Privacy Protection in Data Collection via Randomized Response Procedures

by Jichong Chai

B.E. in Electronic Engineering, June 2011, Zhejiang University
M.S. in Statistics, May 2014, The George Washington University

A Dissertation submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

May 19, 2019

Dissertation directed by

Tapan K. Nayak
Professor of Statistics

The Columbian College of Arts and Sciences of The George Washington University certifies that Jichong Chai has passed the Final Examination for the degree of Doctor of Philosophy as of April 9th, 2019. This is the final and approved form of the dissertation.

Privacy Protection in Data Collection via Randomized Response Procedures

Jichong Chai

Dissertation Research Committee:

Tapan K. Nayak, Professor of Statistics, Dissertation Director

Sudip Bose, Associate Professor of Statistics, Committee Member

Emre Barut, Assistant Professor of Statistics, Committee Member

Acknowledgements

I would like to first and foremost express my appreciation to my research advisor Professor Nayak, to whom I can never thank enough. I met Professor Nayak as a student enrolled in his course, STAT 6202 during my Master program in 2013. His insightful and knowledgeable lectures not only taught me concepts and theory, but also humbled and inspired me in better understanding the essence of Statistical inference. It was through that memorable experience that encouraged me to pursue a Ph.D. degree in Statistics.

Two years later, I was awarded the opportunity to work closely with Professor Nayak. His mentorship has helped me tremendously in developing my independent research and honing my skills. He has taught me the method of developing new research problems by reading literatures, which is a very crucial part of independent research. Through the direction of Professor Nayak, I have become accustomed to making complexed scientific concepts accessible to audiences of various backgrounds, and have also learned to enunciate my ideas and arguments by coherent writing. These abilities are and will prove beneficial to my career and future endeavors. Over the years, Professor Nayak has been very kind and patient in helping me stay on track, for which I am extremely grateful. It has been an honor to be taught and advised by Professor Nayak.

Furthermore, I would like to also extend my thanks and appreciate to Professors Sudip Bose and Emre Barut for being taking on the task as my dissertation readers. Professor Bose read my dissertation very carefully, making correction to many typos and other errors, both technical and grammatical. Professor Barut has provided me with many constructive suggestions with content and structure of my dissertation.

In addition, I would like to thank Professor Reza Modarres and Dr. Emanuel

Ben-David from The Census Bureau for serving as my dissertation examiners as well as Professor Yinglei Lai for being the committee's chair. I would like to also thank my friends in the Statistics Department, especially Dr. Cheng Zhang, who helped me a lot at the beginning of my research with many insightful discussions that were beneficial to my research.

Finally, I would like to thank two important people who have always been there for me: my parents. Their unconditional support and encouragement has helped me overcome the many difficulties I have been face with through my success and life long journey.

Disclaimer

This dissertation is based in part on the previously published articles listed below. I have permission from my co-authors to use the works listed below in my dissertation.

Chai, J., and Nayak, T. K. (2018). A criterion for privacy protection in data collection and its attainment via randomized response procedures. *Electronic Journal of Statistics*, **12**, 4264-4287. https://doi.org/10.1214/18-EJS1508

Chai, J., and Nayak, T. K. (2019). Minimax Randomized Response Methods for Providing Local Differential Privacy. *U.S. Census Bureau Research Report Series, Statistics*, **04**.

Abstract of Dissertation

Privacy Protection in Data Collection via Randomized Response Procedures

Randomized response (RR) methods have long been suggested for protecting respondents' privacy in statistical surveys. However, how to set and achieve privacy protection goals have received little attention. We give a full development and analysis of the view that a privacy mechanism should ensure that no intruder would gain much new information about any respondent from his response. Formally, we say that a privacy breach occurs when an intruder's prior and posterior probabilities about a property of a respondent, denoted p and p_*, respectively, satisfy $p_* < h_l(p)$ or $p_* > h_u(p)$, where h_l and h_u are two given functions. An RR procedure protects privacy if it does not permit any privacy breach. We explore effects of (h_l, h_u) on the resultant privacy demand, and prove that it is precisely attainable only for certain (h_l, h_u). This result is used to define a canonical strict privacy protection criterion, and give practical guidance on the choice of (h_l, h_u). Then, we characterize all privacy satisfying RR procedures and compare their effects on data utility using sufficiency of experiments and identify the class of all admissible procedures. For linear unbiased estimation, we derive privacy preserving minimax procedures. We address optimal choices for both the RR mechanism (or design) and the estimator. A minimax design is a t-subset design (with a special structure) and it can be implemented fairly easily. We also study mixtures of t-subset designs mainly to examine the RAPPOR method, which is used notably by Google and Apple. We note inadmissibility of the RAPPOR design and offer some suggestions for improving both the design and the customary estimator.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CSIP: Canonical Strict Information Privacy

DLP: Differentially Local Privacy

DP: Differential Privacy

MLE: Maximum Likelihood Estimator

PBR: Privacy Breach Region

PRAM: Post-randomization method

RAPPOR: Randomized Aggregatable Privacy Preserving Ordinal Response

RR: Randomized Response

SIP: Strict Information Privacy

TPM: Transition Probability Matrix

Chapter 1

Introduction and Dissertation Overview

Over past few decades, there have been revolutionary advances in computing, storage and network technologies. Businesses, organizations and governments collect, store and process vast amount of data to gain scientific, economic and societal knowledge and make business and policy decisions. Simultaneously, with the ongoing big data revolution, concerns about privacy and data confidentiality have been increasing substantially. With massive growth of data collection, people are becoming increasingly reluctant to participate in surveys or allow others to use their personal information. Protecting privacy and personal information is essential for legal reasons and for upholding public trust and support. To encourage survey participation and improve accuracy, privacy protection must be taken into consideration when collecting individual level data. Several books, e.g. Willenborg and de Waal (2001), Aggarwal and Yu (2008), Hundepool et al. (2012) and Torra (2017), and many papers discuss various privacy and confidentiality protection methods such as grouping, data swapping, cell suppression, imputation and response randomization.

It is important to discuss the meaning of the words "privacy" and "confidentiality". While privacy and confidentiality have often been used synonymously, those should be distinguished due to some important differences (see Nayak et al., 2015). In legal terms, privacy is a person's right to freedom from intrusion into his/her information. Privacy emerges as a desire to share no or only obscured information with a data collector. Thus, privacy protection should occur at the time of data collection. In contrast, confidentiality is an obligation to prevent unauthorized access to private information. People often give their information trusting that their data will be used by researchers and policy makers only to learn about the population as a whole and not about any individual. Privacy applies to individuals whereas confidentiality applies to the data, which may be addressed after data collection. One important technical (and practical) implication is that one may examine the whole dataset for choosing a suitable method for confidentiality protection. In contrast, methods for privacy protection need to be selected before data collection. Essentially, the main purpose of both privacy and confidentiality protection is to control statistical disclosure. Roughly speaking, there are two types of disclosures, predictive disclosure and identity disclosure.

Identity disclosure is one of the top confidentiality breach concerns. It occurs the record of a survey respondent is exactly identified in released data by an intruder. It is insufficient to remove all direct identifiers, such as name, address and SSN from the original data, since a unit might also be identified by matching some other key variables that are included in the survey such as age or zip code, which are easily available from other sources. In statistics, identity disclosure was discussed by many researchers such as Bethlehem et al. (1990), Reiter (2005) and Shlomo and Skinner (2010). Recently, Nayak et al. (2018) developed a Post-randomization method (PRAM) procedure that strictly controls identification risks at a very little loss of data utility.

On the other hand, predictive disclosure happens when the data gives the intruder extra information such that a sensitive attribute of a respondent can be predicted in high probability. Note that confidentiality protection may control both kind of disclosure, but privacy protection only concerns predictive disclosure risk, since the data collector already knows the identity of the respondent. Consequently, some concepts, such as k-anonymity (Sweeney, 2002), l-diversity (Machanavajjhala et al., 2006) and t-closeness (Li et al., 2007), and related methods are applicable for protecting confidentiality but not privacy, as they focus on identity disclosure. In this dissertation, we mainly concentrate on the privacy protection when collecting data on categorical variables.

Randomized response (RR), the main technique that we use in this dissertation, is a method initially introduced by Warner (1965) to encourage honest participation in survey interview by protecting privacy of respondents. The real answers of respondent are perturbed using a prescribed randomization mechanism before sending to the data collector, by which respondents' privacy is protected. The data collector then makes inferences based on the perturbed data. To be specific, when collecting categorical data, the original category is randomized into a new category by the respective prescribed transition probability. Clearly each original category has the corresponding transition probabilities vector, and these vectors form a transition probability matrix (TPM) P, which determines all statistical properties of the RR mechanism.

Since Warner's pioneering work, statisticians worked on RR theory and methods for protecting privacy in interview surveys. More recently, privacy concerns have expanded significantly, largely in reaction to pervasive data collection from surveys, administrative records, customer information, on-line activities, transactions, searches and postings. That has stimulated strong interest in privacy research in other fields, especially in computer science; see e.g., Aggarwal and Yu (2008), Chen et al. (2009) and Fung et al. (2010). In particular, it has enhanced research interests in RR methods

and their scope of applications. Agrawal and Srikant (2000) first introduced the term "Privacy-Preserving Data Mining" and suggested randomization methods in data perturbation. Rizvi and Haritsa (2002), Evfimievski et al. (2004), Agrawal et al. (2009) and Erlingsson et al. (2014) and others have suggested RR procedures for addressing privacy challenges in emerging contexts such as association rule mining, classification and on-line data collection.

Randomization procedures can also be applied in confidentiality protection, such as Post-randomization method (PRAM) (see Gouweleeuw et al., 1998), where the randomization procedure is chosen and applied to the observed (unperturbed) data set. Many papers, e.g., Van den Hout and Van der Heijedn (2002), Van den Hout and Elamir (2006) and Cruyff et al. (2008), discuss properties, variations and applications of the method. PRAM and RR are closely related as both methods randomize true responses using a transition probability matrix P. However, PRAM and RR are not equivalent (see Nayak et al., 2015 and Nayak and Adeshiyan, 2016). RR is for privacy protection whereas PRAM aims to protect confidentiality. Consequently, one important distinction is that P is fixed in RR, but in PRAM, it can be chosen based on the observed (unperturbed) data set. In particular, in invariant (or unbiased) PRAM, P must depend on the data. How the data dependency of P affect statistical inferences was investigated by Nayak and Adeshiyan (2016).

RR methods are used for privacy protection, but there are various RR mechanisms that are used in applications. A data collector should choose a randomization mechanism based on privacy protection goals. Clearly, larger privacy requirement needs stronger data perturbation. Although the RR topic has been investigated for over 50 years, how to set privacy protection goals and how to choose an RR mechanism to achieve the stated privacy goals have received only modest attention and yielded inadequate guidance. To answer these questions, we first need to articulate what are privacy and its protection. As we noted, privacy emerges as a desire to share no or

only obscured information with a data collector. The vagueness of this concept is the main difficulty in setting privacy goals, as different people may interpret it differently. We shall set a clear, precise, and strongly relevant privacy protection goal and define suitable privacy measures in Chapter 3.

Data perturbation will inevitably cause information loss and impact data utility. In general, we want to minimize the impact of data perturbation as long as we assure that the privacy protection goals are achieved. To be specific, we want to select a suitable perturbation procedure under the privacy criterion such that the statistical information loss is minimized. However, the statistical information loss is difficult to formally define, so it is unrealistic to devise a unified and explicit data utility measure to cover all scenarios and data structures. Thus, people use various utility measures for different purposes. In Warner's randomized response survey, data collectors are often interested in estimating the distribution of a single dichotomous variable. Hence, data utility can be measured by the mean square error of estimator. However, when considering general categorical variables, there are various loss functions to use for estimating several parameters simultaneously. In the area of privacy preserving data mining, data are used for various purposes such as association rule mining or classification. Therefore, using the estimation risk as utility measure is not suitable. In brief, the data utility depends on the utility (loss) function, which is usually different according to the specific scenario and purpose.

A utility function for perturbed data is not necessarily the same as the original unperturbed problem. One may not always treat the perturbed data as the original data and use the same utility measure. In an RR survey, the main purpose is to estimate the distribution of a sensitive variable. The data collector cannot access the original data and has to use a perturbed version. However, this does not mean the data collector would use the perturbed data directly. They would estimate the distribution and reconstruct the original data because they know the randomization

procedure. The reconstructed data would be used in future analysis, so the evaluation of data utility should be based on the reconstructed data. It seems that the data utility should depend on the comparison of the reconstructed data and the original unperturbed data. However, the unperturbed data is not observable, and it is just a random sample of the population. Therefore, the data utility should be defined on population level via a statistical sampling model, not merely by a comparison of the original and reconstructed data.

On the other hand, the reconstruction cannot be done without knowing the randomization procedure P. As an example, the randomization procedure P depends on the data in invariant PRAM. Therefore, data publishers usually do not provide the randomization procedure because releasing the randomization procedure may lead to disclosure. Moreover, data agencies may only release some summary statistics instead of the whole dataset, which makes reconstruction impossible. Hence, data users have to treat the perturbed dataset (or outputs) as the original, and the measure of data utility should be based on the comparison of the original and perturbed data directly.

Ideally, we want to find a universally optimal randomization procedure under all possible data utility measures. However, we will see later that such procedures usually do not exist. Therefore, we should use some "good" procedures based on some reasonable criteria. One approach of solving this problem is to choose a reasonable data utility measure and then choose an optimal P by this utility measure. As we discussed, the data utility measure should be logically sound and highly relevant to the actual problem.

Another approach for choosing "good" procedure focuses only on P, regardless of the inferential goals. Data may be used and analyzed in different ways and for various purposes. It may not even be possible to anticipate all future usage of the data at the time of the survey. Especially in privacy preserving data mining area, evaluating the

utility loss for data perturbation based on the purpose could be more complicated than parameter estimation. Therefore, most of the popular utility functions in this area depend only on P, though it may not always be relevant to the actual problem. In this dissertation, we shall explore both approaches and propose our utility measures.

The rest of the dissertation is organized as follows. In Chapter 2, we introduce Warner's RR method for binary variables and some other methods. We introduce the generalization of RR in categorical variables and the basic notation setup for the dissertation.

In privacy protection, a data collector is just an intruder, so the intruder may have some prior information about a target respondent. Privacy breach happens when the intruder gets much new information from the data about the respondent's attributes. The best way to quantify information is by subjective probability. In Chapter 3, we briefly review some privacy measures using subjective probability and generalize those measures to set a compelling privacy protection goal. The goal is to guarantee that no intruder will gain much new information about any respondent from any his/her response. We formalize this idea in full generality and propose Strict Information Privacy (SIP).

In Chapter 4, we give explicit guidance on choosing RR mechanisms based on the privacy protection goal. We make characterization on SIP and explore its properties and implications. We show that an RR mechanism P provides SIP completely depends on its parity. Conversely, parity can be used to measure privacy level of an RR procedure. In summary, parity plays a core role in choosing RR mechanisms.

Under specified privacy level, one approach for choosing RR procedures is data utility comparison. Intuitively, if an RR procedure A is dominated by some other procedure P under all utility measures, we should use P instead of A. In Chapter 5, we investigate this idea by formally defining admissibility using Blackwell's notion of

sufficiency of experiments. If there exists a TPM C such that $A = CP$, P is sufficient for A. Because of the additional randomization, A cannot be more informative than P. Hence, A is inadmissible and should be excluded from the possible choices. We give a full characterization of all admissible procedures, which is the main result of this chapter.

In Chapter 6, we derive some optimality results under specified data utility measures. Section 6.1 considers a criterion that only depends on P. In Section 6.2, we consider linear unbiased estimation. We show that under squared error loss and a given design P, the best linear estimator usually does not exist, as the locally best linear estimator may depend on the unknown parameter. Therefore, we use the minimax risk criterion proposed by Duchi et al. (2018) on linear unbiased estimation. We show that t-subset designs are the optimal under this criterion. We also give a corresponding minimax estimator.

In Chapter 7, we give a practical implementation for t-subset design. We also study mixtures of t-subset designs mainly to examine the RAPPOR method, which is used notably by Google and Apple. We note inadmissibility of the RAPPOR design and offer some suggestions for improving both the design and the customary estimator.

We make some remarks and discuss future research in Chapter 8.

Chapter 2

A Review of Randomized Response Procedures

As mentioned above, our work is mainly about privacy protection in categorical data collection. Discussion of confidentiality protection, such as methods that aim to control identity disclosure, is out of the range of this dissertation. The data perturbation method that we use in this dissertation is Randomized Response (RR). In this chapter, we first introduce Warner's pioneering method, and some subsequent works in the literature. Then, we present our notation and setup for the dissertation.

2.1 A Review of Warner's Method and Previous Works

Randomized Response is a well-known data perturbation method in survey sampling. It was first proposed by Warner (1965) to protect survey participants' privacy. Suppose a surveyor want to collect some sensitive information from the participants. However, he cannot ask the sensitive question directly because a participant may

not be willing to tell the truth. For instance, if the survey question is "Are you a drug user?", the proportion of the answer "Yes" will be much smaller than the real proportion, since drug users tend to conceal this sensitive information. This leads to serious biases in statistical results, and those biases are difficult to evaluate. In addition, some people may refuse to participate in the survey to protect their privacy.

Warner's RR scheme is as follows. Suppose A is a sensitive category (e.g. drug user) that is of interest to the surveyor. Each respondent is asked to randomly choose one of the following two questions Q_1 and Q_2 with probabilities p and $1-p$, where p is pre-chosen and a part of the survey design.

Q_1: Do you belong to A?

Q_2: Do you belong to A^C?

Then, the respondent is asked to answer "Yes" or "No" without telling the true question to the surveyor. Note that the surveyor does not know which question is answered, so the privacy of respondent is protected. As a result, the surveyor can assume the response is true and make inferences based on survey data.

Suppose a respondent belongs to A. He/she will respond "Yes" when Q_1 is answered and "No" when Q_2 is answered. So we can conclude $P(\text{Yes}|A) = p$ and $P(\text{No}|A) = 1-p$. Similarly, $P(\text{Yes}|A^C) = 1-p$ and $P(\text{No}|A^C) = p$. Using these conditional probabilities, we obtain the probability of "Yes" response as

$$\begin{aligned} P(\text{Yes}) &= P(\text{Yes}|A)P(A) + P(\text{Yes}|A^C)P(A^C) \\ &= p\pi + (1-p)(1-\pi) \\ &= (2p-1)\pi + (1-p) = \lambda, \textit{say}. \end{aligned} \tag{2.1}$$

Let $\hat{\lambda}$ be the relative frequency of "Yes" in the sample, which is an unbiased estimator

of λ. Then, using (2.1) an unbiased estimator of π is

$$\hat{\pi} = \frac{\hat{\lambda} - (1-p)}{2p-1}.$$

It should be noted that the estimator of $\hat{\pi}$ is not the MLE, as it may fall out of [0,1]. The MLE is obtained by setting $\hat{\pi}$ to 0 or 1 when it is smaller than 0 or larger than 1, respectively (see Nayak, 1994).

Warner's scheme is not the unique possible choice. Greenberg et al. (1969) proposed another method that uses an unrelated non-sensitive question "Q_3: Do you belong to B?" instead of Q_2. Note that the transition probability matrix P is known only if $P(B)$, the probability of answering "Yes" to Q_3, is known. In privacy protection, there is no need to use an RR method with unknown P, by which one cannot easily estimate π from one sample. For RR methods used for binary characteristics, Nayak (1994) proposed a unified framework, which shows that the essence of RR procedures is the choice of conditional probabilities $P(\text{Yes}|A)$ and $P(\text{No}|A^C)$. He also proposed admissible designs and investigated implementations of RR.

The original RR is used in dichotomous variables, and then extended to categorical variables in Abul-Ela et al. (1967) and Warner (1971). We refer to the books Chaudhuri and Mukerjee (1988), Chaudhuri (2010) and Fox (2016) for comprehensive discussions and further references. RR methods can also be used with quantitative characteristics, which is quite different from categorical cases in perturbation. This dissertation focuses only on categorical cases.

As we mentioned, the essence of an RR procedure is a matrix of transition probabilities. Designing an RR procedure reduces to choosing P suitably, taking both privacy protection and data utility (or accuracy of estimators of π from RR data) into account. However, many papers compared RR designs by comparing only the

variance of estimators under competing designs. That is misleading. A fair comparison should also require a common level of privacy protection. Some authors did that, but mainly for binary characteristics, see e.g., Anderson (1976), Fligner et al. (1977), Nayak (1994) and Nayak and Adeshiyan (2009). For a general categorical variable, the RR literature does not give much guidance on how to choose P. We believe that the choice of P was not addressed properly because privacy measures and precise privacy protection goals were not developed until very recently. We shall introduce some related privacy measures in the literature and set our privacy protection goals in the next chapter.

