redaction. Finally, we compare our implementations of the more general pixel-space exponential mechanisms to the Laplace mechanism [65].

6.8.1 Comparisons Using AAM and GGN-Based Methods of Obfuscation

We begin with a comparison of our proposed AAM and GGN implementations of Section 6.3.3 to k -same implementations following the designs of k -same-m [82] and k -same-net [136]. We employ these comparisons to observe the relative performances of differential privacy and k -same obfuscation in terms of a trade-off between re-identification risk and utility. We additionally test the resilience of differential privacy against parrot attacks and composition attacks.

Generative Model Training and Datasets

We have implemented AAM obfuscation using AAM-API [185] for model training and generation of output. For GNN obfuscation, we have built on top of the DeconvFaces [70] network which implements the concept of up-convolution for the generation of images of input classes [47]. For both models, we have applied differential privacy as described in Sections 6.3.1 and 6.3.3. For the k -same implementation in the AAM, we have followed the process of k -same-m [82]. For the GNN, we have followed the approach of k -same-net [136]. In both cases, we have implemented clustering as described for k -same-m. This deviates from the use of a proxy gallery as described for k -same-net. It is important to note that, while a proxy gallery can reduce re-identification risk, it involves a step that is not captured by the k -same privacy guarantee. Thus, in the absence of a privacy guarantee that incorporates this detail, we omit the use of a proxy gallery in order to focus our experiments on the formalized aspects of the privacy guarantees. Similarly, although we have shown in Section 6.3.4 how a GNN can be used to apply differential privacy to unseen identities, we use only identities from the training set in our experiments. Deviations from an original identity induced by the approximation process may lead to a decrease in re-identification risk in a manner that is not captured by the formal privacy guarantee. We have therefore made these choices in the interest of comparing strictly the privacy protection achieved due to the formalized aspects of the methods of obfuscation.

We apply each method of obfuscation to two different datasets - RAFD [112] and KDEF [129]. These datasets provide frontal facial images of subjects wearing same coloured shirts. The use of same coloured shirts prevents bias in re-identification from the exploitation of information in unique clothing. The RAFD and KDEF datasets contain images of 67 and 70 subjects, respectively, and provide a variety of facial expressions. Due to apparent issues with lens exposure in the KDEF dataset, we have removed two of the subjects from our experiments.

The GNN architecture accepts class vectors for identity and facial expression as input. The RAFD and KDEF datasets are therefore highly suitable for this network. We have trained the network for 1000 epochs on each of the datasets to obtain models capable of reproducing these identities. An example of obfuscated output is shown in Figure 6.5. The AAM has the advantage of being able to approximate previously unseen identities. To use this in our experiments, we have trained a model for each dataset using the other dataset (e.g., RAFD as training data for KDEF)

with pre-processing to adjust the colour saturation of the training data in order to better match the target data. Since the training requires annotations of facial landmarks, we have employed OpenFace [11] to compute high accuracy approximations of the landmarks.

Although we do not employ the GNN approximation of unseen identities in our experiments, we include an example of some approximations in Figure 6.6 for reference. These approximations were produced by setting aside 10 identities from the RAFD dataset and training the GNN on the remaining 57. We then ran a linear program for each of the 10 omitted identities to approximate them as a weighted sum of the 57 training identities. As the training set is relatively small, we expect that the quality of the approximation can be improved through the use of larger and more diverse training sets. We also note that any minor deviations in the approximated identity are not of great significance since the depicted identity is ultimately to be obfuscated.

Figure 6.5: Obfuscation via the GNN on the RAFD dataset. The top row employs differential privacy and the bottom row employs k -same obfuscation.

Figure 6.6: Approximation of unseen identities. Original images are shown in the top row with the corresponding approximations in the bottom row.

Re-identification Risk

To measure re-identification risk, we have employed VGGFace D [161], a deep convolutional neural network which has been shown to achieve excellent facial identity classification accuracy. This simulates how an attacker might leverage machine learning models to launch an attack on obfuscated images. We have trained a separate model for each dataset, using the neutral and sad expressions for each identity for validation and the remaining expressions for training. Following a later released

Figure 6.7: Identity classification accuracy for the methods of obfuscation

note about the network training [180], we employ Xavier initialization [78] for the layer weights. To improve the robustness of the models, we have also augmented the datasets by creating two additional versions of each image - one with increased contrast and one with decreased contrast.

In our experiments, we generate obfuscated images having a neutral facial expression. We measure re-identification risk based on the accuracy of the top 1 guesses of the VGGFace network. Given that differential privacy is a stochastic process, for each combination of a privacy parameter and an identity to be protected, we have generated 10 obfuscated instances over which we take the average of the re-identification risk. We measure overall re-identification risk for a given privacy parameter as the average risk over all individuals in the dataset. Since the k -same approaches are deterministic, we produce only a single output image per identity and then take the average re-identification risk over the whole dataset. We additionally measure the baseline identity classification accuracy on the original (i.e., unobfuscated) data for reference. The results are shown in Figure 6.7.

Under differential privacy, a lower value of ϵ implies a stronger level of privacy whereas with k -same obfuscation, a higher value of k implies stronger privacy. The plotted data in Figure 6.7

shows the expected trend of reduced re-identification risk as the level of privacy, determined by the respective privacy parameters, is strengthened. In contrast to the typical ϵ values applied to differentially private mechanisms for databases, the values used in our experiments may appear unusually high. The larger magnitude is simply a side-effect of the normalization for the model vector, resulting in the interpretation of ϵ on a different scale. Refer back to Section 6.6.1 for discussion on the interpretation of ϵ in this context.

It is immediately notable that the AAMs (the red plots in Figure 6.7) are ineffective at privacy protection, even under severe privacy settings (i.e., low values of ϵ and high values of k). This is due to the contextual information outside of the facial contour such as the hair, ears and neck. In an alternative setting, the obfuscated face could be released on its own (e.g., against a black background) to prevent leaking this contextual information. While this would greatly improve the privacy protection, it would omit a great deal of useful information in the image and would lead to an output that is not very visually pleasing. We expect that in most practical situations, it is desirable to release a full version of the image. We therefore recommend against the use of AAMs for obfuscation, at least in a form that does not obfuscate the full head.

The k -same AAM results clearly demonstrate the violation of the k -same privacy guarantee since the re-identification risk is well above the theoretical maximum of $\frac{1}{k}$. We again note that if we only consider information within the contour of the face, the guarantee is not violated, however, such a guarantee is not useful in most practical situations. With differential privacy, we maintain a meaningful privacy guarantee for images that contain such information.

Preservation of Utility

To compare the methods of obfuscation in terms of utility, we focus on the ability to extract useful, non-sensitive information from the obfuscated output. Specifically, we measure classification accuracy for gender and facial expression in the obfuscated images.

We begin with the popular task of gender recognition [147]. As forms of demographic classification may be desirable for data mining purposes, we consider high classification accuracy to reflect good utility. To this end, we employ a pre-trained model of convolutional neural network for the classification of gender in facial images [115]. Our intent is to compare differential privacy and k -same obfuscation in terms of a privacy-utility trade-off. Yet, the two methods of obfuscation use proprietary privacy parameters which cannot be directly compared as a measure of privacy. We therefore plot the gender classification accuracy as a function of identity classification error in order to abstract away from the proprietary privacy parameters.

Given the poor obfuscation achieved by the AAMs, we omit them from this comparison. To highlight the ability of GNNs to incorporate properties relevant to image utility into the network architecture, we have also created a modified version of the architecture that preserves gender in the obfuscated output. To do so, we have created an input layer having two classes that specify the gender in the image. By training a model with gender labels, it learns to separate features relevant to gender from those relevant to identity. This enables us to focus obfuscation only on the features relevant to identity while leaving the gender feature vector untouched. An example of gender-preserving obfuscation is shown in Figure 6.8. In the interest of a fair comparison between the methods of obfuscation, we employ the modified architecture both for differential privacy as well as k -same obfuscation.

The results of the gender classification comparisons are shown in Figure 6.9. From these results, we see that the basic models for differential privacy and k -same obfuscation (shown by the blue and red plots, respectively) suffer a degradation in classification accuracy as the level of privacy is strengthened (i.e., as identity classification error increases). Comparing the gender-preserved models

Figure 6.8: Gender-preserving obfuscation via the GNN on the RAFD dataset. The top row employs differential privacy and the bottom row employs k -same obfuscation.

Figure 6.9: Gender classification accuracy for the methods of obfuscation

to their basic counterparts (shown by the yellow and purple plots), we see a large improvement in the classification accuracy, suggesting that this is an effective approach for the preservation of specific properties in the obfuscated output. In some cases, the classification accuracy of the obfuscated images has surpassed that of the original data. This is a result of the explicit specification of gender labels in the network input which can lead to obfuscated identities that more prominently display these features.

