# Achieving Transparency Report Privacy in Linear Time

CHIEN-LUN CHEN and LEANA GOLUBCHIK, University of Southern California
RANJAN PAL, University of Michigan

An accountable **algorithmic transparency report (ATR)** should *ideally* investigate (a) *transparency* of the underlying algorithm, and (b) *fairness* of the algorithmic decisions, and at the same time preserve data subjects' *privacy*. However, a provably formal study of the impact to data subjects' privacy caused by the utility of releasing an ATR (that investigates transparency and fairness), has yet to be addressed in the literature. The far-fetched benefit of such a study lies in the methodical characterization of privacy-utility trade-offs for release of ATRs in public, and their consequential application-specific impact on the dimensions of society, politics, and economics. In this paper, we first investigate and demonstrate potential privacy hazards brought on by the deployment of transparency and fairness measures in released ATRs. *To preserve data subjects' privacy, we then propose a linear-time optimal-privacy scheme*, built upon standard **linear fractional programming (LFP)** theory, for announcing ATRs, subject to constraints controlling the tolerance of privacy perturbation on the utility of transparency schemes. Subsequently, we quantify the privacy-utility trade-offs induced by our scheme, and analyze the impact of privacy perturbation on fairness measures in ATRs. To the best of our knowledge, this is the first analytical work that simultaneously addresses trade-offs between the triad of privacy, utility, and fairness, applicable to algorithmic transparency reports.

CCS Concepts: • **Security and privacy → Information accountability and usage control**; **Privacy protections;**

Additional Key Words and Phrases: Privacy, algorithmic transparency, fairness, linear fractional programming

**ACM Reference format:**
Chien-Lun Chen, Leana Golubchik, and Ranjan Pal. 2022. Achieving Transparency Report Privacy in Linear Time. *J. Data Inform. Quality* 14, 2, Article 8 (February 2022), 56 pages.
https://doi.org/10.1145/3460001

8

## 1 INTRODUCTION

In the era of big data and **machine learning (ML)**, automated data processing algorithms are widely adopted in many fields for classification, prediction, or decision-making tasks due to huge volumes of input data and successful performance of ML approaches. Ongoing concerns and social uproar about the transparency and fairness of such decision-making have been raised by the media, government agencies, foundations, and academics over the past decade [14, 68]. On a technical

This work was supported in part by the NSF CNS-1616575, NSF CNS-1939006, NSF CNS-1816887, and ARO W911NF1810208 awards.
Authors' addresses: C.-L. Chen, Amazon.com Services LLC, 10300 Campus Point Dr, San Diego, CA 92121, USA; email: chienluc@amazon.com; L. Golubchik, University of Southern California, 941 Bloom Walk, Los Angeles, CA 90089, USA; email: leana@usc.edu; R. Pal, University of Michigan, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109, USA; email: palr@umich.edu.

1936-1955/2022/02-ART8 $15.00
https://doi.org/10.1145/3460001

note, it has been shown in example studies that ML algorithms can be biased when (i) a dataset used to train ML models reflects society's historical biases [86], e.g., only a few female presidential nominees in the U.S. history, or (ii) because ML algorithms have much better understanding of the majority groups and poor understanding of the minority groups [12]. Thus, as we rapidly move forward to a data-driven age where a significant amount of day-day decision making in personal and professional spheres might be automation-driven, it would make great sense to often know the reasons behind certain decisions in order to understand if they are being treated fairly. Unfortunately, most decision processes today are often opaque, making it difficult to rationalize why certain decisions are made and whether they favor or disfavor certain individuals or groups.