2.2 Our RR Setup

This section will introduce the basic RR notation setups in this dissertation. We consider a categorical survey variable X with the set of possible categories $\mathcal{S}_X = \{c_1, \ldots, c_k\}$. Let $\pi_i, i = 1, \ldots, k$, denote the population level relative frequencies of $c_1, \ldots, c_k$, which are unknown. We collect data to estimate $\pi = (\pi_1, \ldots, \pi_k)'$ and make other inferences about π. If there are several variables, a categorical variable can be generated by cross-classification.

To protect privacy, an RR survey asks each respondent to use a given random mechanism to generate and report a perturbed version Z of his/her true value of X. Denote the output space by $\mathcal{S}_Z = \{d_1, \ldots, d_m\}$. The transition probabilities $p_{ij} = P(Z = d_i | X = c_j), i = 1, \ldots, m, j = 1, \ldots, k$, are prespecified and embedded in the randomization device. The matrix $P = ((p_{ij}))$, called the transition probability matrix (TPM), determines all statistical properties of the RR mechanism, and designing an RR survey essentially reduces to choosing P. Thus, we shall identify an RR procedure by its underlying P. We shall require that each row of P contains at least one nonzero element to define m and $\mathcal{S}_Z$ unambiguously. Let $\lambda_i = P(Z = d_i)$,

and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)'$ denotes the distribution of Z. Then

$$\lambda_i = P(Z = d_i) = \sum_{j=1}^{k} P(Z = d_i | X = c_j) P(X = c_j) = \sum_{j=1}^{k} p_{ij} \pi_j,$$

and hence

$$\lambda_{m \times 1} = P_{m \times k} \pi_{k \times 1}. \tag{2.2}$$

Note that if the ith row of P is zero, then $P(Z = d_i)$ is always zero and d_i is irrelevant. The columns of P are also called channel distributions; see Duchi et al. (2018). Note that the sample spaces $\mathcal{S}_X$ and $\mathcal{S}_Z$ of X and Z, respectively, need not be the same, or even have the same cardinality. For example, $m = 2^k$ in the RAPPOR algorithm of Erlingsson et al. (2014). We briefly introduce the basic procedure of RAPPOR as follows.

RAPPOR is an RR design for privacy protection, which is currently used by Google and Apple. RAPPOR maps true categories into indicator vectors. To be specific, it represents each true category with a row vector $X = (X_1, \ldots, X_k)$ whose ith component is 1 if the true category is c_i and 0 otherwise. We shall denote the sample size by n and then the unperturbed data, i.e., true categories of n sampled units, can be given as an $n \times k$ matrix $\mathcal{D}_*$, with each row showing the true category of one unit. In perturbation, each component will change independently with a specified probability p, so the randomized response is $Z = (Z_1, \ldots, Z_k)$ and the data generated using RAPPOR can be given as a matrix $\mathcal{D}$ of order $n \times k$. It is seen that the output space has dimension $m = 2^k$.

In general, we shall denote the sample frequency of d_i by S_i, for $i = 1, \ldots, m$. We shall assume simple random sampling, in which case the frequency vector $S = (S_1, S_2, \ldots, S_m)'$ has a multinomial distribution, viz. $S \sim mult(n, \lambda)$. The relative frequency vector $\hat{\lambda} = S/n$ is a natural (and method of moments) estimator of λ.

Usually, inferences about λ can be translated into inferences about π, via (2.2). For example, if P is square and nonsingular, an estimator $\tilde{\lambda}(S)$ of λ gives the estimator $\tilde{\pi} = P^{-1}\tilde{\lambda}(S)$ for π. Note that if $m < k$, or more generally if the columns of P are not linearly independent, the model for Z is not identifiable with respect to π and hence π is not estimable.

Remark 2.1. *In a parametric model, if π is a function of fewer parameters, identifiability might hold even when $m < k$. However, identifiability with respect to π ensures that the data set can be analyzed using different models for various (possibly unforeseen) purposes. Thus, inference and data utility considerations suggest to use $m \geq k$ and P with rank k.*

In an estimation problem, one should consider risk for utility comparison. Generally, risk depends on three factors: (a) the RR design, (b) the estimator and (c) the loss function. If considering $\tilde{\pi} = P^{-1}\tilde{\lambda}(S)$ mentioned above and variance (squared loss), we get

$$\begin{aligned} \mathbf{V}(\tilde{\pi}) &= \frac{P^{-1}(D_\lambda - \lambda\lambda')(P^{-1})'}{n} = \frac{P^{-1}D_\lambda(P^{-1})' - \pi\pi'}{n} \\ &= \frac{P^{-1}D_\lambda(P^{-1})' - D_\pi}{n} + \frac{D_\pi - \pi\pi'}{n}, \end{aligned} \tag{2.3}$$

where D_λ is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_k$ and D_π is defined similarly (see Chaudhuri and Mukerjee, 1988, p.43). The first term is the "variance inflation" by the RR procedure, and the second term is due to simple random sampling.

Obviously the design is the pivotal factor as the estimator depends the RR design P. However, a reasonable estimator usually may not be obvious as in this example. If P is not a square matrix, such as the RAPPOR design, P^{-1} does not exist so (2.3) does not hold. Consequently, we need to take both the RR design P and

the corresponding estimator into account in utility comparison. Furthermore, as we mentioned in Chapter 1, the loss function is also an important factor. In Chapter 5, we shall consider some admissibility results using Blackwell's (1951, 1953) notion of *sufficiency of experiments*, which is agnostic about inferential goals and loss functions. In Section 6.1, we consider a specific utility measure only based on P. In Section 6.2, we shall use total variance as the loss function and derive the optimal design and a corresponding estimator, under the minimax criterion.

Chapter 3

A General Criterion for Privacy Protection

In this chapter, we focus only on privacy protection. Privacy violations occur in many forms depending on data type, privacy desires and intruders' knowledge and behavior. Thus, various privacy concepts and measures have appeared in the literature, including identity disclosure, differential privacy, k-anonymity and l-diversity (see Chen et al., 2009). However, these measures apply to data, so they are mainly about confidentiality but not privacy. As Kifer and Lin (2012) noted, most privacy measures are developed intuitively and can lead us astray, and thus one should use privacy criteria that are logically sound and practical. We shall introduce two such privacy measures in the literature.

3.1 A Review of Two Privacy Measures

In the preceding framework, Evfimievski et al. (2003) defined ρ_1-to-ρ_2 privacy, taking a Bayesian view. For a target respondent B, suppose an intruder R's prior probability

of $X = c_j$ is α_j, and let $\alpha = (\alpha_1, \ldots, \alpha_k)'$. Note that an intruder's prior α about a target may be quite different from π. For a given prior α, $P_\alpha(Z = d_i) = \sum_{l=1}^{k} \alpha_l p_{il}$ and the posterior probability of $X = c_j$ given B's response $Z = d_i$, is

$$P_\alpha(X = c_j | Z = d_i) = \frac{P_\alpha(X = c_j, Z = d_i)}{P_\alpha(Z = d_i)} = \frac{\alpha_j p_{ij}}{\sum_{l=1}^{k} \alpha_l p_{il}}. \tag{3.1}$$

Also, R's prior and posterior probabilities of any $Q \subseteq \mathcal{S}_X = \{c_1, \ldots, c_k\}$ are:

$$P_\alpha(X \in Q) = \sum_{j: c_j \in Q} \alpha_j \quad \text{and} \quad P_\alpha(X \in Q | Z = d_i) = \sum_{j: c_j \in Q} P_\alpha(X = c_j | Z = d_i). \tag{3.2}$$

For brevity, we shall denote $P_\alpha(X \in Q)$ by $P_\alpha(Q)$ and $P_\alpha(X \in Q | Z = d_i)$ by $P_\alpha(Q|d_i)$. The following two criteria require that $P_\alpha(Q)$ and $P_\alpha(X \in Q | Z = d_i)$ are not too different by certain measures.

Definition 3.1. *(Evfimievski et al., 2003) Let $0 < \rho_1 < \rho_2 < 1$ be two numbers.*

(a) An RR procedure is said to permit an upward ρ_1-to-ρ_2 privacy breach with respect to $Q \subseteq \mathcal{S}_X$ and a prior distribution α if for some $1 \leq i \leq m$ with $P_\alpha(Z = d_i) > 0$,

$$P_\alpha(Q) < \rho_1 \quad \textit{and} \quad P_\alpha(Q|d_i) > \rho_2. \tag{3.3}$$

Similarly, a procedure is said to admit a downward ρ_2-to-ρ_1 privacy breach if

$$P_\alpha(Q) > \rho_2 \quad \textit{and} \quad P_\alpha(Q|d_i) < \rho_1$$

for some d_i with $P_\alpha(Z = d_i) > 0$.

(b) An RR procedure is said to provide ρ_1-to-ρ_2 privacy protection if it does not permit an upward ρ_1-to-ρ_2 privacy breach or a downward ρ_2-to-ρ_1 privacy breach with respect to any Q and any prior α.

Definition 3.2. *(Nayak et al., 2015) For a given $\beta > 1$,*

(a) an RR procedure admits a β-factor privacy breach, with respect to $Q \subseteq \mathcal{S}_X$ and a prior α if $P_\alpha(Q) > 0$ and

$$\frac{P_\alpha(Q|d_i)}{P_\alpha(Q)} > \beta \quad or \quad \frac{P_\alpha(Q|d_i)}{P_\alpha(Q)} < \frac{1}{\beta} \tag{3.4}$$

for some d_i such that $P_\alpha(Z = d_i) > 0$.

(b) An RR procedure provides β-factor privacy if it does not permit a β-factor breach with respect to any Q and any α.

The above two criteria are very strong, as they require no privacy breach for any d_i, Q and α. Thus, no answer (d_i) of a respondent B would give "much" new information to any intruder R (characterized by α) about any property (Q) of B with respect to X. In practice, values of (ρ_1, ρ_2) and β should be chosen based on the sensitivity of X and privacy concerns. Here, the β-factor privacy is simpler as it requires us to specify only one number (β). Interestingly, the strict privacy requirements of the two criteria are achievable, as summarized below.

Definition 3.3. *(Nayak et al., 2015) The ith row parity of P is defined as*

$$\eta_i(P) = \max\left\{\frac{p_{ij}}{p_{il}} \;\middle|\; j, l = 1, \ldots, k\right\} = \frac{\max_j\{p_{ij}\}}{\min_j\{p_{ij}\}}, \tag{3.5}$$

with the convention $0/0 = 1$ and $a/0 = \infty$ for any $a > 0$.

Furthermore, the parity of P is defined as $\eta(P) = \max_i\{\eta_i(P)\}$.

Clearly, $\eta(P) \geq 1$ and it is finite if and only if all elements of P are positive. The concept of parity is very similar to γ-amplification of Evfimievski et al. (2003). We state the following theorems by the concept of parity.

Theorem 3.1. *(Evfimievski et al., 2003) A sufficient condition for an RR procedure*

with transition probability matrix P to guarantee ρ_1-to-ρ_2 privacy is:

$$\eta(P) \leq \frac{\rho_2(1-\rho_1)}{\rho_1(1-\rho_2)}. \tag{3.6}$$

Theorem 3.2. *(Nayak et al., 2015) An RR procedure guarantees β-factor privacy if and only if*

$$\eta(P) \leq \beta.$$

We shall see later that (3.6) is also a necessary condition for P to provide ρ_1-to-ρ_2 privacy. We should mention that Boreale and Paolini's (2015) concept of "worst-case breach" is essentially the same as β-factor breach. They also proved a version of Theorem 3.2.

Also, both ρ_1-to-ρ_2 and β-factor privacy are achievable. For any given η_0, it is possible to construct P with parity η_0; see Evfimievski et al. (2003) and Agrawal et al. (2009). In particular, for $m \geq k$, one P with $\eta(P) = \eta_0$ is obtained by taking

$$p_{ii} = \frac{\eta_0}{\eta_0 + m - 1}, i = 1, \ldots, k,$$

and

$$p_{ij} = \frac{1}{\eta_0 + m - 1} \quad \forall i \neq j.$$

3.2 Comments on ρ_1-to-ρ_2 and β-factor Privacy

We shall further discuss ρ_1-to-ρ_2 and β-factor privacy and develop a general criterion. We begin with some logical and practical features of Definitions 3.1 and 3.2. The ρ_1-to-ρ_2 and β-factor privacy criteria are very strong, but it should be noted that those are applicable only when an intruder R knows his/her target B's value of Z. Typically, this happens at data collection time, with R being the data collector.

In commercial data mining context, Agrawal et al. (2009) refer to this as business-to-customer (or B2C) privacy. Definitions 3.1 and 3.2 are not applicable if R gets access only to an anonymized version of the original data set, where B's records cannot be ascertained with certainty. In other words, ρ_1-to-ρ_2 and β-factor privacy criteria presumes disclosure of B's identity to an intruder, so it is purely for predictive disclosure protection.

As we discuss next, logically ρ_1-to-ρ_2 and β-factor privacy directly address the core of privacy concern, which is: how much *information* an intruder might gain about a respondent from his/her response (possibly perturbed)? One compelling view of information, as Basu (1988) articulated, is: "Information is what information does. It changes opinion." Furthermore, opinion can be expressed precisely only using subjective probability. An intruder's prior and posterior probabilities describe respectively his/her initial and revised opinion, after learning a respondent's reported value. These constitute a strong argument that privacy should be discussed in terms of intruders' prior and posterior probabilities (instead of technical information measures, e.g., mutual information and f-divergence, that were developed in other contexts). Definitions 3.1 and 3.2 coincide with the above view and are thus highly relevant to privacy considerations.

The changing of a prior to posterior occurs only through the likelihood function, and the change is small if the likelihood function is relatively flat. In our setting, for response d_i, the likelihoods for $c_1, \ldots, c_k$ are $p_{ij}, j = 1, \ldots, k$, and they are fairly close to each other when $\eta_i(P)$ is small. Consequently, the likelihood functions for all possible responses are fairly flat if and only if $\eta(P)$ is fairly small. This comes out precisely in Theorems 3.1 and 3.2.

We now mention a connection to the following concept (see, Duchi et al., 2018) of differential local privacy.

Definition 3.4. *(Duchi et al., 2018) An RR method provides ϵ-differentially local privacy (ϵ-DLP), for $\epsilon > 0$, if*

$$\max\left\{\frac{P(Z \in S|X = c_j)}{P(Z \in S|X = c_l)} \;\middle|\; S \subseteq \mathcal{S}_Z, 1 \leq j, l \leq k\right\} \leq \exp(\epsilon). \tag{3.7}$$

It can be seen that (3.7) is equivalent to $\eta(P) \leq \exp(\epsilon)$. So, in view of Theorem 3.2, ϵ-DLP and β-factor privacy are equivalent, with $\beta = \exp(\epsilon)$. An equivalency of ϵ-DLP and ρ_1-to-ρ_2 privacy can be observed similarly. Clearly, the equivalence holds as each one corresponds to an upper bound on $\eta(P)$, but the thinking behind Definitions 3.1, 3.2 and 3.4 are different. As we discussed, Definitions 3.1 and 3.2 are based on the idea of information gain, while Definition 3.4 is an extreme case of Differential Privacy, briefly discussed next.

Differential Privacy (DP), introduced by Dwork (2008), is a privacy criterion that has received huge attention, especially in Computer Science literature. DP concerns leakage of information about any respondent from releasing data summaries or inferences, which are essentially statistics. Let D denote the dataset, and $T(D)$ denote the statistic. DP uses a randomization mechanism M to perturb the statistic $T(D)$, and release the perturbed output $M(T(D))$. Denoting the output $M(T(D))$ as $\mathcal{K}(D)$, we state the criterion of DP below.

Definition 3.5. *An output mechanism $\mathcal{K}$ provides ϵ-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \in Range(\mathcal{K})$,*

$$\frac{P(\mathcal{K}(D_1) \in S)}{P(\mathcal{K}(D_2) \in S)} \leq \exp(\epsilon). \tag{3.8}$$

Clearly, $\mathcal{K}$ is determined by both the output statistic T and the randomization mechanism M. The output statistic T is usually determined by a client's query, which is a common summary statistic such as mean, median, or a count of specific properties

in the dataset. Once T is determined, M is properly chosen by the data agency to satisfy DP.

We shall discuss differences and the relation between the Definition 3.4 and 3.5. First, DP is applied to protect confidentiality, not privacy. In the context of DP, the true dataset is known to the data agency (or server), while in RR the true sensitive categories are unknown. Second, outputs are different. In DP, the output is usually just a perturbed summary statistic, but in RR, outputs are the randomized responses of all respondents, which consist of the whole dataset. By RR, one can use the perturbed dataset for various purposes with privacy preserved. Third, RR is an extreme case of DP. In RR, the output statistic is the microdata, i.e. $T(D) = D$, and M is a perturbation that apply P to each individual independently. It can be shown that for an independent RR mechanism, ϵ-DLP and ϵ-DP are equivalent.

3.3 The Proposed Privacy Criterion

The main goal of this section is to explore the central idea of ρ_1-to-ρ_2 and β-factor privacy in full generality. Figure 1 gives a helpful geometrical perspective of ρ_1-to-ρ_2 and β-factor privacy. The two shaded rectangles represent the privacy breach region (PBR) of ρ_1-to-ρ_2 privacy, as any (prior, posterior) pair, to be denoted generically by (p, p_*), falling in this region signifies a privacy breach. The two shaded triangles constitute the PBR of β-factor privacy.

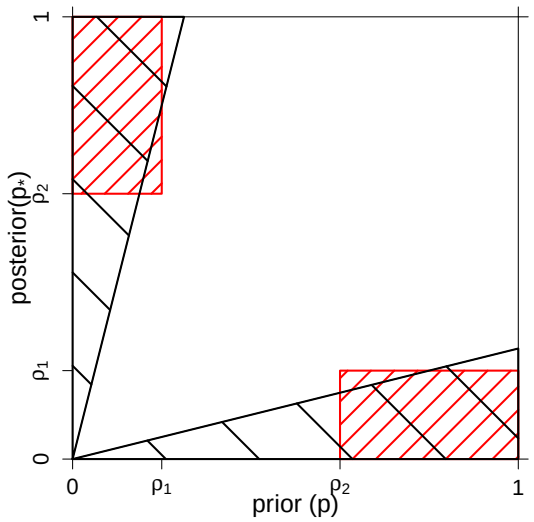

Figure 1: ρ_1-to-ρ_2 and β-factor privacy breach regions.

In practice, visual inspection of various PBRs might help to choose the parameter values, e.g., (ρ_1, ρ_2) or β, of a privacy criterion, and also to compare different privacy guarantees. Naturally, a larger PBR implies a stronger privacy guarantee. Among two PBRs in Figure 1, none is a subset of the other one, but as the β-factor PBR has a larger area and covers most of the other PBR, one might reasonably consider it stronger. As such, two overlapping PBRs, as in Figure 1, are not comparable, but we shall see in Chapter 4 that privacy demands of any two PBRs can be compared meaningfully. The shape of PBR may not necessarily be square or triangle, which are just special cases. In fact, it can be in any reasonable shape. Thus, we now consider a general privacy breach region W, as shown by the shaded region in Figure 2, and require that no (prior, posterior) pair must fall in W. So, the unshaded part (W^c) is the privacy holding region. Describing the down and up privacy breach boundaries of W with two functions h_l and h_u, we introduce the following.

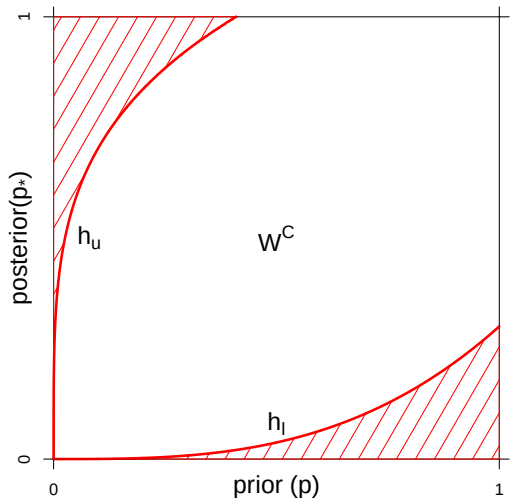

Figure 2: A general privacy breach region.