Another common task is the detection of facial expressions in images [192]. To compare utility in this context, we measure the classification accuracy for facial expressions by using pre-trained neural network models [6] intended for this task. For these experiments, we perform classification for six expressions: happiness, sadness, surprise, disgust, fear and a neutral expression. As the pre-trained models have different strengths and weaknesses with respect to their abilities to classify different expressions, we have employed them in combination to achieve high levels of classification accuracy on our data. We have found that the best results are achieved by first using the Model-4 [6] architecture trained on the RAFDB [118] dataset to detect disgust, followed by the Model-4 architecture trained on the SFEW [42] dataset to detect fear. If neither of these expressions were detected, we then take the sum of the two vectors of model predictions and select the highest prediction as the detected expression. As the models are not trained to detect contempt and give very poor accuracies for detection of anger in our datasets, we exclude these two expressions from our experiments.

For each expression, we generate a full set of obfuscated images as described in the experiments for re-identification risk (Section 6.8.1), providing the expression label as input to the network in order to apply the chosen expression to the obfuscated output. An example of expression-preserving output is shown in Figure 6.10. We measure both identity classification accuracy and expression classification accuracy as the average over all six expressions for each privacy parameter used in both differential privacy and k -same obfuscation. We then plot the expression classification accuracy as a function of the identity classification accuracy in order to once again abstract away from the privacy parameters and compare the utility of the two methods of obfuscation. The results are shown in Figure 6.11. As with the experiments on gender classification accuracy, when the chosen attributes are expressly preserved by the network, we observe classification accuracies that exceed the baseline accuracies measured on the unobfuscated datasets. This is again likely due to the network displaying the relevant features more prominently than the original images, making the task of the classification network easier. In these experiments, differential privacy (shown by the blue plot) generally shows better utility than k -same obfuscation (shown by the red plot).

Across both utility experiments, the overall comparison between the information preserved in differential privacy and k -same obfuscation appears to be inconclusive in these results. Some experiments show better results for differential privacy while other experiments show better results for k -same obfuscation. The k -same results are also more difficult to assess given the sporadic nature of the plots. This is likely due to changes in clusters between each level of obfuscation which can greatly impact classification accuracy. It is clear that the utility is also data-dependent given the variations in the results seen on the two datasets. Notably, many subjects in the KDEF dataset, including some males, have long hair whereas all subjects in the RAFD dataset have short hair. The males with long hair in KDEF may have contributed to the lower gender classification accuracy.

Parrot Attacks

Methods of k -same obfuscation are resistant to parrot attacks as a direct consequence of the process of obfuscation. Differential privacy, due to its resistance to breaches via post-processing of obfuscated data, similarly possesses a theoretical resistance to parrot attacks. Yet, the implications

Figure 6.10: Expression-preserving obfuscation via the GNN on the RAFD dataset. The top row employs differential privacy and the bottom row employs k -same obfuscation.

Figure 6.11: Facial expression classification accuracy for the methods of obfuscation

of parrot attacks in practical terms are less clear. A standard classification network applied to images obfuscated via differential privacy is unlikely to perform as well as it could if it exploited public knowledge about how randomization mechanisms function. An attacker could instead train a classification network for facial recognition using instances of (differentially private) obfuscated images as the training set. In this way, the network might achieve higher classification accuracy than a network trained on unobfuscated instances as it has learned to better identify features in the presence of noise. Yet, differential privacy is a stochastic method of obfuscation, so beyond the presence of noise and its approximate magnitude, there is little the network can learn in terms of predictability of differentially private output. We therefore hypothesize that a sufficiently high degree of noise induced by a differentially private mechanism can render the practical implications of a parrot attack negligible.

To observe the degree to which re-identification risk is impacted on differentially private output by parrot attacks, we have trained the VGGFace network for classification of obfuscated instances at specific privacy parameter (ϵ) values. This requires training a separate model for each privacy parameter value on each dataset. We then compare the classification accuracy achieved by the parrot attacks (shown by the red plot in Figure 6.12) to the accuracy of the models trained on the unobfuscated instances (shown by the blue plot in Figure 6.12). At high values of ϵ , there is an increase in classification accuracy for the parrot attacks as the network has become better suited to ignoring small amounts of noise. However, as the value of ϵ decreases, the gap rapidly closes and the trend reverses (at roughly $\epsilon = 300$ for the datasets we have employed), with the parrot attack showing lower classification accuracy than the model trained on unobfuscated data. This is likely due to the higher magnitude of noise destroying many of the useful patterns that the network otherwise learns in the training data. As a result, we expect that for reasonable configurations of privacy parameters that would be used in practice, parrot attacks would provide little, if any, advantage to an attacker.

Figure 6.12: Identity classification accuracy for a parrot attack

Composition Attacks

One of the key advantages that differential privacy provides over k -anonymity is the property of secure composition, which ensures that the privacy guarantee is never violated in the scenario of

uncoordinated releases of sensitive data. To demonstrate the resilience of our proposed method of obfuscation against composition attacks, we simulate such an attack and measure the identity classification accuracy for both our implementation of differential privacy and k -same obfuscation.

We consider a scenario in which an image is uploaded to two different platforms, each of which apply facial obfuscation in an uncoordinated manner. Through observation of the non-obfuscated portions of the image (e.g., the background of the image), an attacker could determine that the two obfuscated images originally depicted the same individual, enabling them to perform a composition attack. To simulate this, we split a dataset into two subsets of equal size. We select the subsets such that they have a non-empty intersection but also contain many identities that are not found in the other subset. We then train a generative network model on each of the subsets of individuals. This provides us with two models which have some identities in common but which will obfuscate those identities in different ways due to having been trained on differing subsets. These models represent two organizations which will perform obfuscation in an uncoordinated manner. We then use the models to obfuscate only the identities that are present in both subsets. The two obfuscated images produced for each identity represent the two images that an attacker would examine in a composition attack.

As the RAFD and KDEP datasets are too small to achieve any reasonable diversity of identities once they are further reduced to subsets, we combine both datasets together and draw the subsets from the combined dataset. To mitigate bias in identity classification from differences in the controlled settings of the two original datasets, we first crop all images to capture only the faces and then adjust the saturation and contrast of the images to match more closely. From the combined dataset of 135 individuals, we select 50 individuals to be shared across both subsets and split the remaining individuals evenly between them. We match the male to female ratios of both subsets to the full set of images but otherwise select the individuals for each subset at random.

To perform a composition attack on a pair of obfuscated images, we first provide both images as input to the VGGFace network and add together the two resultant vectors of prediction values for the full set of identities. For k -same obfuscation, we then take the intersection of the identities from each cluster used by the two models during obfuscation and select the identity within this intersection that has the highest prediction value from the summed vectors. In practice, an attacker will not necessarily know the exact identities used in each cluster with certainty but could likely determine them with reasonable accuracy by taking the top k predictions from their facial recognition network on each of the obfuscated images. By using the exact identities in our experiments, we effectively test the worst-case scenario.

When launching a composition attack against differentially private output, the attacker no longer has the concept of clusters of identities to use to their advantage. The original image may have depicted any of the potential identities. The attacker may still use the two images to increase the accuracy of their prediction but the additional information does not provide them with any means to violate the differential privacy guarantee. To simulate an attacker using the additional information in this context, we select the highest prediction value from the summed vectors of predictions for each pair of images. Since we generate ten instances of obfuscated output for each identity, we average the prediction accuracy over all ten pairs of images for each identity.

The results are shown in Figure 6.13, comparing the re-identification risk in each method of obfuscation for a single obfuscated image (shown by the blue plot) to the risk from a pair of images on which a composition attack has been performed (shown by the red plot). With the k -same obfuscation, we observe a significant increase in re-identification risk which greatly exceeds the theoretical maximum value of $\frac{1}{k}$. This clearly demonstrates a violation of the k -same privacy guarantee. With differential privacy, we also observe an increase in re-identification risk, as would be expected due to the additional information provided to the attacker, however, the property of secure composition

ensures that the privacy guarantee is preserved. Furthermore, the gap between the single image re-identification risk and composition risk is less significant on the differentially private images than the k -same images, and becomes marginal at high levels of privacy (e.g., when $\epsilon < 300$ in Figure 6.13). Since privacy parameters that are conducive to a high level of privacy would typically be selected in most realistic scenarios, we expect that the degradation in privacy due to multiple instances of differentially private output being distributed in practice would be minimal.

Figure 6.13: Identity classification accuracy for a composition attack

6.8.2 Comparisons Using GAN-Based Methods of Obfuscation

We now turn to a comparison of GAN-based methods of obfuscation. Here, we have implemented our proposed GAN of Section 6.4 combined with our approach to PCA-basis obfuscation of Section 6.5. Unlike the methods of obfuscation compared in Section 6.8.1, our GAN-based approach to facial obfuscation is able to handle a much more varied distribution of input images. In these experiments, we no longer restrict ourselves to highly constrained datasets, allowing instead for variation in aspects such as pose and image background content. We begin by providing the training details used for our model and the experimental setup. We then give an empirical evaluation of the level of privacy obtained through our model over a range of potential choices for the privacy budget ϵ . Lastly, we provide a comparative evaluation against GAN-based implementations of k -same obfuscation and in-painting facial redaction in terms of a trade-off between re-identification risk and measures of utility.

Training Details

We train our model using the CelebA [126] dataset, a collection of 202,599 images annotated with various facial attributes. We align the training images using HD CelebA Cropper [130] and generate pose parameters using RingNet [173]. We normalize the pose parameters such that the training annotations have zero mean and unit variance.