Providing an **algorithmic transparency report (ATR)** by data controllers and third party regulatory agencies to decision-facing individuals is one way to investigate whether decisions made in a *blackbox* are fair and transparent [26, 36, 71] - an immediate application area of considerable social impact being explainable AI for medical diagnoses [50, 77] to enable medical personnel to better understand and interpret diagnostic reports, and to justify vulnerabilities of deployed AI models through domain expertise. This is a popular topic in research and there have been works in the last decade that have developed methodologies to reduce opaqueness in decision making [28, 46] and improve on its fairness relative to certain protected attributes[1] [63]. The notion of transparency has also made its way into recently implemented policies for data protection such as the EU **General Data Protection Regulation (GDPR)**, and the **California Consumer Privacy Act** of 2018 (**CCPA** or AB-375) - both of which regulate the processing of collected personal or non-personal data of any data subject (the natural person to whom the data and the decision process relate) [44]. More specifically, any data controller shall inform data subjects before collecting their data, and is required to clearly explain the purpose of collecting data and how data will be processed, upon data subjects' requests ("*right to explanation*" and "*right to non-discrimination*") [44]. *However, a major side effect of providing transparency and fairness guarantees to the decision-facing clients is an unwanted risk to the privacy of other clients in a database.* To this end, there exists substantial literature pointing out potential privacy threats in ML [69], including membership attacks [80], training data extraction (model inversion attack) [39, 40], model extraction [87], and so on. However, for ATRs, although it has been pointed out that transparency, proposed by legislature to protect people's rights, may hurt privacy [8, 23], *it is yet to be made methodically clear how transparency can hurt privacy.*

**Goal** - An *accountable* ATR, especially for automated ML decision processes, should ideally include *transparency* of the underlying algorithm, ability to inspect *fairness* of the algorithmic decisions, and most importantly, preserve data subjects' *privacy* ("*right to privacy*" [20]), as depicted in Figure 1. *Our goal in this paper is to work towards this goal and study the corresponding trade-offs between the triad elements.*

In this paper, we investigate this problem and explicitly show that data subjects' private information can be inferred via various transparency schemes and fairness measures in announced ATRs.

**Research Contributions** - We make the following contributions in this research paper:

- We explicitly demonstrate inference attacks on data subjects' private information using a synthetic and a real dataset and show that such attacks can be performed on various transparency schemes *without strong assumptions* of adversaries' knowledge. These instances expose the possible aspects of algorithmic transparency that could hurt data subjects' privacy and subsequently have negative socio-political implications (See Section 3 and Appendix B).

[1]Protected attributes form a subset of attributes, to which any decision process should not show preference, in any instance. It may contain public attributes (gender, race, etc.) and/or private/sensitive attributes (health conditions, gene, etc.).

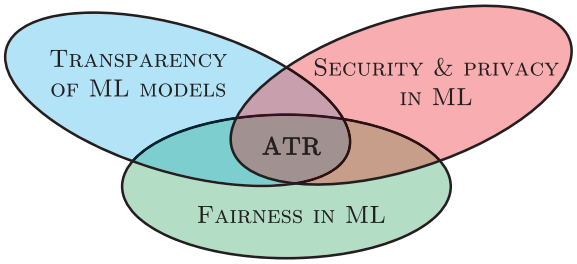


Fig. 1. A depiction of the realm of accountable ATRs.