Definition 3.6. *Let h_l and h_u be two functions from $[0,1]$ to $[0,1]$ such that $0 \leq h_l(a) \leq a \leq h_u(a) \leq 1$ for all $0 \leq a \leq 1$. An RR procedure is said to provide strict information privacy (SIP) with respect to h_l and h_u, to be abbreviated (h_l, h_u)-SIP, if*

$$h_l(P_\alpha(Q)) \leq P_\alpha(Q|d_i) \leq h_u(P_\alpha(Q)) \tag{3.9}$$

for $i = 1, \ldots, m$ and all $\alpha, Q \subseteq \mathcal{S}_X$.

Clearly, the general idea is that for privacy protection, if the prior probability of an event is p, then its posterior probability must be between $h_l(p)$ and $h_u(p)$. Obviously, this covers Definitions 3.1 and 3.2 as special cases. Mathematically, we do not need to put additional conditions on h_l and h_u, but intuitively, they should be nondecreasing. If a change between prior and posterior is defined as a privacy breach, any larger change must also be a privacy breach. Definition 3.6 specifies a *privacy demand* with

its PBR being

$$W[h_l, h_u] = \{(p, p_*), 0 \leq p, p_* \leq 1 : p_* < h_l(p) \text{ or } p_* > h_u(p)\}.$$

Essentially, the (prior, posterior) combination in the PBR will never happen by using an RR procedure that provides SIP.

After having the privacy criterion SIP, we need to consider its attainment. In the next chapter, we shall characterize all RR procedures that can provide the specified privacy demand.

Chapter 4

Characterization of Strict Information Privacy

In Chapter 3, we focused on the privacy demand and proposed our criterion, Strict Information Privacy (SIP). In this chapter, we shall characterize all RR procedures that provide SIP for a specified privacy demand and investigate its implications. If an RR procedure P can provide SIP for a given PBR, the *privacy provided* by that RR procedure should satisfy the specified privacy demand. Therefore, we should start with an RR procedure P first. Intuitively, some (prior, posterior) combination will never happen by using the RR procedure, and it can also be described by a privacy breach region as defined next.

Definition 4.1. *We define the PBR of any RR procedure P as the collection of all non-attainable (prior, posterior) pairs under P, and denote it by W_P. Thus, W_P is the complement (with respect to the unit square) of P's privacy holding region:* $\{(p, p_*), 0 \leq p, p_* \leq 1 : P_\alpha(Q) = p \text{ and } P_\alpha(Q|d_i) = p_* \text{ for some } d_i, \alpha \text{ and } Q \subseteq \mathcal{S}_X\}$.

Note that W_P, the PBR of P, is the privacy provided by the RR procedure P, and should be distinguished from a general privacy breach region W. The next definition

gives a connection between these two concepts.

Definition 4.2. *We shall call a general privacy breach region W precise if there exists an RR procedure P such that $W_P = W$.*

The preceding two definitions are useful for comparing and matching privacy demand with privacy provided by different procedures. Clearly, an RR procedure P provides (h_l, h_u)-SIP if and only if $W[h_l, h_u] \subseteq W_P$. However, if $W[h_l, h_u]$ is not precise, to guarantee (h_l, h_u)-SIP one must use an RR procedure P for which W_P is *strictly larger* than $W[h_l, h_u]$, and in such cases, we should report W_P, the PBR of the procedure actually used, to communicate the privacy guarantee precisely and maximally. This also implies that to determine privacy requirement we should think only in terms of precise PBRs. These observations raise some natural questions, such as: What are the precise PBRs? Which procedures satisfy a given precise PBR? For given h_l and h_u, is there a minimal W_P satisfying $W[h_l, h_u] \subseteq W_P$? We answer these questions in the following sections of this chapter.

4.1 The Equivalence of Upward and Downward Privacy

We begin with some analytic simplifications of the (h_l, h_u)-SIP criterion. First, note that

$$P_\alpha(Q) = 0 \Rightarrow P_\alpha(Q|d_i) = 0 \text{ and } P_\alpha(Q) = 1 \Rightarrow P_\alpha(Q|d_i) = 1, \text{ for all } d_i.$$

So, (3.9) holds automatically when $P_\alpha(Q)$ is 0 or 1, and to establish (h_l, h_u)-SIP, we need to verify (3.9) only for all $\alpha, Q \subseteq \mathcal{S}_X$ such that $0 < P_\alpha(Q) < 1$. Second, observe

that the first $\leq$ in (3.9) is equivalent to

$$1 - h_l(P_\alpha(Q)) \geq 1 - P_\alpha(Q|d_i) \Leftrightarrow P_\alpha(Q^C|d_i)) \leq 1 - h_l(1 - P_\alpha(Q^C)).$$

So, the first $\leq$ in (3.9) holds for all $Q \subseteq \mathcal{S}_X$ if and only if

$$P_\alpha(Q|d_i) \leq 1 - h_l(1 - P_\alpha(Q)) \text{ for all } Q \subseteq \mathcal{S}_X,$$

i.e., the condition for downward privacy breach is equivalent to an upward privacy breach criterion. (Evfimievski et al. (2003) made a similar observation for ρ_1-to-ρ_2 privacy.) Combining the two upward breach conditions and defining

$$[h_l \star h_u](a) = \min\{h_u(a), 1 - h_l(1 - a)\}$$

for $0 \leq a \leq 1$, we obtain the following:

Lemma 4.1. *Let h_l and h_u be as in Definition 3.6 and $[h_l \star h_u]$ be defined as above. Then, an RR procedure P provides (h_l, h_u)-SIP if and only if*

$$P_\alpha(Q|d_i) \leq [h_l \star h_u](P_\alpha(Q)) \tag{4.1}$$

for all $i = 1, \ldots, m$ and all α and $Q \subseteq \mathcal{S}_X$ such that $0 < P_\alpha(Q) < 1$.

The conditions $0 \leq h_l(a) \leq a \leq h_u(a) \leq 1$ imply that $a \leq [h_l \star h_u](a) \leq 1$ for all $0 \leq a \leq 1$. In view of Lemma 4.1 and the preceding discussions, we define a privacy criterion more succinctly only in terms of upward breaches as follows.

Definition 4.3. *Let $h : [0, 1] \to [0, 1]$ be a function satisfying $a \leq h(a) \leq 1$ for all $0 \leq a \leq 1$. An RR procedure is said to provide canonical strict information privacy*

with respect to h, to be abbreviated h-CSIP, if

$$P_\alpha(Q|d_i) \leq h(P_\alpha(Q)) \tag{4.2}$$

for $i = 1, \ldots, m$ and all α and $Q \subseteq \mathcal{S}_X$ such that $0 < P_\alpha(Q) < 1$.

It can be seen that h-CSIP also provides the downward privacy guarantee that

$$P_\alpha(Q|d_i) \geq \tilde{h}(P_\alpha(Q))$$

for $i = 1, \ldots, m$ and all α and $Q \subseteq \mathcal{S}_X$, where $\tilde{h}(a) = 1 - h(1-a), 0 \leq a \leq 1$. Thus, the upper and lower boundaries of the PBR of h-CSIP are given by h and $\tilde{h}$, respectively. Lemma 4.1 shows that for any h_l and h_u, (h_l, h_u)-SIP and $[h_l \star h_u]$-CSIP are equivalent, in the sense that if an RR procedure guarantees one of the two, it must also guarantee the other one. However, the PBR given by some (h_l, h_u)-SIP can be a proper subset of the PBR of the corresponding $[h_l \star h_u]$-CSIP. This is illustrated in Figure 3, where the PBR of $[h_l \star h_u]$-CSIP is the PBR of (h_l, h_u)-SIP (shown as the region shaded with solid lines) plus the two dotted lined parts A and B. The two PBRs would be identical only when $h_l(a) = 1 - h_u(1-a), 0 \leq a \leq 1$. While Definition 4.3 is technically most concise, in real applications, it might be more convenient to specify h_l and h_u, defining lower and upper privacy breaches, and then take $h = [h_l \star h_u]$. Subsequently, we shall explore only h-CSIP, because it is the analytical crux of any privacy criterion as seen above.

4.2 The Characterization of h-CSIP

To characterize all RR procedures that provide h-CSIP for any given h, we first introduce the following result, which says that to assure (4.2) for all $Q \subseteq \mathcal{S}_X$, it

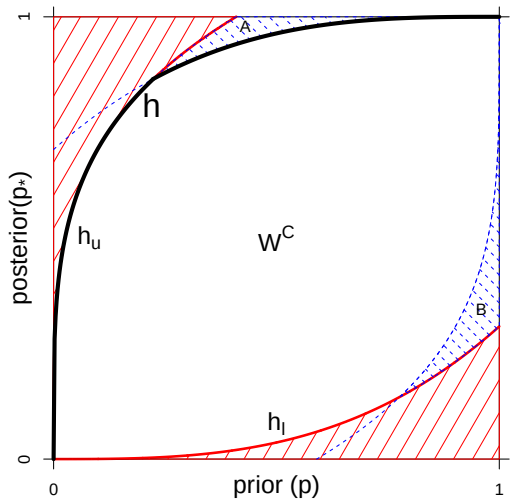

Figure 3: Connection between privacy breach regions of (h_l, h_u)-SIP and $[h_l \star h_u]$-CSIP.

suffices to verify (4.2) only for $\{X = c_j\}$, i.e., for the atomic events of X.

Lemma 4.2. *An RR procedure P provides h-CSIP, with a specified h, if and only if*

$$P_\alpha(X = c_j | Z = d_i) \leq h(P_\alpha(X = c_j)) \tag{4.3}$$

for all i, j and α such that $P_\alpha(X = c_j) > 0$ and $P_\alpha(Z = d_i) > 0$.

Proof. The 'only if' part of the lemma follows readily. So, we shall prove only the 'if' part. Suppose (4.3) holds. Now, take any α, $Q \subseteq \{c_1, \ldots, c_k\}$ and d_i such that $P_\alpha(Q) > 0$ and $P_\alpha(Z = d_i) > 0$. Suppose $c_q \in Q$ is such that $p_{iq} \geq p_{ij}$ for all j such that $c_j \in Q$. Consider the prior $\tilde{\alpha}$ with elements: $\tilde{\alpha_j} = \alpha_j$ if $c_j \notin Q$, $\tilde{\alpha_q} = P_\alpha(Q)$, and

$\tilde{\alpha}_j = 0$ for all other j. Then, we have $P_{\tilde{\alpha}}(X = c_q) = \tilde{\alpha}_q = P_\alpha(Q)$ and

$$\begin{aligned} P_\alpha(Q|Z = d_i) &= \frac{\sum_{j:c_j \in Q} \alpha_j p_{ij}}{\sum_{j:c_j \in Q} \alpha_j p_{ij} + \sum_{j:c_j \notin Q} \alpha_j p_{ij}} \\ &\leq \frac{p_{iq}(\sum_{j:c_j \in Q} \alpha_j)}{p_{iq}(\sum_{j:c_j \in Q} \alpha_j) + \sum_{j:c_j \notin Q} \alpha_j p_{ij}} \\ &= \frac{\tilde{\alpha}_j p_{iq}}{\sum_{j=1}^k \tilde{\alpha}_j p_{ij}} = P_{\tilde{\alpha}}(X = c_q|Z = d_i) \\ &\leq h(P_{\tilde{\alpha}}(X = c_q)) = h(P_\alpha(Q)), \end{aligned}$$

where the first inequality holds by the fact that for $a > 0$, $\psi(x) = \frac{x}{x+a}$ is an increasing function of x over $(0, \infty)$ and the second inequality follows from (4.3). □

It suffices to characterize all P that satisfy (4.3). For any given h, we define

$$B(h) = \inf_{0<p<1} \Big(\frac{1-p}{p}\Big)\Big(\frac{h(p)}{1-h(p)}\Big), \tag{4.4}$$

where we take $h(p)/[1 - h(p)] = \infty$ when $h(p) = 1$. (Alternatively, we can take the infimum over $\{0 < p < 1 : h(p) < 1\}$.)

Lemma 4.3. *A necessary and sufficient condition for an RR procedure P to satisfy (4.3) for all i, j and α is that $\eta(P) \leq B(h)$.*

Proof. We shall prove that P satisfies (4.3) if and only if $\eta_i(P) \leq B(h)$ for $i = 1, \ldots, m$. Take any (fixed) i, and note that (4.3) holds if $\alpha_j = 0$ or 1, or $p_{ij} = 0$. Take any j such that $p_{ij} > 0$ (which exists as each row of P contains at least one nonzero element). Then, for $0 < \alpha_j < 1$, we can write:

$$P(X = c_j|Z = d_i) = \frac{\alpha_j p_{ij}}{\sum_{l=1}^k \alpha_l p_{il}} = \Big[1 + \Big(\frac{1-\alpha_j}{\alpha_j}\Big)\frac{1}{p_{ij}} \sum_{l:l\neq j} \Big(\frac{\alpha_l}{1-\alpha_j}\Big)p_{il}\Big]^{-1}. \tag{4.5}$$

In view of (4.5), for our fixed i and chosen j, (4.3) holds for all α if and only if

$$\sum_{l:l\neq j}(\frac{\alpha_l}{1-\alpha_j})(\frac{p_{il}}{p_{ij}}) \geq (\frac{\alpha_j}{1-\alpha_j})(\frac{1-h(\alpha_j)}{h(\alpha_j)}) \tag{4.6}$$

for all α such that $0 < \alpha_j < 1$.

Letting $w_l = \dfrac{\alpha_l}{1-\alpha_j}$ for $l \neq j$, it is seen that (4.6) is equivalent to

$$\sum_{l:l\neq j} w_l(\frac{p_{il}}{p_{ij}}) \geq (\frac{\alpha_j}{1-\alpha_j})(\frac{1-h(\alpha_j)}{h(\alpha_j)})$$

for all $0 < \alpha_j < 1$ and all $\{w_l\}$ such that $0 \leq w_l \leq 1$ and $\sum_{l\neq j} w_l = 1$. This holds if and only if

$$\inf_{\{w_l\}}(\sum_{l:l\neq j} w_l(\frac{p_{il}}{p_{ij}})) \geq \sup_{0<p<1}(\frac{p}{1-p})(\frac{1-h(p)}{h(p)}). \tag{4.7}$$

The infimum in (4.7) is $\min\{\frac{p_{il}}{p_{ij}} \mid l = 1,\ldots,k, l \neq j\}$. Moreover, it must be positive in order to satisfy (4.7), because $h(p)$ cannot be 1 for all $0 < p < 1$ and hence the right side of (4.7) is positive. This implies that p_{ij} must be positive for all $j = 1,\ldots,k$. So, for our fixed i, (4.3) holds for all j and α if and only if (4.7) holds for $j = 1,\ldots,k$, or equivalently,

$$\min\{\frac{p_{il}}{p_{ij}} \mid j,l = 1,\ldots,k\} \geq \sup_{0<p<1}(\frac{p}{1-p})(\frac{1-h(p)}{h(p)}). \tag{4.8}$$

Note that both sides of (4.8) are positive and finite, and the above inequality can be recognized as $[\eta_i(P)]^{-1} \geq [B(h)]^{-1}$, which yields $\eta_i(P) \leq B(h)$. □

The main result follows readily from the two lemmas.

Theorem 4.1. *For any given h, an RR procedure P provides h-CSIP if and only if $\eta(P) \leq B(h)$.*

The necessary and sufficient condition in Theorem 4.1 depends on h only through

$B(h)$ and on P only through its parity $\eta(P)$. Thus, in h-CSIP context, $B(h)$ quantifies the privacy demand of h and $\eta(P)$ is the privacy level of P. In particular, Theorems 3.1 and 3.2 can be obtained from Theorem 4.1 by calculating $B(h)$ for relevant h functions. We can measure of the privacy demand of any general PBR, with downward and upward breach boundaries h_l and h_u, as $B(h)$, where $h = [h_l \star h_u]$; recall that $[h_l \star h_u](a) = \min\{h_u(a), 1 - h_l(1-a)\}, 0 \leq a \leq 1$. Using this measure, we can compare the privacy demands of any two PBRs, even when they overlap as in Figure 1. Likewise, we can compare the privacy level of all RR procedures P using parity.

Consider an RR procedure P with $\eta(P) = \gamma > 1$. Then, by Theorem 4.1 and (4.4), P guarantees h-CSIP for all h such that

$$\gamma \leq \Big(\frac{1-p}{p}\Big)\Big(\frac{h(p)}{1-h(p)}\Big) \quad \text{for all } 0 < p < 1,$$

or equivalently $h(p) \geq h_{(\gamma)}(p)$, where $h_{(\gamma)}(p)$ is defined as

$$h_{(\gamma)}(p) = \frac{\gamma p}{1 + (\gamma - 1)p}, \quad 0 < p < 1. \tag{4.9}$$

Thus, $h_{(\gamma)}(.)$ is the up breach boundary of any P with parity γ. The corresponding down boundary is $\tilde{h}_{(\gamma)}(p) = 1 - h_{(\gamma)}(1-p)$. Note that the PBR of P is determined only by its parity. So, all P with a common parity have the same PBR. It also follows that W is a precise PBR if and only if its up and down breach boundaries are $h_{(\gamma)}$ and $\tilde{h}_{(\gamma)}$, respectively, for some $\gamma > 1$. As γ increases, $h_{(\gamma)}(p)$ shifts upward and the PBR gets smaller, as shown in Figure 4.

Let $\mathcal{H} = \{h_{(\gamma)}(.); \gamma > 1\}$, i.e., the class of all function of the form $h_{(\gamma)}(.)$. Then, h-CSIP with all $h \in \mathcal{H}$ represent all precise PBRs, which are most relevant to choosing privacy requirement and communicating privacy guarantee. A logical conclusion is that for strict privacy protection, we should think only in terms of h-CSIP and limit

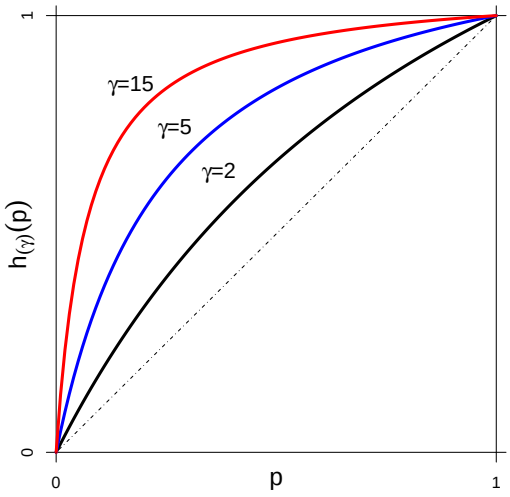

Figure 4: Plots of precise up breach boundaries $h_{(\gamma)}$.

h to $\mathcal{H}$. A practical meaning of $h_{(\gamma)}$-CSIP may not be immediate, but as we show next, this amounts to imposing a bound on all Bayes factors.

For given α, the prior odds of Q is $P_\alpha(Q)/[1-P_\alpha(Q)]$ and its posterior odds given $Z = d_i$ is $P_\alpha(Q|d_i)/[1 - P_\alpha(Q|d_i)]$. Now, take any $\gamma > 1$ and consider the following privacy requirement:

$$\frac{P_\alpha(Q|d_i)}{1 - P_\alpha(Q|d_i)} \frac{1 - P_\alpha(Q)}{P_\alpha(Q)} \leq \gamma \tag{4.10}$$

for all α, Q and d_i such that $0 < P_\alpha(Q) < 1$ and $P_\alpha(Z = d_i) > 0$. The left side of (4.10) is the ratio of posterior odds of Q to its prior odds, or the Bayes factor for testing $X \in Q$ against $X \notin Q$; see Kass and Raftery (1995) for a very informative discussion of Bayes factor. Thus, (4.10) requires all Bayes factors to be less than or equal to γ. Considering Q^C, it can be seen that (4.10) also implies that all Bayes factors are at least γ^{-1}. In summary, (4.10) requires all Bayes factors to be between γ^{-1} and γ. The above criteria is analogous to β-factor privacy; while (4.10) uses the ratio of posterior and prior odds, β-factor privacy uses the ratio of the two probabilities. In

other words, β-factor privacy uses probability scale whereas (4.10) uses odds scale.

By routine algebra, it can be seen that (4.10) is equivalent to

$$P_\alpha(Q|d_i) \leq \frac{\gamma P_\alpha(Q)}{1+(\gamma-1)P_\alpha(Q)}. \tag{4.11}$$

Notice that the right side of (4.11), considered as a function of $P_\alpha(Q)$, is the same as the function $h_{(\gamma)}(.)$ defined in (4.9). So, $h_{(\gamma)}$-CSIP is equivalent to the privacy requirement of (4.10) and γ can be interpreted conveniently as the upper bound on all Bayes factors (and the lower bound is γ^{-1}). It is also seen that any *precise* PBR corresponds to the privacy requirement of (4.10), with a matching value of γ. Based on previous discussions we reach the following practical conclusions.