To prevent the preserved vector of facial attributes from leaking too much information, we choose only to keep the attributes for gender. Following AttGAN, we uniformly perturb the facial attributes to generate the modified vector $b = [b_1, \dots, b_m]$ given to the decoder during training. To generate

the perturbed vector of pose parameters $q = [q_1, \dots, q_k]$ we draw from a normal distribution using the original vector p as the location parameter and a scaling parameter of 1. We follow the AttGAN configuration of hyperparameters with the exception of λ_{cls_g} which we reduce to 1 due to the reduced number of attributes, and λ_{rec} which we increase to 200 for better preservation of visual quality when dealing with noisy encodings. We set our additional objective coefficients of λ_{pose_r} and λ_{pose_g} to 3 and 30, respectively. We have selected these values through experimental tuning to determine a configuration that achieves a reasonable trade-off between the various objectives.

When adding noise to the image encodings during training, we randomly draw a value of ϵ for each batch of images using a uniform distribution over the range [100,1000]. We use a threshold of 1.3 for the ratio of mechanism scaling parameters to standard deviation values. This exceeds the ratio of 0.9 that we use at the time of obfuscation in order to ensure that the model is trained to easily handle this magnitude of noise. If any elements of the noisy encodings fall outside of the range of values observed in the training data, we remap the out-of-bounds values to the boundary of the range. This is done to better preserve visual quality in the obfuscated images by restricting the encodings to remain closer to the distribution learned by the model.

Experimental Setup

We aim to investigate three measures of success in the images obfuscated by our approach: 1) privacy, 2) visual quality, and 3) utility. To measure privacy, we use FaceNet [174], a neural network that has reported excellent results in the classification of facial identities. To measure visual quality, we use SSIM [213], a measure of structural similarity between images which is inspired by principles of the human visual system. Lastly, to measure utility in terms of preservation of specific information, we employ a neural network trained for the task of gender classification. For this, we use the DeepFace [176] framework which implements the DEX [169] gender classification model.

We run our experiments on the FaceScrub [148] dataset, a collection of unconstrained setting images across 530 different identities. Since not all identities have the same number of images we take 50 images per individual to avoid bias due to over or under-representation of some individuals. This leaves us with 506 identities which meet the minimum image requirement. We split the data using 40 images per identity for training FaceNet and the remaining 10 for testing the obfuscation. For all experiments, we pre-process the data with an identical process to that used for our training data, using Dlib [106] to generate the facial landmarks that are needed for the alignment process. In order to obtain the facial attribute annotations needed by our model, we use our classification network to generate the annotations.

We compare our proposed method of obfuscation against two alternative approaches which also offer a formalized privacy guarantee: k -same obfuscation and in-painting. We have implemented k -same obfuscation within the encoding space of our trained model using the clustering algorithm as specified for k -same-m [82]. As such, our implementation of k -same obfuscation gains all the benefits of our proposed GAN architecture and serves as a comparison strictly between the formal privacy guarantees of differential privacy and k -same obfuscation. For obfuscation via in-painting, we use the publicly available pre-trained DeepPrivacy [102] model. A visual comparison of obfuscated output from each of these methods is shown in Figure 6.14.

During obfuscation, should any of the elements in the encodings of the test images fall outside of the bounds defined by the encodings of our training data, we remap the out-of-bounds elements of the testing encodings to boundary of the training data range. This step ensures that the encodings are suitable for the distance measure that we use. The configuration of the mechanism is based on the assumption that we know the range of each element in the encoding in order to appropriately scale the magnitude of the noise applied to the elements. However, in practice, we can only approximate

Figure 6.14: Examples of output images from the three methods of obfuscation applied in our experiments. The top row shows our proposed method of differentially private obfuscation using $\epsilon = 200$. The middle row shows the in-painting approach of DeepPrivacy. The bottom row shows k -same obfuscation using $k = 5$. Only one obfuscated image per identity is shown for k -same as it is a deterministic process.

the range using our training data. The adjustment of out-of-bounds values in the test data ensures that all encodings we encounter in practice conform to the assumed range. With a sufficiently large training set, occurrences of new encodings with out-of-bounds elements are unlikely to be common. Furthermore, any adjustments made in this way will have only a minor impact on the original encoding and thus a negligible impact on the obfuscated output given that noise is to be subsequently added regardless. After the application of noise to the encoding, we perform the same process of adjusting out-of-bounds values as described for the handling of noisy training encodings.

Re-Identification Risk

We begin with an evaluation of the impact the privacy budget ϵ has on re-identification risk. To calculate the level of re-identification risk, we pass obfuscated instances of images to a trained FaceNet model and take the average identity classification accuracy over all instances. Since differentially private obfuscation is stochastic, we generate 3 obfuscated instances per test image. We plot the accuracy as a function of ϵ to provide empirical results on the level of privacy attained at various choices of the privacy parameter.

We use FaceNet to execute a parrot attack on the obfuscated images, training a different model for each value of ϵ we test in order to exploit the ability of a classification network to learn patterns in methods of obfuscation. To do so we take the set of training set of images for FaceNet and pass them through the obfuscation process (configured with the targeted value of ϵ), again generating 3 instances per image. The obfuscated output is used to train the FaceNet model such that it learns to classify images subject to the targeted obfuscation as best as it is able.

The results are shown in Figure 6.15. The baseline classification accuracy achieved by FaceNet on the unobfuscated testing set is 92.6%. We do not show the other methods of obfuscation in this experiment since they cannot be plotted as a function of the privacy parameter used by our method. To provide a meaningful comparison against the other methods, the subsequent experiments examine the methods of obfuscation in terms of a privacy/utility trade-off.

Figure 6.15: Results on the impact of the privacy budget ϵ on re-identification risk. Identity classification accuracy is plotted as a function of ϵ . Recall from Section 6.6.1 that the multiplicative bound in the differential privacy guarantee is given by $e^{\epsilon d(X_1, X_2)}$ where the distance $d(X_1, X_2)$ between a pair of identities represented by X_1 and X_2 is most meaningfully interpreted in terms of nearby identities (e.g., at distance 0.01).

The plotted values in Figure 6.15 show a significant drop in classification accuracy from the baseline. This demonstrates the ability of our proposed method of obfuscation to provide a strong level of privacy, even when subjected to a parrot attack. Although the classification accuracy does not drop as low as that of random guessing, this is to be expected if the mechanism is to retain useful information in the obfuscated output. The intention with differential privacy is not to guarantee that no identifiable information is leaked but rather to limit the distinguishability of the released information. Therefore, some of the released information can naturally be exploited by the parrot-trained classification model. However, the degree of success the model achieves at classification is greatly hampered, as we have shown empirically.

Visual Quality

We next turn to the goal of preservation of visual quality in the obfuscated images. The measure of SSIM can be used to determine the similarity of an obfuscated image to its original instance. A measure of 1 indicates identical images whereas a measure of 0 indicates no structural similarity, therefore, higher values imply better visual quality. We plot the average SSIM over all obfuscated instances as a function of identity classification accuracy. This allows us to compare the methods of obfuscation in a manner that abstracts from approach-specific privacy parameters. The resulting plots can be interpreted as a representation of the trade-off between privacy and visual quality where high SSIM with low classification accuracy is desirable.

Each plotted data point is the result of applying a method of obfuscation with a particular privacy parameter configuration. For differential privacy, we apply a range of from 100 to 1000 in increments of 100 for the privacy budget. For k -same obfuscation, we apply values of 2, 3, 4, 5, 10, 20 and 50 k . Unlike differential privacy and k -same, DeepPrivacy has no configurable privacy parameter. Therefore, rather than producing a plot, it produces a single point of data. For each method of obfuscation, we calculate the identity classification accuracy when subjected to a parrot

Figure 6.16: A comparison of the trade-off between visual quality and re-identification risk. SSIM is plotted as a function of identity classification accuracy. There is only a single data point for DeepPrivacy since it has no configurable privacy parameter.

attack.

As the approach of DeepPrivacy is stochastic in nature, we generate 3 obfuscated instances per test image as with differential privacy. Since k -same is deterministic, we generate a single obfuscated instance per test image. Due to a requirement of k -same obfuscation for each identity to appear at most once in the set of images to be obfuscated, we partition the k -same testing set into 10 subsets, each of which contains a single image of each identity. These galleries are then further subdivided based on equivalence classes for the gender attribute. This is necessary to ensure that clusters do not contain images from different equivalence classes as this would directly violate the k -same guarantee given that the obfuscation GAN is designed to reveal these attributes.

The results are shown in Figure 6.16. Unsurprisingly, DeepPrivacy (shown by the red point) offers the highest level of SSIM in its obfuscated output, but at the cost of having the highest re-identification risk, since it modifies only a tightly cropped area containing the face rather than the whole head. In contrast, obfuscation via our proposed GAN modifies the hair, forehead, ears and neck as well. Furthermore, unlike DeepPrivacy which aims to match the new face to the unmodified facial contour, our GAN is able to make adjustments in skin tone. This leads to lower similarity to the original images and thus a lower SSIM score. However, the greater degree of modification is critical in achieving a sufficient level of privacy in the obfuscated output. A classification network trained for a parrot attack is able to recognize that the outer contour of the face remains invariant under tightly-cropped facial redaction and will thus focus on the features in the contour when attempting to classify identities. This is witnessed by the significantly higher identity classification accuracy measured for DeepPrivacy compared to the other methods of obfuscation. We argue that the drop in visual similarity to the original images under the more thorough modification of our GAN is a necessary sacrifice in order to achieve a reasonable degree of privacy.

The methods of differential privacy (shown by the blue plot in Figure 6.16) and k -same obfuscation (shown by the orange plot in Figure 6.16) appear to follow a similar trend to each other in the trade-off between re-identification risk and visual quality, likely due in part to the use of the same GAN for the generation of the output images. However, differential privacy demonstrates a slightly better trade-off.

Figure 6.17: A comparison of the trade-off between utility and re-identification risk. Gender classification accuracy is plotted as a function of identity classification accuracy. There is only a single data point for DeepPrivacy since it has no configurable privacy parameter.

Utility

Beyond the generation of images with high visual quality, we also aim to preserve specific features in the obfuscated images. Here, we test for the ability to preserve the depicted gender. The FaceScrub dataset provides a gender label for each image and has an equal number of male and female identities. Using these labels as the ground truth, we measure the gender classification accuracy and plot this as a function of identity classification accuracy. Low identity classification accuracy with high gender classification accuracy is desirable. The results are shown in Figure 6.17. For reference, we include the baseline gender classification accuracy on the unobfuscated set of test images.

Differential privacy and k-same obfuscation show a very similar trend in the preservation of gender. Beneath 3% identity classification accuracy, there is a slight drop-off in the gender classification accuracy for differential privacy. This is due to the fact that as the level of privacy is increased, the degree of noise required for differentially private obfuscation also increases, leading to greater modifications in the depicted identity. However, the gender classification accuracy remains very close to baseline accuracy regardless. Both methods of obfuscation in fact provide accuracy above the baseline for most of the plotted points. This is due to the attribute preservation training objective of the GAN which aims to produce output images that clearly depict the specified attributes.

In contrast to the methods which use our GAN, DeepPrivacy obfuscation suffers from a significant drop in gender classification accuracy. This is due to the fact that it does not use auxiliary inputs to selectively preserve certain types of information as we do in our proposed model. While methods of in-painting could be extended with such inputs, this would likely further exacerbate the issue of high identity classification accuracy.

Discussion

As demonstrated through our experimental results, our proposed GAN is able to achieve a favourable trade-off between privacy and utility. Contrarily, the approach of redaction and in-painting used by DeepPrivacy falls short in the ability to sufficiently hamper re-identification risk. Without a configurable privacy parameter, methods of in-painting have no means to further reduce re-identification

risk aside from widening the area of redaction. This type of all-or-nothing approach does not lend itself well to balancing privacy with utility.

The comparison between differential privacy and k -same obfuscation demonstrated comparable levels of utility from the two methods of obfuscation. Empirically, the two methods appear to perform very similarly when using the same GAN. Yet, differential privacy achieves stronger protection than k -same obfuscation against composition attacks, as shown in Section 6.8.1 and is better suited for protection against inferences in more practical scenarios where an attacker is likely to exploit background knowledge. Given the stronger theoretical properties of differential privacy combined with comparable levels of utility to k -same obfuscation, we propose that our method of obfuscation offers a better overall privacy-utility trade-off.

6.8.3 Privacy-Utility Trade-offs in Pixel-Space

Lastly, we compare our proposed use of a pixel-space MSE exponential mechanism to an existing implementation [65] of a Laplace mechanism in pixel-space. The original implementation of the pixel-space Laplace mechanism used a neighborhood size parameter to assist users in choosing an appropriate value for ϵ . In our experiments, we set the neighborhood size to a fixed value. This has no impact on the performance of the mechanism other than modifying the range of appropriate ϵ values needed to achieve desired levels of privacy. As the work in [65] proposes the combination of pixelization with differential privacy, we incorporate pixelization into our experiments as well. We test both mechanisms under three different pixelization settings: 4×4 , 2×2 , and 1×1 (i.e., no pixelization). As it is known that blurring a pixelized image can improve the ability of humans to recognize the image content [90], we additionally test versions of the images that have been blurred as a post-processing step after obfuscation has been applied. Recall that post-processing cannot impact the differential privacy guarantee [56]. However, this does not preclude the possibility of an impact on both re-identification risk and utility.

Given the pronounced visual modifications induced to the images through the application of noise in pixel-space, we apply the parrot attack concept of training a neural network for each level of obfuscated output. A network trained only on clean instances of images would be unable to cope with high levels of pixelization and noise, and would thus indicate overly optimistic levels of privacy. To this end, we employ a network architecture (A.4 of [134]) which has been previously used in the context of defeating pixelization via parrot attacks and has also been used to test the pixel-space Laplace mechanism [65].

We begin by investigating the impact of blurring obfuscated images on re-identification risk. We plot re-identification risk, measured as the accuracy of the facial identification network, as a function of the privacy parameter ϵ . In Figure 6.18, we show plots for blurred and non-blurred obfuscated images generated from the exponential and Laplace mechanisms at various levels of pixelization. The blur operation seems to have little, if any, impact on the network accuracies. For each mechanism variant, both the blurred and original versions are within a few percent accuracy of each other with no clear pattern for one being consistently better across all variants and privacy parameters. The minor differences can likely be attributed to the stochastic natures of the network training process and obfuscation process combined with relatively small training datasets. We hypothesize that the neural network is able to sufficiently learn any patterns present in the data to achieve on its own the same benefits that blurring presents for humans.

We next study the utility of the obfuscated images, measured in terms of MSE and SSIM, as a function of ϵ . The results are shown in Figure 6.19 where we plot the utility measures for the mechanism variants, again for both blurred and non-blurred obfuscated images. The differences in utility between the blurred and non-blurred images are significant. In all instances, blurring the

Figure 6.18: Comparison of the re-identification risk for blurred vs non-blurred obfuscated images at three different pixelization settings.

images provides a large improvement in utility for both measures. These measures are calculated directly on the pixel intensities of the obfuscated images and therefore are able to benefit greatly from the effect of the blur.

Given the improvements in utility for the blurred images, we use only blurred images when comparing the privacy-utility trade-offs of the mechanisms. Given that the re-identification risk is impacted by other factors than the privacy parameter, such as the level of pixelization, we plot utility as a function of re-identification risk in order to directly compare the mechanism variants. The results are shown in Figure 6.20. Interestingly, the lower levels of pixelization show a much more favourable trade-off in the context of SSIM and a similar trend can be seen at high levels of privacy (e.g., above 95% identity classification error) for MSE. We expect that this is due to the neural network’s ability to counteract much of the effect of pixelization via the parrot attack. As the experiments of [65] only considered re-identification risk and utility in isolation from each other, this result was not apparent from their experiments. We therefore recommend that pixelization is not used in combination with differential privacy. Rather, the randomization mechanism should be directly applied to the image in its original resolution.

With respect to the SSIM plots of Figure 6.20, the exponential mechanism consistently achieves a large improvement in utility over the Laplace mechanism. The exponential mechanism also outperforms the Laplace mechanism in terms of the MSE plots of Figure 6.20 at high levels of privacy. While the Laplace mechanism provides a better trade-off for MSE at lower levels of identity classification error, the loss in privacy may not be desirable in practice. Based on these results, the exponential mechanism without pixelization provides the best overall trade-off. If one is strictly interested in reducing MSE at moderate levels of privacy, use of some pixelization with the exponential mechanism may be beneficial. However, given that SSIM is generally accepted as a better measure of visual quality, our recommendation is to apply the exponential mechanism without the use of any pixelization.

Figure 6.19: Comparison of the utility for blurred vs non-blurred obfuscated images at three different pixelization settings.

Figure 6.20: Comparison of utility as a function of re-identification risk for blurred obfuscated images at three different pixelization settings.

6.9 Future Work

As the field of machine learning, and more specifically the study of GANs, is advancing rapidly, we expect that greater preservation of visual quality and utility in the obfuscated images can be attained by applying our methods to cutting-edge architectures and training methodologies. Improved handling of extreme variation in pose could also be achieved through the augmentation of the model training data with samples that capture a more balanced distribution over such types of variation. Furthermore, the investigation of alternate privacy budget allocation schemes within the PCA basis may prove fruitful in achieving improved privacy-utility trade-offs.

Investigation into the variability of non-sensitive that is not explicitly preserved by the generative model employed for obfuscation may also be of interest. For instance, if perceived age is later deemed important for data mining or data analysis purposes but was not set out as a preservation goal at the time of obfuscation, one may wish to know by how much the depicted age is expected to vary. Similarly, the impact of bias in the training data on the preservation of such unspecified attributes would be valuable to know in some circumstances.