(1) While ρ_1-to-ρ_2 and β-factor privacy and more generally (h_l, h_u)-SIP are intuitively sensible, we should discuss, assess and communicate privacy only in terms of h-CSIP with $h \in \mathcal{H}$, or equivalently in terms of bounds on Bayes factors as in (4.10). Both the graphical representation, as in Figure 4, and the Bayes factor interpretation should be helpful for choosing suitable values of γ in practical applications. Kass and Raftery (1995) recommend that one should interpret a Bayes factor 20 or larger as strong evidence, which suggests that values around 20 might be suitable for γ in our context.

(2) Satisfying any privacy requirement reduces strictly to using a procedure with a sufficiently small parity, as stated in Theorem 4.1. We can always find a procedure to provide required privacy, but not uniquely because for any $\gamma > 1$, there exists many P with $\eta(P) = \gamma$. We should compare data utility to choose one procedure among all privacy satisfying procedures. We discuss this in the next chapter.

(3) Our results also bring out a new interpretation of ϵ-DLP (see Definition 3.4). Recall that an RR procedure P provides ϵ-DLP if and only if $\eta(P) \leq \exp(\epsilon)$. Then,

in view of Theorem 4.1 and subsequent discussions, ϵ-DLP is equivalent to $h_{(\gamma)}$-CSIP, with $\gamma = \exp(\epsilon)$. Thus, ϵ-DLP can be explained by the PBR of the corresponding $h_{(\gamma)}$-CSIP.

Chapter 5

Comparison of Data Utility

In earlier chapters, we observed that a randomization procedure P provides strict privacy protection if and only if $\eta(P) \leq \gamma$, where $\gamma > 1$ is determined by the privacy requirement. In this chapter, we shall compare data utility of privacy satisfying RR, namely all P with $\eta(P) \leq \gamma$. However, before that we shall introduce another logical restriction on P.

Recall that $P_{m\times k}$ must be a transition probability matrix (TPM) and each row of P must contain at least one nonzero value. We also argue that no two rows of P should be proportional to each other. The ith row is the (nonparametric) likelihood function when $Z = d_i$. So, from likelihood perspective, if rows i and j are proportional, the statistical information from observing $Z = d_i$ and $Z = d_j$ are the same and the two outcomes (and the corresponding rows) should be merged. Alternatively, two proportional rows can be viewed as obtained from randomly splitting one outcome into two. (This is analogous to irrelevantly splitting one choice into two in discrete choice analyses; e.g., splitting "bus" into "blue bus" and "green bus" in mode of transportation choice.) Also, if proportional rows are allowed, then $\mathcal{S}_Z$ and m cannot be defined uniquely.

With the natural constraints discussed above, the class of all privacy preserving procedures, at a desired level γ, is:

$$\mathcal{C}_\gamma = \{P_{m\times k} : m \geq 2, \eta(P) \leq \gamma \text{ and } P \text{ has no proportional rows}\}.$$

As we noted earlier, one may also impose $m \geq k$ and $rank(P) = k$, for model identifiability. However, these are not needed for our results. Intuitively, we should compare data utility to select a procedure from $\mathcal{C}_\gamma$ for application.

However, "data utility" is difficult to define and measure as the data may be used and analyzed in different ways and for various purposes. It may not even be possible to anticipate all future usage of the data at the time of the survey. Recognizing this, we shall first discuss some admissibility results using Blackwell's (1951, 1953) notion of *sufficiency of experiments*, which is agnostic about inferential goals and loss functions.

5.1 The Blackwell's Criterion

Adopting Blackwell's (1951, 1953) criterion to our context, we introduce the following:

Definition 5.1. *For two randomization procedures $A_{r\times k}$ and $P_{m\times k}$, we say that P is at least as informative (or good) as A, to be denoted $P \succeq A$, if there exists a transition probability matrix $C_{r\times m}$ such that $A = CP$. In this case, P is also said to be sufficient for A.*

If $P \succeq A$ and also $A \succeq P$, then A and P are equivalent and will be denoted $P \sim A$. We say that P is better than A and write $P \succ A$ if $P \succeq A$ but A is not sufficient for P, i.e., $A \not\succeq P$.

It is easy to see that if C and P are TPMs, then $A = CP$ is also a TPM.

The intuitive idea behind Definitions 5.1 is that if $A = CP$, then the procedure A is equivalent to further randomizing (by C) the output of P, and because of the additional randomization, A cannot be more informative than P. Mathematically, it follows that if P is sufficient for A, then for any inference problem with a given loss function and any inference rule δ based the data from A, there exists a rule δ_* based on P such that the risk of δ_* is never larger than the risk of δ (see Lehmann, 1988).

As Definition 5.1 is agnostic about inferential goals and loss functions, one should use P instead of A. However, a fair comparison should take both privacy and data utility into consideration. In privacy literature, Blackwell's criterion has also been used by Kairouz et al. (2016b) to derive results in . We propose the following definition, which connects Blackwell's criterion to our main result in Chapter 4:

Definition 5.2. *A randomization procedure $A \in \mathcal{C}_\gamma$ is said to be inadmissible within $\mathcal{C}_\gamma$ if there exists $P \in \mathcal{C}_\gamma$ such that $P \succ A$. Otherwise, A is called admissible.*

It is seen that if both A and P satisfy the privacy criterion, but P is always better than A in data utility, we should not use A. Naturally, one should use only admissible procedures and we shall characterize all admissible procedures in the next section. Before that, we have several remarks to make and results to introduce:

Remark 5.1. *The restriction that our TPMs must not contain proportional rows can be further justified as follows. Consider a procedure $A_{m\times k}$ and suppose its first two rows, denoted a_1 and a_2 are proportional and $a_1 = \delta(a_1 + a_2), 0 < \delta < 1$. Construct $P^*_{(m-1)\times k}$ by merging the first two rows of A. Then, A and P^* are equivalent, as $A = CP^*$ and $P^* = C^*A$ with C and C^* defined as:*

$$C = \left(\begin{array}{c|c} \delta & 0 \\ 1-\delta & 0 \\ \hline 0 & I \end{array}\right) \quad \textit{and} \quad C^* = \left(\begin{array}{cc|c} 1 & 1 & 0 \\ \hline 0 & 0 & I \end{array}\right).$$

We can repeat this process to eliminate all proportional rows and thus obtain a P such that P has no proportional rows and $P \sim A$.

Remark 5.2. *Intuitively, permuting the rows of $P_{m\times k}$, i.e., relabeling the elements of $\mathcal{S}_Z$, should have no effect on either privacy or data utility. Mathematically, this holds easily. Specifically, it can be seen that if $C_{m\times m}$ is a permutation matrix and $A = CP$, then (i) $\eta(A) = \eta(P)$ and (ii) $A \sim P$ (as C^{-1} is a also a permutation matrix and hence TPM).*

Interestingly, the converse is also true, as the following result states. The proof uses Lemma 5.1, which will be stated and proved later.

Theorem 5.1. *Two procedures $A_{m\times k}, P_{r\times k} \in \mathcal{C}_\gamma$ are equivalent if and only if $m = r$ and $A = CP$, where C is a permutation matrix.*

Proof. The 'if' part follows easily as noted in Remark 5.2. To prove the 'only if' part, suppose A and P are equivalent, i.e., there exist two TPMs $C_{m\times r}$ and $C^*_{r\times m}$ such that $A = CP$ and $P = C^*A$. Then, $(C^*C)_{r\times r}$ and $(CC^*)_{m\times m}$ are TPMs. Also, $C^*CP = C^*A = P$ or $(C^*C - I)P = 0$, and similarly, $(CC^* - I)A = 0$. These imply, by Lemma 5.1 (given below), that both (CC^*) and (C^*C) are identity matrices and consequently we must have $m = r$ (since both of them are of full rank) and $C^{-1} = C^*$. Now, since both are TPMs, C must be a permutation matrix (Minc, 1988, p. 3). □

Lemma 5.1. *Suppose $B_{m\times m}$ and $P_{m\times k}$ are two transition probability matrices, P has no zero or proportional rows and $(B - I)P = 0$. Then, $B = I$, the identity matrix of order m.*

The proof of this lemma is quite long and technical. We shall use the following concepts and results in matrix algebra and Markov chain to prove it.

Definition 5.3. *(Chakravarti, 1975) A square matrix $B_{m\times m}$ is said to be reducible if there exists a permutation matrix Q such that*

$$Q^{-1}BQ = \begin{bmatrix} R & 0 \\ L & N \end{bmatrix}, \tag{5.1}$$

where R and N are square matrices. Otherwise, B is called irreducible.

If B is the TPM of a Markov chain, then B is irreducible means that one can always find a path between any two states. Note that $Q^{-1}BQ$ permutes the diagonal entries of B by exchanging the corresponding row and columns. We call this *diagonal permutation* in the following. In (5.1), if R or N are still reducible, they can be further reduced to the above form through diagonal permutation. Actually, if B is reducible, then through diagonal permutation, we can get a block lower-triangular matrix with irreducible diagonal blocks.

Theorem 5.2. *(Chakravarti, 1975) If a non-negative matrix $B_{m\times m} = ((b_{ij}))$ is irreducible, then the matrix $F = B - D(r)$ must have rank $m-1$, where $D(r)$ is the diagonal matrix with entries $(r_1, r_2, \ldots, r_m)$, and $r_j = \sum_{i=1}^{m} b_{ij}$.*

Definition 5.4. *(Taussky, 1949) The column j of a square matrix $B = ((b_{ij}))$ is called weakly diagonal dominant, if $\sum_{i\neq j} |b_{ij}| \leq |b_{jj}|$. It is called strictly diagonal dominant if '$<$' holds.*

Theorem 5.3. *(Taussky, 1949) Suppose $B_{m\times m}$ is an irreducible matrix, and all columns of B are weakly diagonal dominant and at least one column is strictly diagonal dominant. Then, B is nonsingular.*

Proof of Lemma 5.1. First, suppose $B_{m\times m}$ is irreducible, if possible. Let V_0 denote the vector space that is orthogonal to the row space of $(B-I)$. Note that if $(B-I)P = 0$, then all columns of P must be in V_0. Applying Theorem 5.2 to B, noting that each

column of B adds to 1 as B is a TPM, we obtain $rank(B-I) = m-1$. This implies that the dimension of V_0 is 1 and hence all columns of P are proportional. Actually, they are identical (as the sum of each column is 1) and hence all rows of P are also identical. This contradicts the assumption that P has no proportional rows. Thus, B cannot be irreducible.

Next, suppose B is reducible. Then, there exists a permutation matrix Q such that $Q^{-1}BQ$ is a block lower-triangular matrix with irreducible diagonal blocks $R_1, R_2, ..., R_g$. If all of these blocks are 1×1 identity matrices, then $B = I$ as $Q^{-1}BQ$ is TPM. If not, suppose R_{t+1} with dimension $s \times s$ is the first block that is not 1×1 identity matrix. This implies all the off-diagonal entries on the first t columns of $Q^{-1}BQ$ must be 0 (when $t \geq 1$). Take such a Q and denote $R = R_{t+1} = ((r_{ij}))$ to obtain

$$Q^{-1}BQ = \begin{bmatrix} I_t & 0 & 0 \\ 0 & R_{s\times s} & 0 \\ 0 & L & N \end{bmatrix}. \tag{5.2}$$

Note that $(B-I)P = 0$ implies that $QQ^{-1}(B-I)QQ^{-1}P = 0$ or

$$(Q^{-1}BQ - I)P^* = 0, \tag{5.3}$$

where $P^* = Q^{-1}P$, which is also a TPM with no proportional rows. In view of (5.2), equation (5.3) implies that $(R-I)P_s = 0$, where P_s consists of the $t+1$ to $t+s$ rows of P^*. Here, each column of P_s is orthogonal to the rows of $(R-I)$, and hence must be in V_1, the orthogonal space to the row space of $(R-I)$. We shall consider two cases to examine $rank(R-I)$ and its implication.

(i) $L = 0$ or L does not exist (i.e., $s = m-t$). Here, R is a TPM. Also, $s \geq 2$, since R is not 1×1 identity matrix. Apply Theorem 5.2 to R and the arguments used earlier (for irreducible B) to see that $rank(R-I) = s-1$. So, the dimension

of V_1 is 1, implying that all columns of P_s are constant multiples of a common vector and consequently all rows of P_s are the same. This contradicts the fact that P^* has no proportional rows. So, L cannot be a null matrix.

(ii) $L \neq 0$. Here, we shall apply Theorem 5.3 to $R^* = (R-I) = ((r^*_{ij}))$. First, R^* is irreducible, as R is so. Next, as each column of the right side of (5.2) adds to 1, we get $\sum_{i=1}^{s} r_{ij} \leq 1$ for $j = 1, \ldots, s$, and '$<$' holds for at least one j, as $L \neq 0$. This shows, in view of $0 \leq r_{ij} \leq 1, r^*_{ii} = r_{ii} - 1$ and $r^*_{ij} = r_{ij}$ for $i \neq j$, that $\sum_{i \neq j}^{s} |r^*_{ij}| \leq |r^*_{jj}|$, for $j = 1, \ldots, s$ and '$<$' holds for some j. Thus, R^* satisfies the conditions of Theorem 5.3 and hence $rank(R^*) = s$, i.e., $(R-I)$ is nonsingular. Now, $(R-I)P_s = 0$ implies that $P_s = 0$, which is a contradiction. From the above discussion of all possible cases we must conclude that $B = I$. □

5.2 Characterization of Admissible Procedures

Since only admissible procedures should be used, we shall now characterize all admissible procedures within $\mathcal{C}_\gamma$. Let $\mathcal{C}^a_\gamma$ denote all P in $\mathcal{C}_\gamma$ satisfying the following two conditions:

C1: $\eta_i(P) = \gamma$ for all i

C2: Each row of P contains exactly two distinct values.

Our main result of this chapter is:

A randomization procedure P is admissible within $\mathcal{C}_\gamma$ *if and only if* $P \in \mathcal{C}^a_\gamma$.

We state the two directions of the result in Theorem 5.4 and 5.5. We first focus on the *if* part of the main result.

Lemma 5.2. *Suppose $C_{m\times r}$ and $P_{r\times k}$ are two TPMs, $\eta(P)=\gamma$ and let $A_{m\times k}=CP$. Then, $\eta(A)\le\gamma$.*

Proof. Note that $\eta(P)=\gamma$ implies that $p_{uj}\le\gamma p_{ul}$ for all u,j and l. So, for all i,j and l,

$$a_{ij}=\sum_{u=1}^{r}c_{iu}p_{uj}\le\sum_{u=1}^{r}c_{iu}(\gamma p_{ul})=\gamma a_{il} \tag{5.4}$$

and thus $\eta_i(A)\le\gamma$ for $i=1,\ldots,m$, and consequently $\eta(A)\le\gamma$. □

This result is intuitive: further randomization should not reduce privacy (by increasing parity). It also exhibits a trade-off between privacy and data utility: if P is at least as informative as A, in the sense of $P\succeq A$, then A provides at least as much privacy (by parity measure) as P.

Theorem 5.4. *Any randomization procedure $A\in\mathcal{C}_\gamma^a$ is admissible within $\mathcal{C}_\gamma$.*

Proof. Take any $A\in\mathcal{C}_\gamma^a$. We shall prove that if any $P\in\mathcal{C}_\gamma$ is sufficient for A, then A must be equivalent to P. Suppose there exist $P_{r\times k}\in\mathcal{C}_\gamma$ and a TPM $C_{m\times r}$ such that $A=CP$. Each row of C must contain at least one nonzero element, as A does not have any zero row. We shall see that $c_{iu}\neq 0$ implies that the uth row of P, denoted p_u, is proportional to a_i, the ith row of A. For each i, as $\eta_i(A)=\gamma$, by C1, there exist j and l such that $a_{ij}=\gamma a_{il}$. For such a_{ij} and a_{il}, equality holds in (5.4) and since $c_{iu}\neq 0$, we must have $p_{uj}=\gamma p_{ul}$. This holds for all j and l such that $a_{ij}=\gamma a_{il}$. Since a_i contains exactly two distinct (nonzero) values, by C2, considering all pairs (a_{ij},a_{il}) with $a_{ij}=\gamma a_{il}$, it is seen that $p_u\propto a_i$.

The preceding result implies that each row of C has exactly one nonzero entry; otherwise, P will have proportional rows. Then, if $m<r$, C must have some zero columns hence would not be a TPM. Also, if $m>r$, at least two rows of C must be proportional, and the corresponding rows of A are also proportional, which is a

contradiction. So, we must have $m = r$ and C must be a permutation matrix, to be a TPM, and thus $A \sim P$. □

We should note that in the preceding proof we not only showed that $A \sim P$ but also that P must be a permutation of the rows of A. Consequently, $P \in \mathcal{C}_\gamma^a$, as $A \in \mathcal{C}_\gamma^a$, and we have following.

Corollary 5.1. *If $A \in \mathcal{C}_\gamma^a$, then no $P \in (\mathcal{C}_\gamma \setminus \mathcal{C}_\gamma^a)$ can be equivalent to A. Stated another way, if $A \in \mathcal{C}_\gamma^a, P \in (\mathcal{C}_\gamma \setminus \mathcal{C}_\gamma^a)$ and $A \succeq P$, then $A \succ P$.*

Now we shall prove the *only if* part of the main result, stated below.

Theorem 5.5. *A randomization procedure $A \in \mathcal{C}_\gamma$ is admissible within $\mathcal{C}_\gamma$ only if A satisfies C1 and C2, i.e., $A \in \mathcal{C}_\gamma^a$.*

We organize the proof in several parts. To prove this theorem, it is equivalent to show that all $A \in (\mathcal{C}_\gamma \setminus \mathcal{C}_\gamma^a)$ are inadmissible within $\mathcal{C}_\gamma$. We should consider all forms of violations of the conditions C1 and C2. We state Lemmas 5.3 and 5.4 to cover all possible violations of C1.

Lemma 5.3. *Suppose $A_{m\times k} \in \mathcal{C}_\gamma$ and $1 < \eta_i(A) < \gamma$ for some i. Then, A is inadmissible.*

Proof. Suppose, without loss of generality, $1 < \eta_1(A) = a_{11}/a_{12} < \gamma$. Then, there exists a row i such that $a_{i1} < a_{i2}$ because each column of A adds to 1. For brevity suppose that $a_{21} < a_{22}$. Construct $P_{m\times k}$ as follows: $p_1 = a_1 + (1-\xi)a_2, p_2 = \xi a_2$, where $\xi \geq 1$ is a constant (to be chosen suitably) and $p_i = a_i, i = 3, \ldots, m$. Note that as all elements of A are positive, implied by $\eta(A) \leq \gamma$, there exists ξ_0 such that P is a TPM for all $1 \leq \xi < \xi_0$. Also, $\eta_i(A) = \eta_i(P)$ for $i = 2, \ldots, m$ and so, any difference in $\eta(A)$ and $\eta(P)$ comes from the difference between $\eta_1(A)$ and $\eta_1(P)$. Next, note

that

$$\eta_1(P) = \max_{ij}\left\{\frac{p_{1i}}{p_{1j}}\right\} = \max_{ij}\left\{\frac{a_{1i}+(1-\xi)a_{2i}}{a_{1j}+(1-\xi)a_{2j}}\right\}$$

is a continuous function of ξ, and for $\xi = 1$, $\eta_1(P) = \eta_1(A) < \gamma$. So, there exists $1 < \xi < \xi_0$ for which $\eta_1(P) \leq \gamma$ and consequently, $\eta(P) \leq \gamma$. Take such a value ξ_* and use that in the construction of P.