We have also raised a number of important privacy considerations in Section 6.6 which are worth investigating. By using our framework and adapting our proposed GAN architecture and training details to other models, one could explore the potential to obfuscate different types of information in images such as the full body of individuals. We have also provided discussion on the potential for leakage of information via the auxiliary inputs to generative models applied for obfuscation. An investigation into more precise theoretical and practical implications of this potential risk could provide valuable insight into the design of appropriate methods of obfuscation.

Lastly, further study of our proposed measure of distance for the generalized privacy guarantee may help to better develop the intuitive understanding of the practical implications of our proposed method of obfuscation. Investigation into the distribution of distance measures over a sample population may provide insight on how to interpret distances in practice, allowing for users to make more informed decisions regarding their selection of the privacy parameter ϵ .

6.10 Summary

In digital images depicting faces, privacy is often a concern. A method of facial obfuscation should offer a provable guarantee of privacy while preserving utility in the obfuscated images. We have identified shortcomings in existing methods of facial obfuscation and have proposed the application of differential privacy to generative models as an alternative. We have developed a framework that provides a meaningful privacy guarantee for such models and we have derived the configuration of a Laplace mechanism that can achieve this privacy guarantee. Our proposed approach provides a strong guarantee of privacy while allowing for the generation of photo-realistic obfuscated images which preserve desired types of information such as the original pose and gender depicted in the input image.

We have provided implementation details for the application of our proposed mechanism to a variety of generative models, focusing on the design of a GAN which is able to provide a high degree of utility in the obfuscation images. Further to this, we have shown how to apply PCA to control the application of noise, allowing the privacy budget to be spent in an efficient manner. To assist in the practical application of our work, we have provided details on the interpretation of the privacy guarantee that we achieve. To then broaden the applicability of our work, we additionally propose a more general pixel-space mechanism for the obfuscation of any image (i.e., not restricted to facial images).

We have provided a wide range of experimental comparisons between implementations of our proposed approaches and other methods of facial obfuscation which also offer formal guarantees of privacy. Through these comparisons, we have demonstrated the resilience of differential privacy against parrot and composition attacks. Furthermore, we have shown that our implementations achieve a strong level of privacy protection while preserving favourable levels of utility in the obfuscated images. The results of these comparisons confirm that we provide a privacy guarantee that addresses the theoretical deficiencies of the k -same guarantee while simultaneously achieving higher levels of utility than other formalized approaches to facial obfuscation.

Chapter 7

Conclusions

Data collection and dissemination are activities which have become indispensable given the numerous benefits derived from the processing and analysis of rich datasets. This trend spans across the domains of socio-economic data, health care data and even datasets of facial images, among many others. Yet, the acquisition of such benefits comes paired with the potential for infringement on personal privacy. Often, datasets contain sensitive information that can be associated with specific individuals. The disclosure of links between sensitive information and identities constitutes a breach in privacy which is evidently undesirable from the perspective of the individuals to whom the data pertains. Such breaches must be carefully studied and accordingly mitigated for a range of reasons spanning from the preservation of participant confidence in the management of data collection to compliance with respect to laws and regulations. Yet, the requirement to protect sensitive information is at odds with the intent to obtain useful datasets. These conflicting goals have driven a diverse field of study referred to as privacy-preserving data analysis.

The work in this thesis falls within the domain of differential privacy, a disclosure control method which provides a formal guarantee of privacy with respect to information released via a randomization mechanism. Such mechanisms can broadly be interpreted as stochastic functions which add controlled noise to dataset query responses in order to limit the distinguishability between potential configurations of the underlying sensitive data which was queried. Our work is split into two major topics. In the first, we study the design of range-adherent differentially private mechanisms. This work is motivated by the potential to improve the utility of noisy mechanism responses by ensuring they adhere to the range of the query which was posed. We develop two main types of mechanisms to this end: truncated normalized Laplace mechanisms and range-constrained linear programming mechanisms. We assert adherence to the differential privacy guarantee through the underlying design of the mechanisms and demonstrate improvements in utility over other range-adherent alternatives through experimental comparisons.

The second major topic of the thesis covers differentially private obfuscation of facial images via generative models. Our aim in this work is to provide a formal guarantee of privacy regarding facial identity while generating realistic obfuscated images which preserve desired visual information such as the depicted pose and gender in the input image. We propose a framework to apply distance-generalized differential privacy to the encodings of images within generative models. This enables us to achieve a differentially private guarantee while producing visually realistic images via the generative models. To preserve a high degree of utility in the obfuscated output, we propose design details for a generative adversarial network tailored to our setting and couple this with an application of principal component analysis used to efficiently control how the privacy budget of the

randomization mechanism is spent. Through experimental comparisons against other methods of facial obfuscation which also offer formal guarantees of privacy, we demonstrate that our proposed approach is able to provide more favourable privacy-utility trade-offs.

Bibliography

- [1] Face Recognition in Low Quality Images: A Survey. *CoRR*, abs/1805.11519, 2018.
- [2] Facebook. <https://about.fb.com/company-info/>, 2020. Accessed: January 19, 2020.
- [3] Facebook by the Numbers: Stats, Demographics & Fun Facts. <https://www.omnicoreagency.com/facebook-statistics/>, 2020. Accessed: January 19, 2020.
- [4] Facebook for Business. <https://www.facebook.com/business/marketing/instagram#>, 2020. Accessed: January 19, 2020.
- [5] Instagram by the Numbers: Stats, Demographics & Fun Facts. <https://www.omnicoreagency.com/instagram-statistics/>, 2020. Accessed: January 19, 2020.
- [6] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool. Covariance Pooling for Facial Expression Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 480–4807, 2018.
- [7] N. R. Adam and J. C. Worthmann. Security-Control Methods for Statistical Databases: A Comparative Study. *ACM Computing Surveys*, 21(4):515–556, 1989.
- [8] C. C. Aggarwal. On k-Anonymity and the Curse of Dimensionality. In *Proceedings of 31st International Conference on Very Large Data Bases*, pages 901–909, 2005.
- [9] L. A. Alexander and T. B. Jabine. Access to Social Security Microdata Files for Research and Statistical Purposes. *Soc Secur Bull*, 41:3–17, 1978.
- [10] M. M. Almasi, T. R. Siddiqui, N. Mohammed, and H. Hemmati. The Risk-Utility Tradeoff for Data Privacy Models. In *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5, 2016.
- [11] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. OpenFace: A General-Purpose Face Recognition Library with Mobile Applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.
- [12] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In *2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 901–914, 2013.
- [13] Fahad Anwar, Ilias Petrounias, Tim Morris, and Vassilis Kodigoianis. Mining Anomalous Events Against Frequent Sequences in Surveillance Videos from Commercial Environments. *Expert Syst. Appl.*, 39(4):4511–4531, 2012.
- [14] P. Arzberger, P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhler, and P. Wouters. An International Framework to Promote Access to Data. *Science*, 303(5665):1777–1778, 2004.

- [15] J. Awan and A. Slavković. Differentially Private Uniformly Most Powerful Tests for Binomial Data. In *Advances in Neural Information Processing Systems 31, Montréal, Canada*, pages 4208–4218, 2018.
- [16] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release. In *Proceedings of the 26th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Beijing, China*, pages 273–282, 2007.
- [17] E. Barnoviciu, V. Ghenescu, S. Carata, M. Ghenescu, R. Mihaescu, and M. Chindea. GDPR compliance in Video Surveillance and Video Processing Application. In *International Conference on Speech Technology and Human-Computer Dialogue*, pages 1–6, 2019.
- [18] A. Basu, T. Nakamura, S. Hidano, and S. Kiyomoto. k-Anonymity: Risks and the Reality. In *Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 983–989, 2015.
- [19] K. Benitez and B. Malin. Evaluating Re-Identification Risks With Respect to the HIPAA Privacy Rule. *J Am Med Inform Assoc*, 17:169–77, 2010.
- [20] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. Face Swapping: Automatically Replacing Faces in Photographs. *ACM Trans. Graph.*, 27(3):39:1–39:8, 2008.
- [21] A. Brand. Public Health Genomics and Personalized Healthcare: A Pipeline from Cell to Society. *Drug Metabol Drug Interact*, 27:121–123, 2012.
- [22] A. Brand and J. A. Lal. European Best Practice for Quality Assurance, Provision and Use of Genome-based Information and Technologies. *Drug Metabol Drug Interact*, 27:177–82, 2012.
- [23] H. Brenner and K. Nissim. Impossibility of Differentially Private Universally Optimal Mechanisms. *SIAM Journal on Computing*, 43(5):1513–1540, 2014.
- [24] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328, 2017.
- [25] Statistics Canada. Individuals File, 2011 National Household Survey (Public Use Microdata Files), 2014. <http://www5.statcan.gc.ca/olc-cel/olc.action?objId=99M0001X2011001&objType=46&lang=en&limit=0> Accessed: Mar, 2015.
- [26] A. Cavallaro. Privacy in Video Surveillance [In the Spotlight]. *IEEE Signal Processing Magazine*, 24(2):168–166, 2007.
- [27] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the Scope of Differential Privacy Using Metrics. In *Privacy Enhancing Technologies*, pages 82–102, 2013.
- [28] Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. Near-optimal Differentially Private Principal Components. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*, pages 998–1006, 2012.
- [29] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward Privacy in Public Databases. In *Proceedings of the 2nd Theory of Cryptography Conference*, pages 363–385, 2005.
- [30] C.-L. Chen, R. Pal, and L. Golubchik. Oblivious Mechanisms in Differential Privacy: Experiments, Conjectures, and Open Questions. In *Proceedings of IEEE Security and Privacy Workshops*, pages 41–48, 2016.

- [31] H. Chi and Y. H. Hu. Face De-Identification Using Facial Identity Preserving Features. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 586–590, 2015.
- [32] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.
- [33] C. Clifton and T. Tassa. On Syntactic Anonymity and Differential Privacy. In *Proceedings of Transactions on Data Privacy*, pages 88–93, 2013.
- [34] European Commission. Commercial Exploitation of Europe’s Public Sector Information. Technical report, European Commission, 2000. <https://ec.europa.eu/digital-single-market/en/news/commercial-exploitation-europes-public-sector-information-pira-study-full-report>.
- [35] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active Appearance Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.
- [36] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth. On the Lambert W Function. *Advances in Computational Mathematics*, 5:329–359, 1996.
- [37] G. Cormode, T. Kulkarni, and D. Srivastava. Constrained Private Mechanisms for Count Data. In *34th IEEE International Conference on Data Engineering, Paris, France*, pages 845–856, 2018.
- [38] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Empirical Privacy and Empirical Utility of Anonymized Data. In *Proceedings of the IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, pages 77–82, 2013.
- [39] William Lee Croft, Jörg-Rüdiger Sack, and Wei Shi. Differentially private obfuscation of facial images. In *Machine Learning and Knowledge Extraction*, pages 229–249, 2019.
- [40] T. Dalenius. Finding a Needle in a Haystack or Identifying Anonymous Census Record. *Journal of Official Statistics*, 2:329–336, 1986.
- [41] D. E. Denning, P. J. Denning, and M. D. Schwartz. The Tracker: A Threat to Statistical Database Security. *ACM Trans. Database Syst.*, 4:76–96, 1979.
- [42] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static Facial Expression Analysis in Tough Conditions: Data, Evaluation Protocol and Benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, 2011.
- [43] I. Dinur and K. Nissim. Revealing Information while Preserving Privacy. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, volume 22, pages 202–210, 2003.
- [44] J. Domingo-Ferrer and V. Torra. Disclosure Risk Assessment in Statistical Microdata Protection Via Advanced Record Linkage. *Statistics and Computing*, 13(4):343–354, 2003.
- [45] J. Domingo-Ferrer and V. Torra. A Critique of k-Anonymity and Some of Its Enhancements. In *Proceedings of the 3rd International Conference on Availability, Security, and Reliability*, pages 990–993, 2008.
- [46] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial Feature Learning. *CoRR*, abs/1605.09782, 2016.
- [47] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox. Learning to Generate Chairs, Tables and Cars with Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):692–705, 2017.

- [48] L. Du, M. Yi, E. Blasch, and H. Ling. GARP-Face: Balancing Privacy Protection and Utility Preservation in Face De-Identification. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014.
- [49] Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. <http://archive.ics.uci.edu/ml>.
- [50] Maeve Duggan. Photo and Video Sharing Grow Online. Technical report, Pew Research Center, 2013. <https://www.pewresearch.org/internet/2013/10/28/photo-and-video-sharing-grow-online/>.
- [51] G. Duncan and L. Stokes. Data Masking for Disclosure Limitation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1:83–92, 2009.
- [52] G. T. Duncan and R. W. Pearson. Enhancing Access to Data While Protecting Confidentiality: Prospects for the Future. *Statistical Science*, 6:219–232, 1991.
- [53] C. Dwork. Differential Privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II, Venice, Italy*, pages 1–12, 2006.
- [54] C. Dwork. Differential Privacy: A Survey of Results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, Xi'an, China*, volume 4978 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 1–19, 2008.
- [55] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, volume 3876, pages 265–284, 2006.
- [56] C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407, 2014.
- [57] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: Optimal Bounds for Privacy-Preserving Principal Component Analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, page 11–20, 2014.
- [58] E. ElSalamouny, K. Chatzikokolakis, and C. Palamidessi. Generalized Differential Privacy: Regions of Priors that Admit Robust Optimal Mechanisms. In *Proceedings of PrakashFest Conference*, volume 8464, pages 292–318, 2014.
- [59] K. E. Emam, F. K. Dankar, A. Neisa, and E. Jonker. Evaluating the Risk of Patient Re-identification from Adverse Drug Event Reports. *BMC Med Inform Decis Mak*, 13, 2013.
- [60] K. E. Emam, F. K. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk. Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records. *Can J Hosp Pharm*, 62, 2009.
- [61] Science Europe. How to Transform Big Data into Better Health. Technical report, 2014. https://www.scienceeurope.org/wp-content/uploads/2015/12/Workshop-Report_MED_Big_Data_web.pdf.
- [62] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting Privacy Breaches in Privacy Preserving Data Mining. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 211–222. ACM, 2003.
- [63] EVPSI. Public Sector Information in Europe. http://www.evpsi.org/evpsifiles/PSI_in_EU_OVERVIEW.pdf.
- [64] L. Fan. Practical image obfuscation with provable privacy. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 784–789, 2019.

- [65] Liyue Fan. Image Pixelization with Differential Privacy. In *Data and Applications Security and Privacy XXXII*, pages 148–162, 2018.
- [66] I. P. Fellegi and J. J. Phillips. Statistical Confidentiality: Some Theory and Applications to Data Dissemination. *Annals of Economic and Social Measurement*, 3:399–409, 1974.
- [67] Thomas S. Ferguson. Linear programming: A concise introduction. <https://www.math.ucla.edu/~tom/LP.pdf>, Accessed: September 9, 2019.
- [68] S.E. Fienberg, A. Rinaldo, and X. Yang. Differential Privacy and the Risk-Utility Tradeoff for Multi-Dimensional Contingency Tables. In *Proceedings of Privacy in Statistical Databases*, pages 187–199, 2010.
- [69] F. Fioretto and P. Van Hentenryck. Constrained-Based Differential Privacy: Releasing Optimal Power Flow Benchmarks Privately. In *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 215–231, 2018.
- [70] M. Flynn. Generating Faces with Deconvolution Networks, 2016. <https://github.com/zo7/deconvfaces>, Accessed: November 1, 2018.
- [71] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-Scale Privacy Protection in Google Street View. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2373–2380, 2009.
- [72] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.*, 42(4), 2010.
- [73] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition Attacks and Auxiliary Information in Data Privacy. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 265–273, 2008.
- [74] Q. Geng, W. Ding, R. Guo, and S. Kumar. Truncated Laplacian Mechanism for Approximate Differential Privacy. *CoRR*, 2018.
- [75] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath. The Staircase Mechanism in Differential Privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.
- [76] Q. Geng and P. Viswanath. The Optimal Mechanism in Differential Privacy. In *Proceedings of IEEE International Symposium on Information Theory*, pages 2371–2375, 2014.
- [77] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally Utility-Maximizing Privacy Mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.
- [78] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 249–256, 2010.
- [79] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.
- [80] Google. Google Maps. <https://www.google.be/maps>, Accessed: February 27, 2019.
- [81] R. Gross and A. Acquisti. Information Revelation and Privacy in Online Social Networks. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pages 71–80, 2005.
- [82] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06)*, pages 161–161, 2006.

- [83] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating Utility into Face De-identification. In *Privacy Enhancing Technologies*, pages 227–242, 2006.
- [84] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent Advances in Convolutional Neural Networks. *Pattern Recognition*, 7:354 – 377, 2018.
- [85] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, pages 5767–5777, 2017.
- [86] M. Gupte and M. Sundararajan. Universally Optimal Privacy Mechanisms for Minimax Agents. In *Proceedings of the 29th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Indianapolis, USA*, pages 135–145, 2010.
- [87] S. Haney, M. Graham, A. Machanavajjhala, M. Kutzbach, J.M. Abowd, and L. Villhuber. Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics. In *Proceedings of ACM Special Interest Group on Management of Data*, pages 1339–1354, 2017.
- [88] M. Hardt, K. Ligett, and F. McSherry. A Simple and Practical Algorithm for Differentially Private Data Release. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pages 2339–2347, 2012.
- [89] M. Hardt and G. N. Rothblum. A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis. In *Proceedings of the Annual IEEE Symposium on Foundations of Computer Science*, pages 61–70, 2010.
- [90] L.D. Harmon. The Recognition of Faces. *Scientific American*, 229(5):71–82, 1973.
- [91] L.D. Harmon and B. Julesz. Masking in Visual Recognition: Effects of Two-Dimensional Filtered Noise. *Science*, 180(4091):1194–1197, 1973.
- [92] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.
- [93] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the Accuracy of Differentially Private Histograms Through Consistency. *Proceedings of the Very Large Data Base Endowment*, 3:1021–1032, 2010.
- [94] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.
- [95] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.
- [96] Kashmir Hill. The Secretive Company That Might End Privacy as We Know It. <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>, 2020. Accessed: 2020-06-03.
- [97] Johannes Hiller, Michael Schuldes, and Lutz Eckstein. Recognition and Pseudonymization of Data Privacy Relevant Areas in Videos for Compliance with GDPR. In *IEEE Intelligent Transportation Systems Conference*, pages 2387–2393, 2019.
- [98] S. T. Holgate, A. Palotie, B. Prainsack, A. Brand, and H. Lehrach. Personalised Medicine for the European Citizen. Technical report, European Science Foundation, 2012. <http://archives.esf.org/coordinating-research/forward-looks/biomedical-sciences-med/current-forward-looks-in-biomedical-sciences/personalised-medicine-for-the-european-citizen.html>.