Finally, note that $P = CA$ and $A = C^{-1}P$, with

$$C = \left(\begin{array}{cc|c} 1 & 1-\xi_* & 0 \\ 0 & \xi_* & 0 \\ \hline 0 & 0 & I \end{array}\right) \quad \text{and} \quad C^{-1} = \left(\begin{array}{cc|c} 1 & 1-1/\xi_* & 0 \\ 0 & 1/\xi_* & 0 \\ \hline 0 & 0 & I \end{array}\right).$$

Now, as $\xi_* > 1, C^{-1}$ is a TPM and hence $P \succeq A$. Also, as C is nonsingular, $P = DA$ only with $D = C$. But, C is not a TPM, as $\xi_* > 1$, and hence A is not sufficient for P. In summary, $P \succ A$ and hence A is inadmissible. □

Lemma 5.4. *Suppose $A \in \mathcal{C}_\gamma$, $\eta_i(A)$ equals 1 or γ for all i, and $\eta_i(A) = 1$ for some i. Then, there exists $P \in \mathcal{C}_\gamma$ such that $P \succeq A$ and P satisfies the condition C1.*

Proof. Note that A can have at most one constant row because $A \in \mathcal{C}_\gamma$ and thus cannot have proportional rows. For notational simplicity, suppose that $\eta_1(A) = 1$, i.e., the all values in row 1 of A are the same, say δ. Then, from likelihood perspective, the response d_1 does not give any information about π. Intuitively, we may eliminate the response d_1 and distribute its probability (proportionally) to other responses. Specifically, construct P, from A, by deleting the first row and multiplying all other elements by $(1-\delta)^{-1}$. It can be seen easily that row parity of the retained rows remain the same and $A_{m\times k} = C_{m\times(m-1)}P_{(m-1)\times k}$, where all elements of the first row of C are δ and the remaining rows constitute $(1-\delta)I_{m-1}$. Thus, P satisfies C1 and $P \succeq A$. □

If P as constructed in Lemma 5.4 also satisfies C2, i.e., $P \in \mathcal{C}_\gamma^a$, then from Corollary 5.1, it follows that $P \succ A$ and hence A is inadmissible. As we shall show next, if P does not satisfy C2, then P is inadmissible, which implies A is also inadmissible. The following lemma also completes the proof of Theorem 5.5.

Lemma 5.5. *Any randomization procedure $A \in \mathcal{C}_\gamma$ that satisfies the condition C1 but not C2 is inadmissible within $\mathcal{C}_\gamma$.*

Proof. Suppose $A_{m\times k} \in \mathcal{C}_\gamma$ satisfies C1 but not C2. Thus, $\eta_i(A) = \gamma$ for $i = 1, \ldots m$, and at least one row of A contains more than two distinct values. For notational simplicity, suppose the first row contains three or more distinct values and a_{11} is a "middle" value, i.e., $t < a_{11} < T$, where $t = \min_i\{a_{1i}\}$ and $T = \max_i\{a_{1i}\}$. Note that $T/t = \gamma$ as A satisfies C1. Let $\delta = (T - a_{11})/(T - t)$. Consider $P^*_{(m+1)\times k}$ whose rows are: $p_1 = \delta(t, a_{12}, \ldots, a_{1k}), p_2 = (1-\delta)(T, a_{12}, \ldots, a_{1k})$ and $p_i = a_{i-1}, i = 3, \ldots, m+1$. It can be verified easily that P^* is a TPM, $\eta_i(P^*) = \gamma$ for $i = 1, \ldots m + 1$ and $A = CP^*$, with $C = \left(\begin{array}{cc|c} 1 & 1 & 0 \\ \hline 0 & 0 & I \end{array}\right)$ and thus $P^* \succeq A$. Repeat the process to eliminate all "middle" values of A and if it creates any proportional rows, add those as per Remark 5.1. The resulting P belongs to $\mathcal{C}_\gamma^a$ and $P \succeq A$. Finally, in view of Corollary 5.1, we can conclude that $P \succ A$ and thus A is inadmissible. □

Our compelling result indicates that one only needs to choose P within $P \in \mathcal{C}_\gamma^a$, which simplifies the choices of P. We should briefly discuss how our results differ from staircase mechanisms, which is proposed by Kairouz et al. (2016b). Staircase mechanisms are similar to the admissible class $\mathcal{C}_\gamma^a$, but it admits rows with parity 1. Also, they shows that the optimal procedure is a staircase mechanism, if the utility measure is the sum of any sublinear functions of rows of P. Actually, these utility measures, such as mutual information or f-divergence, are merely functions of P, without considering the inferential goals. However, our result holds for any

inferential goals and any utility functions, not just for a particular class of utility measures.

Chapter 6

Optimality Results

Generally, the class $\mathcal{C}_\gamma^a$ contains infinitly many P, and sufficiency does not yield a best procedure. So, to find optimum designs and strategies we need additional criteria and measures of data utility. Examples of such criteria and some related results can be found in Agrawal et al. (2009), Kairouz et al. (2016b) and Duchi et al. (2018). However, for $k = 2$, it can be seen easily that if $P_{m\times 2}$ satisfies the condition C1 and has no proportional rows, then we must have $m = 2$. Moreover, $\mathcal{C}_\gamma^a$ consists of only two TPMs, which are also equivalent (by permutation). Thus, both are optimal procedures, one of which is reported below.

Proposition 6.1. *For binary X (i.e., $k = 2$), an optimal procedure at privacy level γ is given by $m = 2, p_{11} = p_{22} = \gamma(\gamma+1)^{-1}$ and $p_{12} = p_{21} = (\gamma+1)^{-1}$.*

For $k \geq 3$, there are infinitly many of admissible procedures, so choosing an optimal procedure from $\mathcal{C}_\gamma^a$ requires specific utility (or loss) functions. In the following section, we shall present one result in a common setting.

6.1 Maximization of Unchanged Probabilities

Frequently, the categories of the survey variable are used as possible response categories, i.e., $m = k$ and $d_i = c_i, i = 1, \ldots k$, and consequently $\mathcal{S}_Z = \mathcal{S}_X$. In such cases, a common desire is to retain the original values as much as possible while meeting the privacy requirement. One mathematical formulation of this idea is to choose $P_{k\times k}$ to maximize $\sum_i p_{ii}$, the trace of P, subject to $\eta(P) \leq \gamma$, where γ is specified. The optimal P for this objective is given below.

Theorem 6.1. *Suppose $P_{k\times k}$ is a TPM with $\eta(P) \leq \gamma$. Then,*
(a) $\sum_{i=1}^{k} p_{ii} \leq \frac{\gamma k}{\gamma+k-1}$ and
(b) P attains the upper bound in (a) if and only if $p_{ii} = \frac{\gamma}{\gamma+k-1}$ for all i and $p_{ij} = \frac{1}{\gamma+k-1}$ for all $i \neq j$.

Proof. Take any $P_{k\times k}$ satisfying $\eta(P) \leq \gamma$, which implies that $p_{ii} \leq \gamma p_{ij}$ for all $i \neq j$. For fixed i, summing over $j \neq i$ and then adding p_{ii} to both sides, we get

$$(\gamma - 1 + k)p_{ii} \leq \gamma \sum_{j=1}^{k} p_{ij} \quad \text{or} \quad p_{ii} \leq \frac{\gamma}{\gamma + k - 1} \sum_{j=1}^{k} p_{ij}.$$

Then, adding both sides of the last inequality over i, and using the fact that for each j, $\sum_{i=1}^{k} p_{ij} = 1$, we obtain the inequality in (a).

The "if" part of (b) is easy to verify. For the "only if" part, the chain of inequalities in the proof of part (a) shows that equality in (a) holds if and only if $p_{ij} = p_{ii}/\gamma$ for all $i \neq j$. This implies that $p_{ij} = a_i/\gamma$ for all $i \neq j$, where $a_1, \ldots, a_k$ denote the diagonal elements of P. Now, as each column of P adds to 1, i.e., $a_j + \frac{1}{\gamma}\sum_{i\neq j} a_i = 1$ we obtain:

$$a_j\big(1 - \frac{1}{\gamma}\big) + \frac{1}{\gamma}\sum_{i=1}^{k} a_i = 1 \quad \text{or} \quad a_j = \frac{\gamma}{1-\gamma}\Big[1 - \sum_{i=1}^{k} a_i\Big]$$

for all $j = 1, \ldots, k$. Thus, we must have $a_1 = \cdots = a_k$ and hence $p_{ii} = \frac{\gamma}{\gamma+k-1}$ for all i

and $p_{ij} = \frac{1}{\gamma+k-1}$ for all $i \neq j$, as each column of P must add to 1. □

Let P_1 denote the optimal TPM (for given k and γ) given above. Thus, the elements of P_1 are: $p_{ii} = \frac{\gamma}{\gamma+k-1}$ for all i and $p_{ij} = \frac{1}{\gamma+k-1}$ for all $i \neq j$. This P_1 has some attractive features and has received much attention. Note that P_1 is in $\mathcal{C}_\gamma^a$ and hence admissible. Agrawal et al. (2009) refer to P_1 as "the Gamma-Diagonal matrix" due to its structure; it has a common diagonal value and also a common off-diagonal value. They also proved an optimality property of P_1, in terms of lowest condition number, among all symmetric positive definite P with $\eta(P) \leq \gamma$. Kairouz et al. (2016b) refer to P_1 as "the randomized response mechanism" and view this as Warner's method for polychotomous variables. The true category does not change with certain probability p, and change to a different category randomly with probability $1 - p$. They also present certain mutual information optimality of P_1 when γ is large enough.

6.2 Minimax RR Strategies

The previous section focus only on the case $m = k$. But if $m \geq k$, such as the RAPPOR design, maximizing the trace of P is not applicable, as P is not square. Also, the estimation of π is more complicated than the case $m = k$, since the estimator is not obvious. In this section, we shall investigate optimum RR strategies that include both RR design and the corresponding estimator for general cases $m \geq k$ under linear unbiased estimation.

6.2.1 The Criterion

As we discussed in the earlier chapter, one should select reasonable criteria to derive optimality results. We shall develop a minimax criterion, using squared error loss and restricting to linear unbiased estimation. Specifically, we shall consider only unbiased

estimators that are linear in T, or equivalently linear in $\hat{\lambda} = T/n$. A linear estimator $\hat{\pi} = L\hat{\lambda}$ is unbiased, i.e., $E(L\hat{\lambda}) = \pi$ or $LP\pi = \pi$ for all π, if and only if $LP = I$, which can hold only if $r(P) = k$ (and $m \geq k$). Conversely, if $r(P_{m\times k}) = k$, there exists $L_{k\times m}$ such that $LP = I$. Thus, we must restrict our attention to RR designs P with $r(P) = k$. Adopting squared error loss, we define the risk function of a linear unbiased RR strategy (P, L) as

$$\begin{aligned}\mathbf{R}(P, L; \pi) &= nE_{P,\pi}\left[\|L\hat{\lambda} - \pi\|^2\right] \\ &= n[tr(\mathbf{V}_{P,\pi}(L\hat{\lambda}))] = tr(L(D_\lambda - \lambda\lambda')L') \\ &= tr(LD_\lambda L') - \sum_{i=1}^{k} \pi_i^2, \end{aligned} \tag{6.1}$$

where $\lambda = P\pi$, for a vector $v = (v_1, \ldots, v_k)'$, D_v denotes the diagonal matrix with diagonal elements $v_1, \ldots, v_k$, the expectation is with respect to both sampling and randomization and the multiplier n is used for notational convenience.

Note that the conclusions of Theorem 5.5 hold also under the added restriction $r(P) = k$. If $r(P) = k$ and P is inadmissible, it follows easily that there exists $A \in \mathcal{C}_\gamma^a$ such that $r(A) = k$ and $P = CA$ for some TPM C, i.e., there exists a more informative design A with $r(A) = k$. Thus, we shall restrict our attention to $\mathcal{C}_\gamma^1$, the class of all admissible procedures P with $r(P) = k$. To be precise, $\mathcal{C}_\gamma^1$ consists of all $P_{m\times k}, m \geq k$, satisfying the conditions

C1: $\eta_i(P) = \gamma$ for $i = 1, \ldots, m$.

C2: Each row of P contains two distinct values.

C3: $r(P) = k$

C4: No two rows of P are proportional to each other.

Note that C1 implies that all elements of P must be positive. A natural goal is

to find $P \in \mathcal{C}_\gamma^1$ and an L such that the risk in (6.1) is minimum. First, consider minimizing (6.1) with respect to L, for given P. If P is square and nonsingular, then $P^{-1}\hat{\lambda}$ is the unique linear unbiased estimator of π (see Chaudhuri and Mukerjee, 1988), hence the optimal L is P^{-1}. For any $P_{m\times k} \in \mathcal{C}_\gamma^1$ with $m > k$, unbiased L is not unique and the following result gives locally optimal linear unbiased estimators.

Proposition 6.2. *For any given $P \in \mathcal{C}_\gamma^1$ and π,*

$$\boldsymbol{R}(P, L; \pi) \geq tr(P' D_\lambda^{-1} P)^{-1} - \sum_{i=1}^{k} \pi_i^2 \tag{6.2}$$

for all L such that $LP = I$, and the lower bound is attained when

$$L = (P' D_\lambda^{-1} P)^{-1} P' D_\lambda^{-1} = L_*, \ \textit{say.} \tag{6.3}$$

Proof. For given $P \in \mathcal{C}_\gamma^1$, take any L such that $LP = I$. In view of (6.1), it suffices to show that $tr(LD_\lambda L') \geq tr(P' D_\lambda^{-1} P)^{-1}$. Let $U = D_\lambda^{-1/2} P$ and $U^- = LD_\lambda^{1/2}$. Then, $U^- U = I$ and $U'U = P' D_\lambda^{-1} P$, and thus

$$\begin{aligned} LD_\lambda L' = U^-(U^-)' &= \left(U^- - (U'U)^{-1}U'\right)\left(U^- - (U'U)^{-1}U'\right)' + (U'U)^{-1} \\ &= \left(U^- - (U'U)^{-1}U'\right)\left(U^- - (U'U)^{-1}U'\right)' + (P' D_\lambda^{-1} P)^{-1}. \end{aligned} \tag{6.4}$$

Now, (6.4) shows that $LD_\lambda L' - (P' D_\lambda^{-1} P)^{-1}$ is non-negative definite and thus $tr(LD_\lambda L') \geq tr(P' D_\lambda^{-1} P)^{-1}$. Moreover, the equality holds if and only if

$$U^- - (U'U)^{-1}U' = 0 \Leftrightarrow L = (P' D_\lambda^{-1} P)^{-1} P' D_\lambda^{-1}$$

. □

Remark 6.1. *Clearly, L_* depends on P and π, but for notational simplicity we do not write that explicitly. The proof of Proposition 6.2 shows that $L_*\hat{\lambda}$ is locally best*

also under D- and E-optimality criteria. Also, (6.2) holds more generally for any P (not limited to $\mathcal{C}_\gamma^1$) and π such that $r(P) = k$ and all elements of $P\pi$ are positive.

The optimum L in (6.3) depends on π, unless P is square and non-singular. So, a uniformly minimum risk estimator among all linear estimators does not exist. This also shows that a uniformly minimum risk RR strategy (P, L) does not exist. As an alternative, we shall use minimaxity to find an optimality RR strategy. Specifically, we shall try to find a strategy $(P_0,\ L_0)$ such that $P_0 \in \mathcal{C}_\gamma^1$, $L_0 P_0 = I$ and

$$\inf_{P \in \mathcal{C}_\gamma^1} \inf_{L} \sup_{\pi} \mathbf{R}(P, L; \pi) = \sup_{\pi} \mathbf{R}(P_0, L_0; \pi). \tag{6.5}$$

In (6.5), for brevity, we do not show the requirement $LP = I$ explicitly. The left side of (6.5) is the minimax value. Duchi et al. (2018) considered a similar approach and derived some asymptotic results for a general class of loss functions. In particular, they obtained minimax rates of convergence for several estimation problems and loss function. In contrast, we shall derive exact minimax procedures, but under linearity, unbiasedness and squared error loss.

6.2.2 Derivation of Minimax Strategies

The main result and the corresponding concept are stated in Theorem 6.4 and Definition 6.1. This subsection gives a full detailed derivation of the main result.

To find a minimax strategy (P, L), we shall first find (P_0, L_0) that minimizes (6.1) at $\pi = (1/k, \ldots, 1/k) = \pi_u$, say, i.e.,

$$\inf_{P \in \mathcal{C}_\gamma^1} \inf_{L} \mathbf{R}(P, L; \pi_u) = \mathbf{R}(P_0, L_0; \pi_u). \tag{6.6}$$

Then, we shall prove that the solution (P_0, L_0) satisfies (6.5). In a sense, the degen-

erate distribution at π_u is least favorable. It is reasonable as the risk from sampling is the largest when $\pi = \pi_u$.

In view of Proposition 6.2, solving (6.6) reduces to finding $P \in \mathcal{C}_\gamma^1$ such that $tr(P'D_\lambda^{-1}P)^{-1}$ is minimum, where $\lambda = P\pi_u$. Note that $P'D_\lambda^{-1}P$ is a symmetric positive definite matrix, and let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_k$ denote its eigenvalues. Then,

$$tr(P'D_\lambda^{-1}P)^{-1} = \sum_{i=1}^{k} \frac{1}{\alpha_i}, \tag{6.7}$$

which suggests that we should try to make α_i's large as we search for an optimum P, viz. a minimizer of (6.7). However, α_i's cannot be arbitrarily large, as they must satisfy certain restrictions. Recall that $\lambda = P\pi_u$ and so $\lambda = P\pi_u = k^{-1}P\mathbf{1}_k$, where $\mathbf{1}_k$ denotes the vector of dimension k all of whose components are 1. So, $P'D_\lambda^{-1}P\mathbf{1}_k = kP'D_{(P\pi_u)}^{-1}P\pi_u = kP'\mathbf{1}_m = k\mathbf{1}_k$. Thus, when $\pi = \pi_u$, $(P'D_\lambda^{-1}P)/k$ is a stochastic matrix and hence the dominant eigenvalue of $P'D_\lambda^{-1}P$ is $\alpha_k = k$ (Lax, 2007, p.241).

Moreover, $\sum_{i=1}^{k} \alpha_i = tr(P'D_\lambda^{-1}P)$ has a tight upper bound, as the following lemma shows. Recall that conditions C1 and C2 imply that for any $P \in \mathcal{C}_\gamma^1$, each row of P contains two distinct values and the ratio of the largest to smallest values is γ (> 1). Subsequently, we shall refer to the smaller (larger) of the two values as the small (large) value.

Lemma 6.1. *For given $\gamma > 1$ and $k \geq 2$, let*

$$f(x) = \frac{k^2(x\gamma^2 + k - x)}{(x\gamma + k - x)^2}, \quad x \geq 0, \tag{6.8}$$

and

$$q = \begin{cases} \lfloor \dfrac{k}{1+\gamma} \rfloor, & \text{if } f(\lfloor \dfrac{k}{1+\gamma} \rfloor) \geq f(\lceil \dfrac{k}{1+\gamma} \rceil) \text{ and } \lfloor \dfrac{k}{1+\gamma} \rfloor \geq 1 \\ \lceil \dfrac{k}{1+\gamma} \rceil, & \text{otherwise.} \end{cases} \tag{6.9}$$

Then, for all $P \in \mathcal{C}_\gamma^1$,

$$tr(P' D_\lambda^{-1} P) \leq f(q), \tag{6.10}$$

and the equality holds if each row of P *contains exactly* q *large values.*

Proof. Take any $P_{m\times k} \in \mathcal{C}_\gamma^1$ and let $P\pi_u = \lambda = (\lambda_1, \ldots, \lambda_m)'$. Then,

$$tr(P' D_\lambda^{-1} P) = tr(D_\lambda^{-1} PP') = \sum_{i=1}^{m} \frac{1}{\lambda_i} \sum_{j=1}^{k} p_{ij}^2. \tag{6.11}$$

Recalling that each row of P contains two distinct values, for $i = 1, \ldots, m$, let s_i denote the 'small' value in the ith row of P, and so the 'large' value is γs_i. Also, suppose that the ith row contains q_i large values and $(k - q_i)$ small values, with $1 \leq q_i \leq k-1$. Note that $\lambda_i = \frac{1}{k}\sum_{j=1}^{k} p_{ij}$, as π_u is uniform. This implies that $q_i\gamma s_i + (k - q_i)s_i = k\lambda_i$ or $s_i = \dfrac{\lambda_i k}{q_i\gamma + k - q_i}$. Thus,

$$\sum_{j=1}^{k} p_{ij}^2 = q_i(\gamma s_i)^2 + (k - q_i)s_i^2 = \lambda_i^2 \frac{k^2(q_i\gamma^2 + k - q_i)}{(q_i\gamma + k - q_i)^2} = \lambda_i^2 f(q_i). \tag{6.12}$$

Combining (6.11) and (6.12), we get

$$tr(P' D_\lambda^{-1} P) = \sum_{i=1}^{m} \lambda_i f(q_i). \tag{6.13}$$

Taking derivative, it can be seen that as x increases, $f(x)$ first increases and then decreases, reaching its maximum at $x = \dfrac{k}{1+\gamma}$. Then, for $x \in \{1, \ldots, k-1\}$, it can be seen that $f(x)$ is maximized at q, as defined in (6.9). So,

$$\sum_{i=1}^{m} \lambda_i f(q_i) \leq (\sum_{i=1}^{m} \lambda_i) f(q) = f(q),$$

which establishes (6.10). Clearly, the upper bound is attained if all rows of P contain

exactly q large values. □

Remark 6.2. *If $f(q) \neq f(q+1)$, the upper bound is attained if and only if $q_i = q$, irrespective of the small value s_i in each row.*

Now, minimizing (6.7) reduces to minimizing $\sum_{i=1}^{k-1}(1/\alpha_i)$, subject to $\sum_{i=1}^{k-1}\alpha_i \leq f(q) - k$, and $\alpha_i > 0, i = 1, \ldots, k-1$, as $\alpha_k = k$. It can be seen easily that $\sum_{i=1}^{k-1}(1/\alpha_i)$ is a strictly Schur-convex function on $\Delta = \{(\alpha_1, \ldots, \alpha_{k-1}) : \sum_{i=1}^{k-1}\alpha_i = f(q) - k, \alpha_i > 0, i = 1, \ldots k-1\}$. So, $\sum_{i=1}^{k-1}(1/\alpha_i)$ is minimized over Δ if and only if $\alpha_i = [f(q) - k]/(k-1), i = 1, \ldots, k-1$ (Marshall et al., 2011). Now, the following conclusion can be reached readily.