- [99] Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Pól Mac Aonghusa. The Bounded Laplace Mechanism in Differential Privacy. *Journal of Privacy and Confidentiality*, 10(1), 2019.
- [100] Xiaoxia Hu, Donghui Hu, Shuli Zheng, Wangwang Li, Fan Chen, Zhaopin Shu, and Lina Wang. How People Share Digital Images in Social Networks: A Questionnaire-Based Study of Privacy Decisions and Access Control. *Multimedia Tools and Applications*, 77(14):18163–18185, 2018.
- [101] Scott E. Hudson and Ian Smith. Techniques for Addressing Fundamental Privacy and Disruption Tradeoffs in Awareness Support Systems. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, pages 248–257, 1996.
- [102] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In *Advances in Visual Computing*, pages 565–578, 2019.
- [103] Xiaoqian Jiang, Zhanglong Ji, Shuang Wang, Noman Mohammed, Samuel Cheng, and Lucila Ohno-Machado. Differential-Private Data Publishing Through Component Analysis. *Trans. Data Privacy*, 6(1):19–34, 2013.
- [104] E. F. Judge. Enabling Access and Reuse of Public Sector Information in Canada: Crown Commons Licenses, Copyright, and Public Sector Information. In Michael Geist, editor, *From “Radical Extremism” to “Balanced Copyright”: Canadian Copyright and the Digital Agenda*, pages 598–642. Irwin Law Inc., 2010.
- [105] Sanjay Kairam, Joseph “Jofish” Kaye, John Alexis Guerra-Gomez, and David A. Shamma. Snap decisions? how users, content, and aesthetics interact to shape photo sharing behaviors. In *2016 CHI Conference on Human Factors in Computing Systems*, page 113–124, 2016.
- [106] Davis E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [107] P. Korshunov and T. Ebrahimi. Using Face Morphing to Protect Privacy. In *10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 208–213, 2013.
- [108] P. Korshunov and T. Ebrahimi. Using Warping for Privacy Protection in Video Surveillance. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pages 1–6, 2013.
- [109] S. Kotz, T. Kozubowski, and P. Krzysztof. *Classical Symmetric Laplace Distribution*, chapter 2. Birkhauser Boston, 2001.
- [110] J. Kröckel and F. Bodendorf. Customer Tracking and Tracing Data as a Basis for Service Innovations at the Point of Sale. In *2012 Annual SRII Global Conference*, pages 691–696, 2012.
- [111] J. Kulynych and D. Korn. The Effect of the New Federal Medical-Privacy Rule on Research. *N Engl J Med*, 346:201–204, 2002.
- [112] O. Langner, R. Dotsch, G. Bijlstra, D.H.J. Wigboldus, S.T. Hawk, and A van Knippenberg. Presentation and Validation of the Radboud Faces Database. *Cognition and Emotion*, 24(8):1377–1388, 2010.
- [113] J. Lee, Y. Wang, and D. Kifer. Maximum Likelihood Postprocessing for Differential Privacy Under Consistency Constraints. In *Proceedings of the 21st ACM International Conference on Knowledge Discovery and Data Mining, Sydney, Australia*, pages 635–644, 2015.
- [114] G. Letournel, A. Bugeau, V. . Ta, and J. . Domenger. Face De-Identification with Expressions Preservation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4366–4370, 2015.

- [115] Gil Levi and Tal Hassner. Age and Gender Classification Using Convolutional Neural Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015.
- [116] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear Counting Queries under Differential Privacy. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 123–134, 2010.
- [117] P. Li, L. Prieto, D. Mery, and P. J. Flynn. On Low-Resolution Face Recognition in the Wild: Comparisons and New Techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019.
- [118] Shan Li, Weihong Deng, and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 28, pages 2584–2593, 2017.
- [119] Tao Li and Lei Lin. AnonymousNet: Natural Face De-Identification With Measurable Privacy. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, page 0, 2019.
- [120] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine. Blur vs. Block: Investigating the Effectiveness of Privacy-Enhancing Obfuscation for Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1343–1351, 2017.
- [121] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. Effectiveness and Users’ Experience of Obfuscation As a Privacy-Enhancing Technology for Sharing Photos. *Proc. ACM Hum.-Comput. Interact.*, 1:1–24, 2017.
- [122] Yuezun Li and Siwei Lyu. De-Identification Without Losing Faces. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, page 83–88, 2019.
- [123] B.-R. Lin and D. Kifer. Information Measures in Statistical Privacy and Data Processing Applications. *ACM Transactions on Knowledge Discovery from Data*, 9(4), 2015.
- [124] F. Liu. Statistical Properties of Sanitized Results from Differentially Private Laplace Mechanism with Univariate Bounding Constraints. *Trans. Data Priv.*, 12(3):169–195, 2019.
- [125] X. Liu, N. Krahnstoeber, T. Yu, and P. Tu. What Are Customers Looking at? In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 405–410, 2007.
- [126] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.
- [127] G. Loomes and R. Sugden. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92(368):805–824, 1982.
- [128] W. Lowrance. Access to Collections of Data and Materials for Health Research. Technical report, Wellcome Trust, 2006. <https://wellcomelibrary.org/item/b20927691>.
- [129] D. Lundqvist, A. Flykt, and A. Öhman. The Karolinska Directed Emotional Faces – KDEF, 1998. ISBN 91-630-7164-9.
- [130] LynnHo. HD CelebA Cropper, 2018. <https://github.com/LynnHo/HD-CelebA-Cropper>, Accessed: February 1, 2020.
- [131] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-Diversity: Privacy Beyond k-Anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, pages 24–35, 2007.

- [132] E. Mackey, M. Elliot, and K. Purdam. The Regulation of the Personal Individual Data Use and Identity in the UK. *Policy Studies*, 25(4):267–281, 2004.
- [133] Kirsten Martin and Katie Shilton. Putting Mobile Application Privacy in Context: An Empirical Study of User Privacy Expectations for Mobile Devices. *The Information Society*, 32(3):200–216, 2016.
- [134] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating Image Obfuscation with Deep Learning. *CoRR*, 2016.
- [135] F. McSherry and K. Talwar. Mechanism Design via Differential Privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science, Providence, USA*, pages 94–103, 2007.
- [136] B. Meden, Z. Emersic, V. Struc, and P. Peer. k-Same-Net: Neural-Network-Based Face De-identification. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOB)*, pages 1–7, 2017.
- [137] L. Meng and Z. Sun. Face De-identification with Perfect Privacy Protection. In *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1234–1239, 2014.
- [138] L. Meng, Z. Sun, A. Ariyaeinia, and K. L. Bennett. Retaining Expressions on De-Identified Faces. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1252–1257, 2014.
- [139] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. *CoRR*, 2020.
- [140] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784, 2014.
- [141] Tom Mitchell. Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression. In *Machine Learning*, chapter 3, pages 1–17. McGraw Hill, 2017.
- [142] Saleh Mosaddegh, Loic Simon, and Frédéric Jurie. Photorealistic Face De-Identification by Aggregating Donors’ Face Components. In *Computer Vision – ACCV 2014*, pages 159–174, 2015.
- [143] National Research Council. *The Role of Scientific and Technical Data and Information in the Public Domain*. The National Academies Press, 2003.
- [144] R. B. Ness. Influence of the HIPAA Privacy Rule on Health Research. *JAMA*, 298:2164–2170, 2007.
- [145] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Comput.-Hum. Interact.*, 13(1):1–36, 2006.
- [146] E. M. Newton, L. Sweeney, and B. Malin. Preserving Privacy by De-Identifying Face Images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.
- [147] Choon Boon Ng, Yong Haur Tay, and Bok-Min Goi. Recognizing Human Gender in Computer Vision: A Survey. In *PRICAI 2012: Trends in Artificial Intelligence*, pages 335–346, 2012.
- [148] Hongwei Ng and Stefan Winkler. A Data-Driven Approach to Cleaning Large Face Datasets. In *IEEE International Conference on Image Processing*, pages 343–347, 2014.
- [149] H. H. Nguyen, J. Kim, and Y. Kim. Differential Privacy in Practice. *Journal of Computing Science and Engineering*, 7:177–186, 2013.

- [150] L. Ninghui, L. Tiancheng, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proceedings of the International Conference on Data Engineering*, pages 106–115, 2007.
- [151] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard G. Medioni. On Face Segmentation, Face Swapping, and Face Perception. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 98–105, 2018.
- [152] K. Nissim and A. Wood. Is Privacy Privacy? *The Royal Society*, 376, 2018.
- [153] Kobbi Nissim, Aaron Bembek, Alexandra Wood, Mark Bun, Marco Gaboardi, Urs Gasser, David R. O’Brien, and Salil Vadhan. Bridging the Gap between Computer Science and Legal Approaches to Privacy. *Harvard Journal of Law & Technology*, 31:687–780, 2018.
- [154] R. Nosowsky and T. J. Giordano. The Health Insurance Portability and Accountability Act: Implications for Clinical Research. *Annu Rev Med*, 57:575–590, 2006.
- [155] Paraskevi Nousi, Sotirios Papadopoulos, Anastasios Tefas, and Ioannis Pitas. Deep Autoencoders for Attribute Preserving Face De-Identification. *Signal Processing: Image Communication*, 81:115699, 2020.
- [156] Anne Oeldorf-Hirsch and S. Shyam Sundar. Social and technological motivations for online photo sharing. *Journal of Broadcasting & Electronic Media*, 60(4):624–642, 2016.
- [157] Association of Academic Health Centers. HIPAA Creating Barriers to Research and Discovery. Technical report, 2008. http://www.aahcdc.org/Portals/41/Series/Issue-Briefs/HIPAA_Creating_Barriers_to_Research.pdf?ver=2017-02-17-134959-563.
- [158] Academy of Medical Sciences. Stratified, Personalized or P4 Medicine: A New Direction for Placing the Patient at the Center of Healthcare and Health Education. Technical report, 2005. <https://acmedsci.ac.uk/file-download/38266-56e6d483e1d21.pdf>.
- [159] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless Person Recognition: Privacy Implications in Social Media. In *Computer Vision – ECCV 2016*, pages 19–35, 2016.
- [160] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual Privacy Protection Methods: A Survey. *Expert Syst. Appl.*, 42(9):4177–4195, 2015.
- [161] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. In *British Machine Vision Conference*, 2015.
- [162] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible Conditional GANs for Image Editing. *CoRR*, abs/1611.06355, 2016.
- [163] Y. Pluijmers. The Economic Impacts of Open Access Policies for Public Sector Spatial Information. Technical report, Spatial Information – Open Access and Cooperation, 2000. https://www.fig.net/resources/proceedings/fig_proceedings/fig_2002/Ts3-6/TS3_6_pluijmers.pdf.
- [164] GNU Project. GNU Linear Programming Kit. Ver: 4.65 (2018-02-16), <https://www.gnu.org/software/glpk/>.
- [165] Rachel Cummings and Deven Desai. The Role of Differential Privacy in GDPR Compliance, 2018. http://pwp.gatech.edu/rachel-cummings/wp-content/uploads/sites/679/2018/09/GDPR_DiffPrivacy.pdf, Accessed: October 2, 2019.
- [166] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *4th International Conference on Learning Representations*, 2016.

- [167] Slobodan Ribaric, Aladdin Ariyaeinia, and Nikola Pavesic. De-identification for Privacy Protection in Multimedia Content: A Survey. *Signal Processing: Image Communication*, 47:131 – 151, 2016.
- [168] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *18th Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015.
- [169] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep EXpectation of Apparent Age from a Single Image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.
- [170] M. A. Rothstein. Is De-Identification Sufficient to Protect Health Privacy in Research. *Am J Bioeth*, 10:3–11, 2010.
- [171] P. Samarati. Protecting Respondents’ Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
- [172] P. Samarati and L. Sweeney. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression. Technical report, 1998.
- [173] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.
- [174] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.
- [175] H. Schuchardt. GLPK for Windows. Ver: 4.65 (2018-03-17), <http://winglpk.sourceforge.net/>.
- [176] Sefik Ilkin Serengil. deepface - A Lightweight Face Recognition and Facial Attribute Analysis Framework for Python, 2020. <https://github.com/serengil/deepface>.
- [177] A. Shabo. Health Record Banks: Integrating Clinical and Genomic Data into Patient-Centric Longitudinal and Cross-Institutional Health Records. *Personalized Medicine*, 4(4):453–455, 2007.
- [178] Jonathon Shlens. A Tutorial on Principal Component Analysis. *CoRR*, 2014.
- [179] T. Sim and L. Zhang. Controllable Face Privacy. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 04, pages 1–8, 2015.
- [180] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015.
- [181] D. K. Singh, V. Swaroop, and M. M. Malaviya. Data Security and Privacy in Data Mining: Research Issues & Preparation. *International Journal of Computer Trends and Technology*, 4, 2013.
- [182] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE*, 94:1948 – 1962, 2006.
- [183] Congzheng Song and Vitaly Shmatikov. Overlearning Reveals Sensitive Attributes. In *8th International Conference on Learning Representations*, 2020.
- [184] J. Soria-Comas and J. Domingo-Ferrer. Optimal Data-Independent Noise for Differential Privacy. *Information Sciences*, 250:200 – 214, 2013.

- [185] M. B. Stegmann. The AAM-API, 2003. <http://www.imm.dtu.dk/~aam/aamapi/>, Accessed: November 27, 2018.
- [186] Jose M. Such, Joel Porter, Sören Preibusch, and Adam Joinson. Photo privacy conflicts in social media: A large-scale empirical study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, page 3821–3832, 2017.
- [187] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018.
- [188] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A Hybrid Model for Identity Obfuscation by Face Replacement. In *Computer Vision – ECCV 2018*, pages 570–586, 2018.
- [189] L. Sweeney. Simple Demographics Often Identify People Uniquely. <http://dataprivacylab.org/projects/identifiability/>, 2000.
- [190] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.
- [191] Kyle Taylor and Laura Silver. Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally. Technical report, Pew Research Center, 2019. https://www.pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf.
- [192] Ying-Li Tian, Takeo Kanade, and Jeffrey F. Cohn. Facial Expression Analysis. In Stan Li and Anil Jain, editors, *Handbook of Face Recognition*, chapter 11, pages 247–275. Springer Science+Business Media Inc., 2005.
- [193] Aleksei Triastcyn and Boi Faltings. Generating Differentially Private Datasets Using GANs. *CoRR*, 2018.
- [194] M. A. Turk and A. P. Pentland. Face Recognition Using Eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.
- [195] European Union. Public Sector Information: A Key Resource for Europe. Green Paper on Public Sector Information in the Information Society., 1998. <http://aei.pitt.edu/1168/>.
- [196] U.S. Department of Health & Human Services. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, 2015. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>, Accessed: February 9, 2018.
- [197] J. Vaidya, B. Shafiq, A. Basu, and Y. Hong. Differentially Private Naïve Bayes Classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, Atlanta, USA*, volume 1, pages 571–576, 2013.
- [198] P. L. Venetianer, Z. Zhang, A. Scanlon, Y. Hu, and A. J. Lipton. Video Verification of Point of Sale Transactions. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 411–416, 2007.
- [199] P. Weiss. Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts. In Julie Esanu and Paul Uhler, editors, *Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium*, chapter 17, pages 69–73. The National Academic Press, 2004.

- [200] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel. A Classification of Location Privacy Attacks and Approaches. *Personal and Ubiquitous Computing*, 18(1):163–175, 2014.
- [201] A. D. Weston and L. Hood. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. *Journal of Proteome Research*, 3(2):179–196, 2004.
- [202] Thomas Winkler and Bernhard Rinner. Security and Privacy Protection in Visual Sensor Networks: A Survey. *ACM Comput. Surv.*, 47(1):1–42, 2014.
- [203] W. E. Winkler. Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems. In J. Domingo-Ferrer and V. Torra, editors, *Proceedings of Privacy in Statistical Databases*, pages 231–246. Springer Berlin Heidelberg, 2004.
- [204] W. E. Winkler. Re-identification Methods for Masked Microdata. In J. Domingo-Ferrer and V. Torra, editors, *Proceedings of Privacy in Statistical Databases*, pages 216–230. Springer Berlin Heidelberg, 2004.
- [205] R. C.-W. Wong, A. W.-C. Fu, K. Wang, P. S. Yu, and J. Pei. Can the Utility of Anonymized Data be Used for Privacy Breaches? *ACM Transactions on Knowledge Discovery from Data*, 5(3), 2011.
- [206] Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-Protective-GAN for Privacy Preserving Face De-Identification. *Journal of Computer Science and Technology*, 34(1):47–60, 2019.
- [207] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially Private Generative Adversarial Network. *CoRR*, 2018.
- [208] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. GANobfuscator: Mitigating Information Leakage Under GAN via Differential Privacy. *IEEE Trans. Information Forensics and Security*, 14(9):2358–2371, 2019.
- [209] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren. Information Security in Big Data: Privacy and Data Mining. *IEEE Access*, 2:1149–1176, 2014.
- [210] X. Yao, X. Zhou, and J. Ma. Differential Privacy of Big Data: An Overview. In *Proceedings of the 2nd IEEE International Conference on Big Data Security on Cloud, 2nd IEEE International Conference on High Performance and Smart Computing, and IEEE International Conference on Intelligent Data and Security*, pages 7–12, 2016.
- [211] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially Private Releasing via Deep Generative Model. *CoRR*, 2018.
- [212] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic Image Completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1438–1447, 2019.
- [213] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.