Lemma 6.2. *A lower bound for $tr(P'D_\lambda^{-1}P)^{-1}$ is $\dfrac{(k-1)^2}{f(q)-k} + \dfrac{1}{k}$, and it is attained if and only if the eigenvalues of $P'D_\lambda^{-1}P$ are*

$$\alpha_k = k \quad \textit{and} \quad \alpha_i = \frac{f(q)-k}{k-1} \textit{ for } 1 \leq i \leq k-1. \tag{6.14}$$

As $P'D_\lambda^{-1}P\mathbf{1}_k = k\mathbf{1}_k$ (observed earlier), the eigenvector of $P'D_\lambda^{-1}P$ corresponding to the eigenvalue $k \ (= \alpha_k)$ is $\mathbf{1}_k$. Using this and the spectral decomposition of $P'D_\lambda^{-1}P$ we get the following alternative perspective of Lemma 6.2.

Lemma 6.3. *The lower bound in Lemma 6.2 is attained if and only if*

$$P'D_\lambda^{-1}P = a_q I + b_q \boldsymbol{1}_k \boldsymbol{1}_k', \tag{6.15}$$

where $a_q = \dfrac{f(q)-k}{k-1}$ and $b_q = 1 - \dfrac{a_q}{k}$.

Proof. By Lemma 6.2 and the fact that eigenvalues of a matrix is unique, it suffices

to show that (6.14) implies (6.15). By spectral decomposition,

$$\begin{aligned} P'D_\lambda^{-1}P &= \alpha_1 \sum_{i=1}^{k-1} e_i e_i' + \alpha_k e_k e_k' \\ &= \alpha_1 \sum_{i=1}^{k} e_i e_i' + (\alpha_k - \alpha_1) e_k e_k' \\ &= a_q I + (k - a_q)\frac{1}{\sqrt{k}}\mathbf{1}_k \frac{1}{\sqrt{k}}\mathbf{1}_k' \\ &= a_q I + b_q \mathbf{1}_k \mathbf{1}_k'. \end{aligned}$$

□

Next, we need to explore existence of $P \in \mathcal{C}_\gamma^1$ satisfying (6.15) and find one, if it exists. For simplicity, consider the situation where $f(q) \neq f(q+1)$. Then, recall that to attain the lower bound in Lemma 6.2, each row of P must contain exactly q large values and $(k-q)$ small values. The positions for q large values can be chosen in $\binom{k}{q}$ ways. It is reasonable to explore RR designs which utilize all possible arrangements of large (and small) values. Wang et al. (2016) and Ye and Berg (2018) studied the following class of RR designs, requiring additionally all small values to be equal.

Definition 6.1. *For any integer t with $1 \leq t \leq k-1$, an RR design $P \in \mathcal{C}_\gamma^1$ is called a t-subset design if*

(i) P has $\binom{k}{t}$ rows,

(ii) P contain exactly 2 distinct values; a large value and a small value and

(iii) each row contains exactly t large and $(k-t)$ small values.

We shall denote a t-subset design by P_t and let $m_t = \binom{k}{t}$. Since proportional rows are not allowed, for each t, P_t is unique up to row permutation. We shall see in the sequel that our minimax RR design is P_q, i.e., P_t with $t = q$, with q as defined in (6.9). To review some basic properties of P_t, denote its small value by s_t; so its large value is γs_t. Clearly, P_t has m_t rows. It can be seen that each column of P_t contains

exactly $\binom{k-1}{t-1} = (\frac{t}{k})m_t$ large values and $\binom{k-1}{t} = (\frac{k-t}{k})m_t$ small values. From this, we can find that

$$s_t = \frac{k}{m_t(t\gamma + k - t)} \tag{6.16}$$

and that the sum of each row is k/m_t.

Remark 6.3. *A t-subset design can be constructed as follows. Consider all $\binom{k}{t}$ subsets of size t of $\mathcal{S}_X = \{c_1, \ldots, c_k\}$, the sample space of X. Call the subsets $d_1, \ldots, d_{m_t}$, where $m_t = \binom{k}{t}$. Thus, each d_i contains a subset of the t categories in $\mathcal{S}_X$. Then, let $p_{ij} = \gamma s_t$ if $c_j \in d_i$, otherwise $p_{ij} = s_t$. Figure 5 illustrates this for $k = 4$, $\gamma = 2$ and $t = 1, 2, 3$, where the columns represent c_1, c_2, c_3 and c_4, respectively, and the subsets are shown to the right of each TPM.*

$$\frac{1}{5}\begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix} \begin{matrix} \{c_1\} \\ \{c_2\} \\ \{c_3\} \\ \{c_4\} \end{matrix} \qquad t = 1$$

$$\frac{1}{9}\begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & 1 & 2 & 1 \\ 2 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 \end{bmatrix} \begin{matrix} \{c_1, c_2\} \\ \{c_1, c_3\} \\ \{c_1, c_4\} \\ \{c_2, c_3\} \\ \{c_2, c_4\} \\ \{c_3, c_4\} \end{matrix} \qquad t = 2$$

$$\frac{1}{7}\begin{bmatrix} 2 & 2 & 2 & 1 \\ 2 & 2 & 1 & 2 \\ 2 & 1 & 2 & 2 \\ 1 & 2 & 2 & 2 \end{bmatrix} \begin{matrix} \{c_1, c_2, c_3\} \\ \{c_1, c_2, c_4\} \\ \{c_1, c_3, c_4\} \\ \{c_2, c_3, c_4\} \end{matrix} \qquad t = 3$$

Figure 5: t-subset designs

Next, we present some additional properties of t-subset designs, for $1 \leq t \leq k-1$. As each row of P_t adds to k/m_t, it follows that

$$P_t\pi_u = P_t\Big(\frac{1}{k}\mathbf{1}_k\Big) = \frac{1}{m_t}\mathbf{1}_{m_t} \quad \Longrightarrow D_{(P_t\pi_u)} = \frac{1}{m_t}I. \tag{6.17}$$

By (6.17), we get $P_t' D^{-1}_{(P_t\pi_u)} P_t = m_t P_t' P_t$. As noted earlier, each column of P_t has $\binom{k-1}{t-1}$ values that are γs_t and the rest are s_t. Any two columns of P_t have the

large value (γs_t) in $\binom{k-2}{t-2}$ common rows, the small value s_t in $\binom{k-2}{t}$ common rows and the remaining rows contain one large and one small value. Using these and routine algebra we can verify that

$$P_t' D_{(P_t\pi_u)}^{-1} P_t = m_t P_t' P_t = a_t I_k + b_t \mathbf{1}_k \mathbf{1}_k', \tag{6.18}$$

where

$$a_t = \frac{f(t) - k}{k - 1} \quad \text{and} \quad b_t = 1 - \frac{a_t}{k}. \tag{6.19}$$

By (6.18), Lemma 6.3 and previous observations we obtain:

Theorem 6.2. *For given k and γ, let q and P_q be as defined earlier and let L_q denote the optimal L in (6.3) for $P = P_q$ and $\pi = \pi_u$. Then,*

$$\inf_{P \in \mathcal{C}_\gamma^1} \inf_L \boldsymbol{R}(P, L; \pi_u) = \boldsymbol{R}(P_q, L_q; \pi_u) = \frac{(k-1)^2}{f(q) - k}. \tag{6.20}$$

This result tells us that (P_q, L_q) is a locally (at $\pi = \pi_u$) optimal RR strategy at privacy level γ. To investigate its properties more generally and to solve the minimax problem of (6.5), we next describe some additional properties of t-subset designs. For a given P_t, the locally (at π_u) optimum L, to be denoted L_t, can be obtained by using (6.17) and (6.18) in (6.3). Specifically,

$$\begin{aligned} L_t &= (P_t' D_{(P_t\pi_u)}^{-1} P_t)^{-1} P_t' D_{(P_t\pi_u)}^{-1} \\ &= m_t (a_t I + b_t \mathbf{1}_k \mathbf{1}_k')^{-1} P_t' \\ &= m_t (a_t^{-1} I - d_t \mathbf{1}_k \mathbf{1}_k') P_t' \\ &= a_t^{-1} (m_t P_t' - b_t \mathbf{1}_k \mathbf{1}_{m_t}'), \end{aligned} \tag{6.21}$$

where $d_t = \dfrac{b_t}{a_t(a_t + k b_t)} = \dfrac{b_t}{k a_t}$ and the last " = " follows from $\mathbf{1}_k' P_t' = (k/m_t)\mathbf{1}_{m_t}'$.

The strategy (P_t, L_t) has an interesting property, as the following lemma shows.

Lemma 6.4. *For any t-subset design P_t and L_t as in (6.21), $tr(L_t D_\lambda L_t')$ is a constant, independent of π, where $\lambda = P_t\pi$.*

Proof. Let $((f_{ij})) = F_{m_t \times m_t} = L_t' L_t$. Then, f_{ii} is the squared length of the ith row of L_t'. Using (6.21) and considering the structure of P_t, we see that in each row of L_t', exactly t values are $a_t^{-1}(m_t\gamma s_t - b_t)$ and $(k-t)$ are $a_t^{-1}(m_t s_t - b_t)$. So, all rows are have the same length and consequently, $f_{11} = \cdots = f_{m_t m_t} = f_0, say$. Now,

$$tr(L_t D_\lambda L_t') = tr(D_\lambda L_t' L_t) = \sum_{i=1}^{m_t} \lambda_i f_{ii} = f_0 \sum_{i=1}^{m_t} \lambda_i = f_0, \tag{6.22}$$

which is independent of π, as the lemma asserts. □

The next two theorems are the main results in this chapter, which give a minimax estimator for given P_t and a minimax strategy satisfying (6.5).

Theorem 6.3. *For any given t-subset procedure P_t, a linear unbiased minimax estimator of π is $L_t\hat{\lambda}$, where L_t is as given by (6.21).*

Proof. First, we note that

$$\begin{aligned}
\sup_\pi \mathbf{R}(P_t, L_t; \pi) &= \sup_\pi \left[tr(L_t D_\lambda L_t') - \sum_{i=1}^{k} \pi_i^2 \right] \\
&= tr(L_t D_\lambda L_t') - \inf_\pi \sum_{i=1}^{k} \pi_i^2 \\
&= \mathbf{R}(P_t, L_t; \pi_u),
\end{aligned} \tag{6.23}$$

as $tr(L_t D_\lambda L_t')$ is independent of π by Lemma 6.4 and $\sum \pi_i^2$ is minimum when $\pi = \pi_u$. Consider any L such that $LP_t = I$. Then, by (6.23) and Proposition 6.2,

$$\sup_\pi \mathbf{R}(P_t, L_t; \pi) = \mathbf{R}(P_t, L_t; \pi_u) \le \mathbf{R}(P_t, L; \pi_u) \le \sup_\pi \mathbf{R}(P_t, L; \pi), \tag{6.24}$$

which proves the theorem. □

The next theorem is the main result in this chapter, which gives a minimax strategy satisfying (6.5).

Theorem 6.4. *For a given privacy level γ, a minimax strategy that solves (6.5) is (P_q, L_q) and*

$$\inf_{P\in\mathcal{C}_\gamma^1}\inf_{L}\sup_{\pi}\boldsymbol{R}(P,L;\pi)=\sup_{\pi}\boldsymbol{R}(P_q,L_q;\pi)=\frac{(k-1)^2}{f(q)-k}, \tag{6.25}$$

where $f(.), q, P_q$ and L_q are as defined earlier.

Proof. By (6.23) and Theorem 6.2 we get

$$\begin{aligned}\sup_{\pi}\mathbf{R}(P_q,L_q;\pi)&=\mathbf{R}(P_q,L_q;\pi_u)\\&=\inf_{P\in\mathcal{C}_\gamma^1}\inf_{L}\mathbf{R}(P,L;\pi_u)\\&\le\inf_{P\in\mathcal{C}_\gamma^1}\inf_{L}\sup_{\pi}\mathbf{R}(P,L;\pi)\\&\le\sup_{\pi}\mathbf{R}(P_q,L_q;\pi).\end{aligned}$$

Hence, the first "=" in (6.25) holds and the second follows readily from (6.20). □

Clearly, the minimax risk in (6.25) is a function of k and γ. From (6.9), $q\approx\dfrac{k}{1+\gamma}$, and hence it can be seen that

$$\text{minimax risk}=\frac{(k-1)^2}{f(q)-k}\approx 4\frac{(k-1)^2}{k}\frac{\gamma}{(\gamma-1)^2}.$$

Figure 6 exhibits the dependence of the minimax risk on k and γ. Approximately, the risk is proportional to k and inversely proportional to γ.

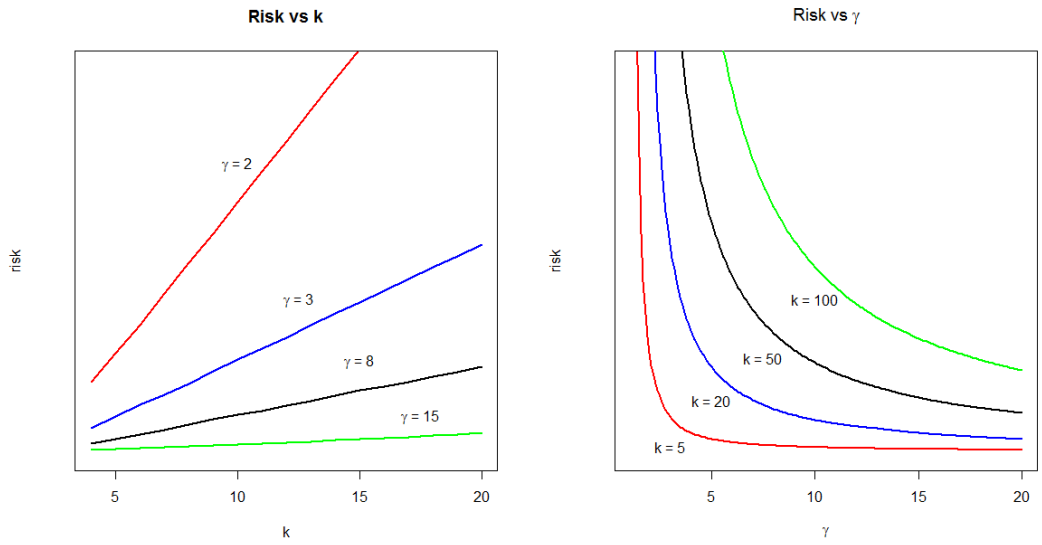

Figure 6: The dependence of minimax risk on k and γ

We should briefly discuss how our results differ from those of Wang et al. (2016) and Ye and Berg (2018). Ye and Berg (2018) also introduced t-subset designs and arrived at P_q and the corresponding moment estimator as optimal strategy, but by considering minimax rate of convergence, similar to Duchi et al. (2018). Wang et al. (2016) considered finding an RR design P, subject to (3.7), that maximizes the mutual information between true and randomized responses under $\pi = \pi_u$ and proved that P_q is an optimal design. They also proposed a methods of moments estimator of π, which coincides with our minimax estimator, as we shall prove in the next chapter.

Chapter 7

Discussions of t-subset and RAPPOR designs

In this chapter, we discuss the implementation of t-subset designs. We give a simpler view of t-subset designs and a simplified form of the minimax estimator. We shall also discuss RAPPOR, the RR method used by Google and Apple. We relate RAPPOR to the mixture of t-subset designs and compare the minimax risk among empirical RAPPOR, minimax RAPPOR and the q-subset strategy. The notations used in this chapter are the same as that in Section 6.2.

7.1 Practical Aspects of t-subset Designs

The mathematical solution of the minimax problem derived in the preceding chapter is a bit abstract and hard to use. To construct and implement the minimax design P_q directly and calculate the estimator $L_q\hat{\lambda}$, following Section 6.2 results literally, we need to calculate q, define $d_1, \ldots, d_m$, compute P_q, apply randomization, obtain $\hat{\lambda}$ and calculate $L_q\hat{\lambda}$. That can be overly burdensome and time consuming because m_q

may easily be very large. Table 1 gives values of q and m_q for several combinations of k and γ. In each cell, the top number is the value of q and the bottom number is m_q. As an example, for $k = 20$ and $\gamma = 2$, we have $q = 7$ and $m_q = 77,520$, and so we shall need to create 77,520 response categories, randomize each true value among 77,520 categories (with very small probabilities) etc.

	γ					
k	1.1	1.5	2	5	10	20
4	2 6	2 6	1 4	1 4	1 4	1 4
6	3 20	2 15	2 15	1 6	1 6	1 6
10	5 252	4 210	3 120	2 45	1 10	1 10
20	10 184,756	8 125,970	7 77,520	3 1140	2 190	1 20

Table 1: values of q and m_t for several combinations of k and γ

Wang et al. (2016) and Ye and Berg (2018) described an alternative and simpler method for using t-subset designs. In the following, we review that approach and present some new results. Using indicator vectors, both original and perturbed data may be presented conveniently as $n \times k$ matrices. Recording each true category with a row vector $X = (X_1, \ldots, X_k)$ whose ith component is 1 if the true category is c_i and 0 otherwise, the unperturbed data from n units yields an $n \times k$ data matrix $\mathcal{D}_*$, with each row showing the true category of one unit. Perturbed data can also be organized as a matrix using the following scheme.

Take any t-subset design P_t with parity γ. Recall that P_t has $m_t = \binom{k}{t}$ rows and each row has t large values (γs_t) and $(k-t)$ small values (s_t). The output variable has m_t categories, which we labeled earlier $d_1, \ldots, d_{m_t}$ (arbitrarily) and attached those to the rows of P_t. The alternative scheme represents the output categories with the k dimensional row vectors that are obtained by replacing the large values by 1 and

small values by 0 in P_t. Specifically, for $i = 1, \ldots, m_t$, the response corresponding to the ith row of P_t is recorded as $(z_{i1}, \ldots, z_{ik})$, where z_{ij} is 1 if $p_{ij} = \gamma s_t$ and zero otherwise. This data representation scheme is the reverse of the construction method noted in Remark 6.3. Note that $\sum_j z_{ij} = t$ for all i and the possible responses are the indicator vectors for all subsets of $\{c_1, \ldots, c_k\}$ of size t. Denoting the randomized response with a vector $Z = (Z_1, \ldots, Z_k)$ and using one row for each respondent, the data from using P_t can be given as a matrix $\mathcal{D}$ of order $n \times k$.

Wang et al. (2016) and Ye and Berg (2018) gave the following algorithm for implementing P_t and generating a data matrix $\mathcal{D}$. For a true response $(x_1, \ldots, x_k)$ the algorithm generates a randomized response $(z_1, \ldots, z_k)$ as follows. Recall that only one of $x_1, \ldots, x_k$ is 1 and the rest are 0. Suppose $x_j = 1$. Then, first using a suitable binary experiment set $z_j = 1$ with probability

$$p = tm_t\gamma s_t/k,$$

else set $z_j = 0$. Next, if $z_j = 1$, randomly select $(t-1)$ of the remaining $(k-1)$ components of z and set those to 1. For $z_j = 0$, assign 1 to t of the remaining components of z, selected at random. In either case, all other components of z are 0. Then, it can be verified that

$$P(Z = (z_1, \ldots, z_k)|x_j = 1) = \begin{cases} \gamma s_t, & \text{if } z_j = 1 \\ s_t, & \text{if } z_j \neq 1 \end{cases}$$

and hence the algorithm implements P_t. Actually, the algorithm can be motivated and justified by facts that each column of P_t contains $\binom{k-1}{t-1}$ large values and $\binom{k-1}{t}$ small values and we have the following:

$$P(Z_i = 1|X_i = 1) = \binom{k-1}{t-1}\gamma s_t = \frac{t}{k}m_t\gamma s_t,$$

$$P(Z=(z_1,\ldots,z_k)|Z_i=1,X_i=1)=\begin{cases}1/\binom{k-1}{t-1}, & \text{if } z_i=1\\ 0, & \text{otherwise,}\end{cases}$$

$$P(Z=(z_1,\ldots,z_k)|Z_i=0,X_i=1)=\begin{cases}1/\binom{k-1}{t}, & \text{if } z_i=0\\ 0, & \text{otherwise.}\end{cases}$$

Both Wang et al. (2016) and Ye and Berg (2018) used method of moments for estimating π from the data matrix $\mathcal{D}$. Let $V'=(V_1,\ldots,V_k)$ denote the vector of column sums of $\mathcal{D}$. For $j=1,\ldots,k$, let n_j denote the original frequency of c_j. Then, it follows that

$$E(V_j|\mathcal{D}_*)=n_jp+\big(n-n_j\big)\Big[p\frac{t-1}{k-1}+(1-p)\frac{t}{k-1}\Big]$$

and unconditionally,

$$E(\frac{V_j}{n})=p\pi_j+\big(1-\pi_j\big)\Big[p\frac{t-1}{k-1}+(1-p)\frac{t}{k-1}\Big], \tag{7.1}$$

which is a linear function of π_j. Recall that $p=tm_t\gamma s_t/k$, $m_t=\binom{k}{t}$ and $s_t=k/[\binom{k}{t}(t\gamma+k-t)]$. So, $p=(t\gamma)/(t\gamma+k-t)$. Using this in (7.1) and standard algebra, one obtains the following method of moments estimator of π_j (for $j=1,\ldots,k$):

$$\tilde{\pi}_j=\frac{(k-1)(t\gamma+k-t)}{t(\gamma-1)(k-t)}\Big(\frac{V_j}{n}\Big)+\frac{1}{k}\Big[\frac{(1-k)(t\gamma+k-t)}{(\gamma-1)(k-t)}+1\Big]. \tag{7.2}$$

Note that the method of moments estimator $\tilde{\pi}$ of π requires only the column totals of the data matrix $\mathcal{D}$ and is very easy to calculate. It can also verified that $\tilde{\pi}$ is an unbiased estimator of π. Another interesting property of $\tilde{\pi}$ is the following:

Proposition 7.1. *The method of moments estimator $\tilde{\pi}$ is also a minimax linear unbiased estimator of π under the t-subset design P_t.*

Proof. We shall simplify our minimax estimator $\hat{\pi} = L_t\hat{\lambda}$ to prove this result.Using (6.21), we get

$$L_t\hat{\lambda} = a_t^{-1}m_tP_t'\hat{\lambda} - (a_t^{-1}b_t)\mathbf{1}_k = a_t^{-1}m_tP_t'\hat{\lambda} - (a_t^{-1} - k^{-1})\mathbf{1}_k.$$

So,

$$\hat{\pi}_j = a_t^{-1}m_t\sum_{i=1}^{m_t} p_{ij}\hat{\lambda}_i - (a_t^{-1} - k^{-1}). \tag{7.3}$$

Represent the response categories using indicator vectors $d_i = (z_{i1}, \ldots, z_{ik}), i = 1, \ldots, m_t$, as discussed above. Recall that $z_{ij} = 1$ if $p_{ij} = \gamma s_t$ and $z_{ij} = 0$ if $p_{ij} = s_t$. Let $B_j = \{d_i : z_{ij} = 1\}$. Then,

$$\sum_{i=1}^{m} P_{ij}\hat{\lambda}_i = \gamma s_t\sum_{i\in B_j}\hat{\lambda}_i + s_t\sum_{i\notin B_j}\hat{\lambda}_i = \gamma s_t\big(\frac{V_j}{n}\big) + s_t\Big(1 - \big(\frac{V_j}{n}\big)\Big)$$

and (7.3) reduces to

$$\hat{\pi}_j = a_t^{-1}(\gamma - 1)m_ts_t\Big(\frac{V_j}{n}\Big) + a_t^{-1}(m_ts_t - 1) + k^{-1}. \tag{7.4}$$

Next, we can verify the following identities:

$$m_ts_t = \frac{k}{t\gamma + k - t}, \quad m_ts_t - 1 = \frac{t(1-\gamma)}{t\gamma + k - t} \quad \text{and} \quad a_t = \frac{kt(\gamma-1)^2(k-t)}{(k-1)(t\gamma + k - t)^2}.$$

Using these and routine algebra (7.4) can be reduced to (7.2). □

In view of the preceding result and (7.2), the minimax estimator π under P_t is $\hat{\pi} = c(\frac{V}{n}) + d$, where c and d are evident from (7.2). Note that V is the sum of n independent realizations of the response vector $Z = (Z_1, \ldots, Z_k)$. So, the variance-covariance matrix of $\hat{\pi}$ is $V(\hat{\pi}) = (c^2/n)\Sigma$, where $\Sigma = ((\sigma_{ij})) = V(Z)$. Since each Z_i's are binary variables, $\sigma_{jj} = P(Z_j = 1)[1 - P(Z_j = 1)]$ and $\sigma_{ij} = P(Z_i = 1, Z_j =$

$1) - P(Z_i = 1)P(Z_j = 1)$ for $i \neq j$. Moreover, the right side of (7.1) is $P(Z_j = 1)$ and simplifying it further we get

$$\begin{aligned} P(Z_j = 1) &= \Big[\frac{t(\gamma m_t s_t - 1)}{k-1}\Big]\pi_j + \frac{t(k - \gamma m_t s_t)}{k(k-1)} \\ &= \frac{t}{(k-1)(t\gamma + k - t)}\Big[(\gamma - 1)(k-t)\pi_j + \{t(\gamma-1) + k - t\}\Big]. \end{aligned}$$

For $t = 1$, $P(Z_i = 1, Z_j = 1) = 0$. For $t \geq 3$, using simpler algorithm for implementing P_t, discussed above, and letting $p = (t\gamma m_t s_t)/k$, we get

$$\begin{aligned} P(Z_i = 1, Z_j = 1) =& \sum_{r=1}^{k} \pi_r P(Z_i = 1, Z_j = 1 | X_r = 1) \\ =& (\pi_i + \pi_j) p \Big[\binom{k-2}{t-2} \div \binom{k-1}{t-1}\Big] \\ &+ (1 - \pi_i - \pi_j)\Big[p\Big\{\binom{k-3}{t-3} \div \binom{k-1}{t-1}\Big\} \\ &+ (1-p)\Big\{\binom{k-3}{t-2} \div \binom{k-1}{t}\Big\}\Big] \\ =& (\pi_i + \pi_j) p \frac{t-1}{k-1} + (1 - \pi_i - \pi_j)\Big[p\frac{(t-1)(t-2)}{(k-1)(k-2)} \\ &+ (1-p)\frac{t(t-1)}{(k-1)(k-2)}\Big] \\ =& \frac{t(t-1)\Big[(k-t)(\gamma-1)(\pi_i + \pi_j) + (t\gamma - 2\gamma + k - t)\Big]}{(k-1)(k-2)(t\gamma + k - t)}. \end{aligned} \tag{7.5}$$

Actually, (7.5) holds for all $1 \leq t \leq k-1$. For $t = 2$, the above derivation remains valid if $\binom{k-3}{t-3}$ is interpreted as 0.

7.2 Mixture of t-subset Designs

This section is motivated by the RAPPOR (randomized aggregatable privacy preserving ordinal response) algorithm proposed by Erlingsson et al. (2014). It is an RR

procedure and has been further discussed by Kairouz et al. (2016a), Fanti et al. (2016), Wang et al. (2017), Ye and Berg (2018) and others. Quite importantly, Google and Apple use RAPPOR for privacy protection. Basic RAPPOR is directly relevant to our context and it works as follows. As in Section 7.1, it represents the true category with an indicator vector $X = (X_1, \ldots, X_k)$. Then, it produces a perturbed output $Z = (Z_1, \ldots, Z_k)$ by changing each component of X independently with probability $p = 1/(\sqrt{\gamma}+1)$. So, the output space has 2^k elements. As we explain next, RAPPOR is a mixture of t-subset designs, with $t = 0, 1, \ldots, k$.

Consider RAPPOR perturbation and let $T = \sum_{j=1}^{k} Z_j$ denote the number of 1's in a randomized response $(Z_1, \ldots, Z_k)$. Then, for any $0 \le t \le k$ and $1 \le j \le k$,

$$\begin{aligned} P(T = t|X_j = 1) &= P(T = t, Z_j = 1|X_j = 1) + P(T = t, Z_j = 0|X_j = 1) \\ &= \binom{k-1}{t-1} p^{t-1}(1-p)^{k-t+1} + \binom{k-1}{t} p^{t+1}(1-p)^{k-t-1}. \end{aligned} \tag{7.6}$$

Since (7.6) is independent of j, it is also the unconditional probability $P(T = t)$, which we shall denote by p_t. Also,

$$P(Z = z, T = t|X_j = 1) = \begin{cases} p^{t-1}(1-p)^{k-t+1}, & \text{if } z_j = 1, \sum z_i = t, \\ p^{t+1}(1-p)^{k-t-1}, & \text{if } z_j = 0, \sum z_i = t. \end{cases}$$

Recall that $p = 1/(\sqrt{\gamma}+1)$ and so $\gamma = [(1-p)/p]^2$. Using this and the above, conditionally on $T = t$ and X we have

$$P(Z = z|T = t, X_j = 1) = \begin{cases} 1/[\binom{k-1}{t-1} + \binom{k-1}{t}\gamma^{-1}] = \gamma s_t, & \text{if } z_j = 1, \sum z_i = t, \\ 1/[\binom{k-1}{t-1}\gamma + \binom{k-1}{t-1}] = s_t, & \text{if } z_j = 0, \sum z_i = t, \end{cases}$$

which are the transition probabilities of the t-subset design with parity γ.

From the preceding observations it follows that RAPPOR perturbation is equiva-

lent to a two step procedure: First draw a value t from $\{0, 1, \ldots, k\}$ with probabilities $p_0, p_1, \ldots, p_k$ and then apply the t-subset design with parity γ. Thus, RAPPOR is a mixture of t-subset designs and its TPM is $P = \left[p_0 P_0' \mid p_1 P_1' \mid \ldots \mid p_k P_k'\right]'$, where P_t is the TPM of the t-subset design, for $t = 0, 1, \ldots, k$.

Note that Theorem 5.5 implies that the basic RAPPOR design is inadmissible, as the two rows of its TPM corresponding to $t = 0$ and $t = k$ have parity 1 (i.e., each row contains a common value). The two associated outputs, i.e., $Z = (0, 0, \ldots, 0)$ and $Z = (1, 1, \ldots, 1)$, give no information about the true category and hence about π. Effectively, RAPPOR throws away the units that yield those two responses. This wastage is minimal for large k, where both p_0 and p_k are small. But, for small k, the loss can be substantial. We can remove those two rows and normalize the TPM to get an admissible design.

Motivated by the preceding discussion, we shall next explore properties of mixtures of t-subset designs, with $t = 1, \ldots, k-1$. The TPM of such a design is a partitioned matrix

$$P_M = \left[w_1 P_1' \mid w_2 P_2' \mid \ldots \mid w_{k-1} P_{k-1}'\right]', \tag{7.7}$$

where $w_j \geq 0$ are the mixing probabilities and $\sum w_j = 1$. Naturally, if $w_j = 0$ for some j, the corresponding rows should be omitted. We may conveniently view P_M as a two-step procedure: first select a value t from $\{1, \ldots, k-1\}$ with probabilities $w_1, \ldots, w_{k-1}$ and then apply the t-subset design. Note that $D_\lambda = diag(P_M \pi)$ is a block diagonal matrix

$$D_\lambda = diag(D_\lambda^{(1)}, D_\lambda^{(2)}, \ldots, D_\lambda^{(k-1)}),$$

where $D_\lambda^{(t)} = diag(w_t P_t \pi)$ for $t = 1, \ldots k-1$.

We can derive the minimax linear unbiased estimator of π under P_M, using ar-

guments similar to those used in Section 6.2. The uniform distribution $\pi = \pi_u$ turns out to be a least favorable distribution in this case too. When $\pi = \pi_u$, $D_\lambda^{-1} = diag(\frac{m_1}{w_1} I_{m_1}, \frac{m_2}{w_2} I_{m_2}, \ldots, \frac{m_{k-1}}{w_{k-1}} I_{m_{k-1}})$ with $m_t = \binom{k}{t}$, and by Proposition 6.2, the locally best unbiased estimator of π is $\hat{\pi} = L_M \hat{\lambda}$, where

$$L_M = \Big(\sum_{t=1}^{k-1} w_t P_t' (D_\lambda^{(t)})^{-1} w_t P_t \Big)^{-1} P_M' D_\lambda^{-1} \tag{7.8}$$

$$= \Big((\sum_{t=1}^{k-1} w_t a_t) I_k + (\sum_{t=1}^{k-1} w_t b_t) \mathbf{1}_k \mathbf{1}_k' \Big)^{-1} P_M' D_\lambda^{-1} \tag{7.9}$$

$$= \Big(a_* I_k + b_* \mathbf{1}_k \mathbf{1}_k' \Big)^{-1} P_M' D_\lambda^{-1}$$

$$= \Big(a_*^{-1} I_k - \frac{b_*}{k a_*} \mathbf{1}_k \mathbf{1}_k' \Big) P_M' D_\lambda^{-1}, \tag{7.10}$$

$a_* = \sum_{t=1}^{k-1} w_t a_t$ and $b_* = \sum_{t=1}^{k-1} w_t b_t$ (and a_t and b_t are as defined in (6.19)). Moreover, $P_M' D_\lambda^{-1} = [m_1 P_1' \mid m_2 P_2' \mid \ldots \mid m_{k-1} P_{k-1}']$, and so

$$L_M = [L_1^* \mid L_2^* \mid \ldots \mid L_{k-1}^*], \text{ with } L_t^* = a_*^{-1}(m_t P_t' - b_* \mathbf{1}_k \mathbf{1}_{m_t}'). \tag{7.11}$$

Note that L_t^* and L_t (in (6.21)) have the same structure, with different constants. So, letting $\lambda = P_M \pi$, it can be seen as in Lemma 6.4 that $tr(L_t^* D_\lambda^{(t)} L_t'^*)$ is independent of π for all t. Now, using $L_M D_\lambda L_M' = \sum_{t=1}^{k-1} (L_t^* D_\lambda^{(t)} L_t'^*)$, we can prove the following:

Lemma 7.1. *Consider any mixture of t-subset designs, P_M as in (7.7), and the corresponding L_M in (7.11), and let $\lambda = P_M \pi$. Then, $tr(L_M D_\lambda L_M')$ is a constant, independent of π.*

This lemma leads to the following result, whose proof is similar to that of Theorem 6.3 and hence omitted.

Theorem 7.1. *Under P_M in (7.7), a linear unbiased minimax estimator of π is $L_M \hat{\lambda}$, where L_M is as in (7.11).*

Next, we give a simpler view of the minimax estimator $L_M\hat{\lambda}$. For $t = 1, \ldots, k-1$, let $\hat{\lambda}^{(t)}$ denote the vector of relative frequencies of the response types that satisfy $\sum_j z_{ij} = t$, i.e., generated by a t-subset design (in the second step of our two-step view of P_M). The data $\mathcal{D}$ can be be represented as $\mathcal{D}' = [\mathcal{D}'_1 \mid \mathcal{D}'_2 \mid \ldots \mid \mathcal{D}'_{k-1}]$, where $\mathcal{D}'_t$ contains all responses generated by the t-subset design. Let n_t denote the sample size of $\mathcal{D}'_t$. Then, using (7.11) we get

$$L_M\hat{\lambda} = \sum_{t=1}^{k-1} L_t^*\hat{\lambda}^{(t)} = \sum_{t=1}^{k-1} a_* m_t P_t'\hat{\lambda}^{(t)} - \sum_{t=1}^{k-1} \frac{n_t}{n}(a_*^{-1} - k^{-1})\mathbf{1}_k.$$

Now, using some results from the proof of Proposition 7.1, we get

$$\begin{aligned}\hat{\pi}_j &= \sum_{t=1}^{k-1}\left[a_*^{-1} m_t s_t(\gamma-1)\frac{V_j^{(t)}}{n} + \frac{n_t}{n}a_*^{-1}(m_t s_t - 1) + \frac{n_t}{n}k^{-1}\right] \\ &= a_*^{-1}\sum_{t=1}^{k-1}\left[m_t s_t(\gamma-1)\frac{V_j^{(t)}}{n} + \frac{n_t}{n}(m_t s_t - 1)\right] + k^{-1},\end{aligned}$$

where $V_j^{(t)}$ is the jth column sum of $\mathcal{D}_t$.

The minimax criterion compares maximum (over the parameter space) risks of competing procedures, and in general, a minimax procedure need not dominate (or be uniformly better) another procedure, i.e., have a uniformly smaller risk. Interestingly, the following theorem shows that the minimax estimator $L_q\hat{\lambda}$ based on the q-subset design dominates the minimax estimator $L_M\hat{\lambda}$ based on any mixture of t-subset designs.

Theorem 7.2. *Let P_M be a mixture of t-subset designs and suppose $P_M \neq P_q$. Then, the strategy (P_M, L_M) is dominated by the minimax strategy (P_q, L_q), i.e., $\boldsymbol{R}(P_q, L_q; \pi) \leq \boldsymbol{R}(P_M, L_M; \pi)$ for all π and the "=" holds if and only if $P_M = P_q$.*

Proof. By (6.1) and lemmas 6.4 and 7.1, the difference of the risk functions of the

two strategies,

$$\mathbf{R}(P_M, L_M; \pi) - \mathbf{R}(P_q, L_q; \pi) = tr(L_M D_\lambda L_M') - tr(L_q D_\lambda L_q'),$$

is independent of π. So,

$$\mathbf{R}(P_M, L_M; \pi) - \mathbf{R}(P_q, L_q; \pi) = \mathbf{R}(P_M, L_M; \pi_u) - \mathbf{R}(P_q, L_q; \pi_u). \tag{7.12}$$

Now, the proof can be completed by noting that if $P_M \neq P_q$, then P_M contains rows that have more than q large values and hence by Theorem 6.2, (7.12) > 0. □

Remark 7.1. *For a mixture design* $P_M = \left[w_0 P_0' \mid w_1 P_1' \mid \ldots \mid w_k P_k'\right]'$ *that includes the two constant rows corresponding to* $t = 0$ *and* $t = k$*, as in RAPPOR design, the preceding results hold with simple changes. In particular, the sums in (7.8) and (7.9) will be over* $t = 0$ *to* k *and* L_M *in (7.11) will include* L_0^* *and* L_k^**. With these changes, theorems 7.1 and 7.2 hold true. Note from (6.8) and (6.19) that* $a_0 = a_k = 0$ *and* $b_0 = b_k = 1$ *and so in (7.10),* a_* *remains the same and* $b_* = (w_0 + w_k) + \sum_{t=1}^{k-1} w_t b_t$*.*

Next, we shall discuss some directions for improving upon the basis RAPPOR strategy. The empirical estimator currently being used with RAPPOR (see Erlingsson et al. (2014) and Ye and Berg (2018)) is

$$\tilde{\pi}_R = \left(\frac{\sqrt{\gamma}+1}{\sqrt{\gamma}-1}\right)\frac{V}{n} - \frac{1}{\sqrt{\gamma}-1}\mathbf{1}_k, \tag{7.13}$$

where V is the vector of column sums of the data matrix, as in Section 7.1. Specifically, the jth component (V_j) of V is the number of responses with $Z_j = 1$. Kairouz et al. (2016a) derived the risk of the empirical estimator as

$$\mathbf{R}(P_R, \tilde{\pi}_R; \pi) = \frac{k\sqrt{\gamma}}{(\sqrt{\gamma}-1)^2} + 1 - \sum_{i=1}^{k} \pi_i^2. \tag{7.14}$$

It can be seen that $\tilde{\pi}_R$ is different from the minimax estimator $\hat{\pi}_R = L_R\hat{\lambda}$ under RAPPOR design, where L_R is similar to (7.11), as noted in Remark 7.1.

Theorem 7.3. *For RAPPOR design,* $\boldsymbol{R}(P_R, L_R; \pi) < \boldsymbol{R}(P_R, \tilde{\pi}_R; \pi)$ *for all* π *and thus, the empirical estimator* $\tilde{\pi}_R$ *in (7.13) is dominated by the minimax linear unbiased estimator* $L_R\hat{\lambda}$.

Proof. By Lemma 7.1 and (7.14), the difference of the two risks

$$\mathbf{R}(P_R, \tilde{\pi}; \pi) - \mathbf{R}(P_R, L_R; \pi) = \frac{k\sqrt{\gamma}}{(\sqrt{\gamma} - 1)^2} + 1 - tr(L_R D_\lambda L_R')$$

is independent of π. So,

$$\mathbf{R}(P_R, \tilde{\pi}_R; \pi) - \mathbf{R}(P_R, L_R; \pi) = \mathbf{R}(P_R, \tilde{\pi}_R; \pi_u) - \mathbf{R}(P_R, L_R; \pi_u) > 0,$$

as in the proof of Theorem 7.2. □

The preceding result shows that the RAPPOR strategy $(P_R, \tilde{\pi}_R)$ can be improved by replacing the empirical estimator by the linear unbiased minimax estimator $\hat{\pi} = L_R\hat{\lambda}$. As we noted earlier, the basic RAPPOR design is inadmissible, as its TPM includes two constant rows. Deleting those two rows and normalizing the TPM we get a modified RAPPOR design that is admissible. Thus, a better idea would be to use this modified design and the corresponding minimax estimator. However, this modified RAPPOR method is worse than the minimax strategy (P_q, L_q), by Theorem 7.2.

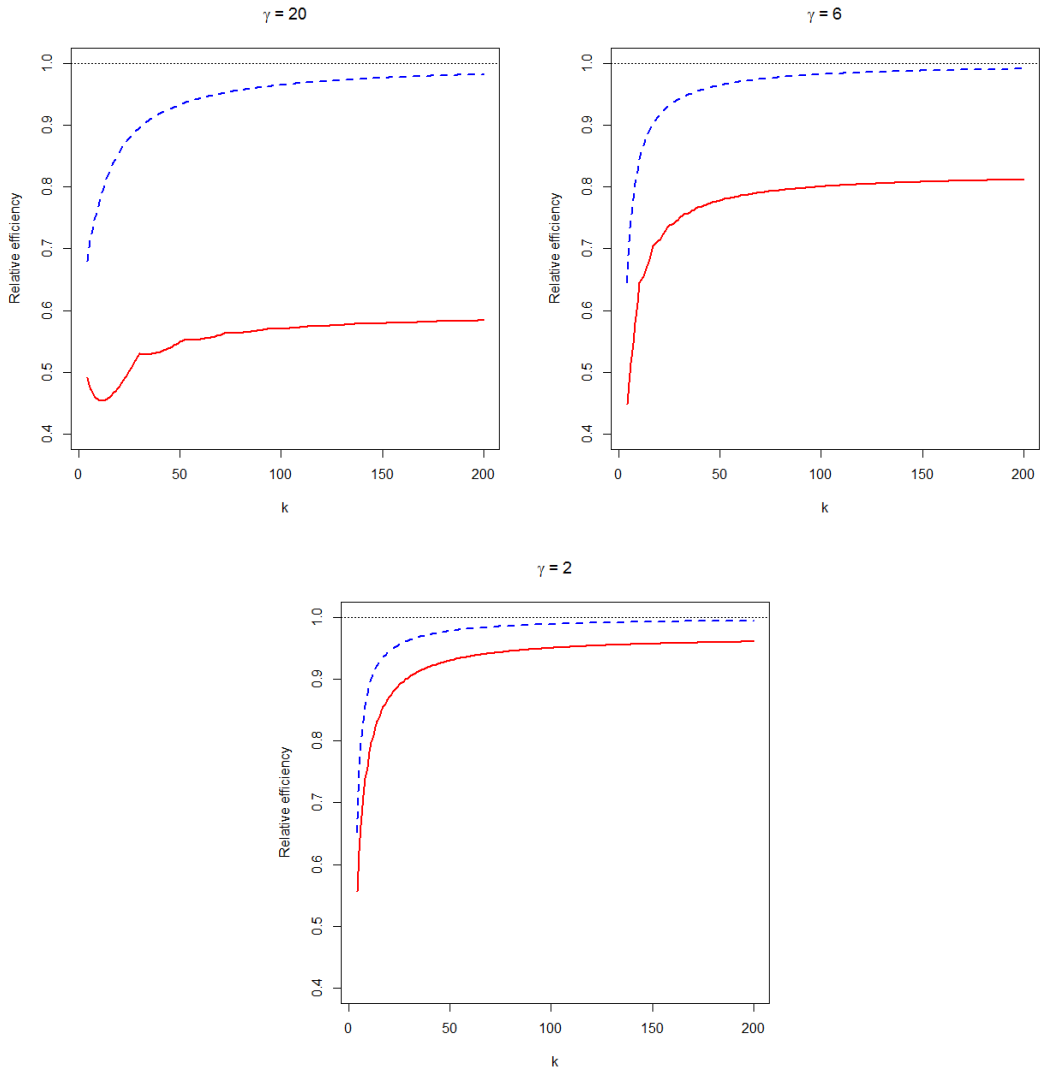

Figure 7: Relative efficiency comparison

We compared the sample size adjusted risks, defined in (6.1), of (P_q, L_q) and (P_R, L_R) with that of $(P_R, \tilde{\pi}_R)$ for some γ and k. The results are presented in Figure 7. The solid curves represent the relative efficiency of $(P_R, \tilde{\pi}_R)$ compared to (P_q, L_q). They show the ratio of the risk of (P_q, L_q) to that of $(P_R, \tilde{\pi}_R)$. Similarly, the dashed curves compare the risks of (P_R, L_R) and $(P_R, \tilde{\pi}_R)$. Thus, they show possible efficiency gain from just replacing the RAPPOR's estimator by the minimax estimator under RAPPOR design, which can be obtained from (7.11). All relative efficiency

curves are always less than 1, consistent with our theoretical results. We see that the minimax strategy is substantially better than the original RAPPOR strategy, especially for moderate to large γ, i.e., in moderate to low privacy paradigm. Under RAPPOR design, the efficiency gain from using the corresponding minimax estimator is noticeable for small to moderate k, depending on γ.

Chapter 8

Discussion and Future Research

In this dissertation, we investigated the logic underlying the ρ_1-to-ρ_2 and β-factor privacy criteria in full generality. We gave new insight and clarity using geometrical representation of privacy breach regions. We introduced the concepts of precise PBR and canonical strict information privacy to accurately describe the privacy demands of any stated criterion. Our Theorem 4.1, which gives necessary and sufficient conditions for attaining desired privacy, is a significant result. It also yields a numerical measure of the privacy demand of any given PBR, and shows that the parity of an RR procedure determines its privacy guarantee. It also gives a set of practically relevant PBRs and tells us to choose one of those in setting privacy requirement in real applications.

We further investigated the data utility of using RR procedure under given privacy level. In Chapter 5, we compared data utility of privacy satisfying RR procedures using sufficiency of experiments, which is a strong criterion that does not rely on any specific loss function or utility measure. The class of all privacy preserving admissible RR procedures, $\mathcal{C}_\gamma^a$, is an important finding. In Chapter 6, we obtained two optimality results: (i) Maximizing the trace of P, the sum of unchanged probability, when $m = k$.

(ii) We developed a minimax criterion, using squared error loss and restricting to linear unbiased estimation. We derived a minimax strategy, which includes both RR design and the corresponding estimator. This investigation identified an interesting class of RR, t-subset designs. In Chapter 7, we gave a simpler view of t-subset designs and a simplified form of the minimax estimator, which coincides with the moment estimator proposed by Wang et al. (2016) and Ye and Berg (2018). It is much easier to implement in practice, as one does not need to generate so many responses and deal with the matrix inverse. We also reviewed and explored RAPPOR, the RR method used by Google and Apple. We related RAPPOR to the mixture of t-subset designs and derived the minimax estimator of RAPPOR. We also showed that under the minimax criterion, the q-subset design is strictly better than RAPPOR, and the improvement is quite considerable under medium and low privacy requirements.

h-CSIP is a compelling privacy protection goal. However, it may be overly stringent in practice. Cell collapsing (or generalization) is a common privacy protection tool, which can be viewed as a special case of RR, with $P_\alpha(Z = d_i|X = c_j) = 1$ if c_j is collapsed within d_i (or d_i contains c_j) and 0 otherwise. But, the parity of any such TPM is infinity, unless $m = 1$, in which case data utility is null. So, cell collapsing cannot give any strict information privacy without totally destroying data utility. It will be useful to modify the criterion by requiring no privacy breach for a subset of properties $\mathcal{Q}$. Actually, intruders may only be interested in some properties, such as the marginal distribution of a sensitive category, but not all Q. We leave choosing $\mathcal{Q}$ and appropriately modifying our results as future research topics.

Another research topic is the balance of privacy and data utility when k is very large, such as in association rule mining or social network data. In association rule mining, suppose there are d items $(a_1, a_2, \ldots, a_d)$ in total and each customer buys some of them. For each item, there are 2 categories: "Yes" and "No". Each purchase record must take one of these two categories for every item, so there are 2^d different

combinations. After cross-classification, the data have $k = 2^d$ categories. However, d is usually not a small value, (i.e. $d = 20$), so $k = 2^d$ is extremely large. As Figure 6 shows, the risk is proportional to k, so the data utility will be very poor for an extremely large k. Indeed, one can increase γ to keep the data utility, but the privacy protection gets weaker. To balance the privacy and data utility, a possible solution is to customize this strict criterion in some way, and choose the TPM that satisfies the new criterion so that data utility can be improved.

Another reason why a customized criterion is needed is that its practical relevance decreases as k increases. The privacy view is based on intruder's prior and posterior comparison, so a full prior vector by intruder is needed. Providing the full prior can be done when the number of original categories k is relatively small, but it is rather tricky when k is large, especially when the original categories are the cross-classification of several variables. We shall further explain this in the context of association rule mining. For any intruders, the prior distribution in each item is a Bernoulli distribution with mean p_i. The whole prior is a vector with dimension 2^d, as it is the distribution of cross-classification. It is easy for an intruder to provide subjective probability of the mean vector for all items, and to provide the dependence between two variables is also possible. However, it is much more complicated to describe the relationship among 3 or more items. Hence, it is formidable for an intruder to articulate the whole prior for all 2^d cells, in which case the posterior cannot be calculated. A practical way to generate the full prior is to assume independence among different variables, so the full prior follows from the marginal priors p_i. However, intruders are usually only interested in some specific sensitive items, so they probably just have marginal priors for the sensitive items, or the dependence of a sensitive item and a related item. In conclusion, how to articulate the posterior without full prior is the main problem in practice.

To solve this problem, a direct idea is to perturb different items independently.

Since each item a_l has only 2 categories "Yes" and "No", it can be perturbed by a 2×2 matrix P_l. Suppose an intruder only has a prior for a specific item and is only interested in its posterior, he/she can simply use the 2×2 matrix P_l for that item to calculate the posterior. Suppose an intruder has a joint prior for several items, the joint randomization procedure should be considered. This can be calculated easily by considering the Kronecker product of these 2×2 TPMs. Also, the parity of each P_l can be unequal, in order to satisfy different privacy requirements for different items. Some items are not sensitive at all, so we even do not need to perturb them. In short, with independent perturbation, intruders can always calculate posterior with partial prior of several variables.

However, if we use independent perturbation with h-CSIP, the TPM would be the Kronecker product of all 2×2 TPMs. It can be shown that the parity of the whole TPM is $\prod_{i=1}^{d} \eta_i$, where η_i is the parity of the 2×2 TPM P_i. Clearly the parity is very large if d is not small, so the privacy cannot be well protected under h-CSIP. Conversely, to satisfy h-CSIP, the parity of each 2×2 TPM have to be very small, which severely impacts the data utility. Actually, every item should be perturbed, even if it is not sensitive, otherwise the parity will be infinity. This is because h-CSIP considers all prior α, and privacy breach happens when the prior has a extreme dependence between two variables, though the dependence may not be reasonable in practice. For example, a possible such prior may imply $P(\text{Buy a gun} \mid \text{Buy bread}) = 1$. We have to perturb bread to avoid privacy breach under this ridiculous prior, but the perturbation of bread decreases the data utility. If we can exclude this kind of "unreasonable" prior, the independent perturbation will work better. Hence, we need to customize h-CSIP by considering only some partial prior instead of the whole prior, which also has more practical relevance as we mentioned earlier. A possible solution is only considering priors that has no dependence between variables, i.e. the priors that is generated by marginal p_i. We leave detailed investigations for future research.

Finally, as we discussed in Chapter 2, the unbiased estimator of π may fall out of the parameter domain. In practice, the estimate should be projected onto probability simplex. Agrawal et al. (2009) suggested a method that first zeroes all negative estimates, and then normalizes the remaining terms such that they sum to 1. Kairouz et al. (2016a) proposed an algorithm to calculate the MLE for γ-diagonal design, which is based on the normalization idea. Taking the projection method into consideration for general t-subset design and investigating its influence in utility is left for future research.

Bibliography

Abul-Ela, A.L.A., Greenberg, G.G., and Horvitz, D.G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association*, **62**, 990-1008.

Aggarwal, C.C. and Yu, P.S. (Eds.) (2008). *Privacy-Preserving Data Mining: Models and Algorithms*, New York: Springer Science and Business Media.

Agrawal, S., Haritsa, J.R. and Prakash, B.A. (2009). FRAPP: A Framework for high-accuracy privacy-preserving mining. *Data Mining and Knowledge Discovery*, **18**, 101-139.

Agrawal, R., and Srikant, R. (2000). Privacy-preserving data mining. In *ACM Sigmod Record* (Vol. 29, No. 2, pp. 439-450). ACM.

Anderson, H. (1976). Estimation of a proportion through randomized response. *International Statistical Review*. **44**, 213-217.

Basu, D. (1988). Likelihood and partial likelihood. In *Statistical Information and Likelihood: A Collection of Critical Essays by Dr. D. Basu*, J.K. Ghosh (ed.), Springer, New York, pp. 313-320.

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38-45.

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 93-102.

Blackwell, D. (1953). Equivalent comparison of experiments. *Annals of Mathematical Statistics*. **24**, 265-272.

Boreale, M., and Paolini, M. (2015). Worst-and average-case privacy breaches in randomization mechanisms. *Theoretical Computer Science*, **597**, 40-61.

Chakravarti, I.M. (1975). On a characterization of irreducibility of a non-negative matrix. *Linear Algebra and Its Applications*, **10**, 103-109.

Chaudhuri, A. (2010). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton: CRC Press.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

Chai, J., and Nayak, T.K. (2018). A criterion for privacy protection in data collection and its attainment via randomized response procedures. *Electronic Journal of Statistics*, **12**, 4264-4287. https://doi.org/10.1214/18-EJS1508

Chai, J., and Nayak, T.K. (2019). Minimax Randomized Response Methods for Providing Local Differential Privacy. *U.S. Census Bureau Research Report Series, Statistics*, **04**.

Chen, B-C., Kifer, D., LeFevre, K. and Machanavajjhala, A. (2009) Privacy-preserving data publishing. *Foundations and Trends in Databases*, **2**, 1-167.

Cruyff, M.J., Van Den Hout, A., and Van Der Heijden, P.G. (2008). The analysis of randomized response sum score variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 21-30.

Duchi, J.C., Jordan, M.I., and Wainwright, M.J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, **113**, 182-201.

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*. Springer, Berlin, Heidelberg, pp. 1-19.

Erlingsson, U., Pihur, V. and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, Scottsdale, Arizona, pp. 1054-1067.

Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in privacy-preserving data mining. *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, San Diego, pp. 211-222.

Evfimievski, A., Srikant, R. Agrawal, R. and Gehrke, J. (2004) Privacy preserving mining of association rules. *Information Systems*, **29**, 343-364.

Fanti, G., Pihur, V. and lfar Erlingsson, U. (2016). Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, **3**, 4161

Fligner, M.A., Policello, G.E. and Singh, J. (1977). A comparison of two randomized response survey methods with consideration for the level of respondent protection. *Communications in Statistics-Theory and Methods*. **6**, 1511-1524.

Fox, J.A. (2016). *Randomized Response and Related Methods: Surveying Sensitive Data*. Thousand Oaks, CA, Sage Publications.

Fung, B.C.M., Wang, K., Chen, R. and Yu, P.S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, **42**, 14.

Greenberg, B.G., Abul-Ela, A.-L.A., Simmons, W.R. and Horvitz, D.G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*. **64**, 520-539.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, **14**, 463-478.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. New York: John Wiley & Sons.

Kairouz, P., Bonawitz, K., and Ramage, D. (2016a). Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning*, New York, pp. 2436-2444.

Kairouz, P., Oh, S., and Viswanath, P. (2016b). Extremal Mechanisms for Local Differential Privacy. *Journal of Machine Learning Research*, **17**, 1-51.

Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Kifer, D. and Lin, B-R. (2012). An axiomatic view of statistical privacy and utility. *Journal of Privacy and Confidentiality*, **4**, 5-49.

Marshall, A.W., Olkin, I., and Arnold, B. (2011). *Inequalities: Theory of Majorization and Its Applications.* New York: Academic press.

Lax, Peter D. (2007). *Linear Algebra and Its Applications.* 2nd ed., Wiley-Interscience.

Lehmann, E. (1988). Comparing location experiments. *Annals of statistics*, **16**, 521-533.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106-115). IEEE.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). ℓ-Diversity: Privacy Beyond κ-Anonymity. *In Proc. 22nd ICDE* (p. 24). IEEE.

Minc, H. (1988). *Nonnegative Matrices*. New York: John Wiley & Sons.

Nayak, T.K. (1994). On randomized response surveys for estimating a proportion. *Communications in Statistics-Theory and Methods*, **23**, 3303-3321.

Nayak, T.K., and Adeshiyan, S.A. (2009). A unified framework for analysis and comparison of randomized response surveys of binary characteristics. *Journal of Statistical Planning and Inference*, **139**, 2757-2766.

Nayak, T.K. and Adeshiyan, S. A. (2016). On invariant post-randomization for statistical disclosure control. *International Statistical Review*, **84**, 26-42.

Nayak, T.K., Adeshiyan, S.A. and Zhang, C. (2016). A Concise Theory of Randomized Response Techniques for Privacy and Confidentiality Protection. *Handbook of Statistics*, **34**, 273-286.

Nayak, T.K., Zhang, C., and Adeshiyan, S.A. (2015). Emerging applications of randomized response concepts and some related issues. *Model Assisted Statistics and Applications*, **10**, 335-344.

Nayak, T.K., Zhang, C., and You, J. (2018). Measuring Identification Risk in Microdata Release and Its Control by Post-randomisation. *International Statistical Review*, **86**, 300-321.

Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, **100**, 1103-1112.

Rizvi, S.J. and Haritsa, J.R. (2002). Maintaining data privacy in association rule mining. In *Proceedings of the 28th international conference on Very Large Data Bases (pp. 682-693)*. VLDB Endowment.

Shlomo, N., and Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *The Annals of Applied Statistics*, **4**, 1291-1310.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **10**, 557-570.

Taussky, O. (1949). A recurring theorem on determinants. *The American Mathematical Monthly*, **56**, 672-676.

Torra, V. (2017). *Data Privacy: Foundations, New Developments and the Big Data Challenge*. New York: Springer.

Van den Hout, A., and Elamir, E.A. (2006). Statistical disclosure control using post randomisation: Variants and measures for disclosure risk. *Journal of Official Statistics*, **22**, 711-731.

Van den Hout, A. and Van der Heijedn, P.G. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, **70**, 269-288.

Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X-Y. and Qiao, C. (2016). Mutual Information Optimally Local Private Discrete Distribution Estimation. *arXiv preprint arXiv:1607.08025*

Wang, T., Blocki, J, Li, N. and Jha, S. (2017). Locally differentially private protocols for frequency estimation. In *Proceedings of 26th USENIX Security Symposium*.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association*, **66**, 884-888.

Willenborg, L.C.R.J. and De Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer.

Ye, M. and Barg, A. (2018). Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, **64**, 5662-5676.