- To protect data subjects' privacy in an ATR, we propose a privacy-aware mechanism perturbing the unveiled rationale of opaque decision-making algorithms to control the amount of disclosed information in ATRs, at the same time providing sufficient utility. Specifically, given a released privacy-preserving ATR, for *honest-but-curious* adversaries which may know (i) the target individuals' public information used as inputs to the decision model, (ii) the target individuals' received decisions (model outputs), and (iii) side-information from auxiliary sources, the maximum confidence of inferring any sensitive information about any data subject is guaranteed not exceeding a predetermined privacy threshold (See Section 4 and 5).
- We study the trade-off between privacy and utility of ATRs. Specifically, we aim to understand the minimum required perturbation/distortion in order to provide a certain privacy guarantee, or the maximum privacy that can be provided/guaranteed subject to utility constraints, which can be formulated as an optimization problem for privacy-utility trade-off in ATRs. In addition, we analyze the impact of privacy perturbations on fairness measures. In this regard, our work provides useful quantitative trade-offs and influences between privacy, transparency, and fairness measures in ATRs (See Sections 5 and 6).
- We deduce that our privacy-utility optimization problem is equivalent to a *generalized* ***linear fractional programming*** *problem* **(LFP)** [16, 94]. Such a problem can in general be solved as a sequence of linear programming feasibility problems, each with pseudo-polynomial time complexity with respect to the number of problem variables (the number of different *decision regions*[2] in our problem), which, however, in the worst case, grows exponentially with the number of input attributes - hence lending the said optimization problem intractable for large record sizes. *However, on a closer investigation*, we figure out that the region of interest in the solution space can be decomposed into disjoint "subspaces" leading to multiple independent sub-problems - each bearing important properties and amenable to propose *closed-form* solutions. Subject to utility constraints, the optimal-privacy protection scheme can thus be solved from the optimization problem efficiently in *linear time* (See Sections 7 and 8).

## 2 APPLICABILITY AND ETHICALITY OF THE PROPOSED PRIVACY SCHEME

In this section, we first describe a range of possible application domains for the proposed ATRs privacy protection scheme. We then discuss potential ethical concerns related to announcing a perturbed "transparency" report, specifically, the conflict between principles of *transparency* and *perturbation* for preserving data subjects' privacy.

[2]The regions of input attributes partitioned by decision rules; see Section 3 for examples.

### 2.1 Applicability

The proposed privacy protection scheme, based on perturbing of decision mappings (Definition 1), can be applied to numerous applications with characteristics that *the decision regions of the application decision process are disjoint finite sets of the input attribute space, i.e., the application decision process can be represented by a finite number of decision mappings*.[3] Although, to our knowledge, to date there is no current instantiation of an *algorithmic* transparency report, we believe that potential applicable domains for our scheme include the following (among others):

- *University admissions and job recruitment*: Fairness in university/college admissions has become a significant public concern, attracting more and more attention from the public as well as the media [37], even though the fact that the definition of "fairness" is still controversial, e.g., whether race should be used in admissions decisions to reflect racial diversity [65, 90]. To respond to the public's concerns, some governments [2] and universities [1] initiated work on *admissions transparency*, providing statistical data from applications (inputs) and admissions (outputs). Currently, we are not aware of instantiations disclosing the admission decision process; however, a negligent report could leak applications' data or records, e.g., (range of) SAT scores, competition records and ranks, extracurricular activities, or volunteer work. Similar circumstances apply to applications and corresponding decisions in job recruitment and other related domains.
- *Credit scores and the associated domains*: When it comes to evaluating and identifying everyone's financial creditworthiness based on credit scores, people have the following concerns: are their credit scores computed/treated fairly [49] and whether they have the ability to identify and contest any (potentially) unfair credit decisions [52]. Similar circumstances apply to credit card applications [27] and a variety of loans, i.e., domains where credit scores may be taken into account. In this paper, we demonstrate potential privacy hazards that could be brought on by a bluntly disclosed ATR in credit card applications via various types of transparency schemes and fairness measures using a synthetic (Section 3) and a real dataset (Appendix B).
- *Medical or pharmocogenetic models*: As noted in [53], a pharmocogenetic model has been built to predict proper dosages for patients based on their clinical histories, demographics, and genotypes. However, it has been shown in [40] that once accurate information about the model is leaked (or obtained through hacking), it can be utilized by an attacker to identify patients' genotypes, which could be exploited to further infer other private information, e.g., risk of getting a particular disease or someone's family ancestry. In the ATR setting, we focus on a related scenario where an adversary has no ability to access the pharmocogenetic model internals but can merely gain model information from an announced ATR. In such a case, our proposed privacy protection scheme can be applied to preserve patients' privacy and their genotype information.
- *Open Government*: It has been reported in [36] that nowadays governments utilize algorithms to detect or to determine a variety of issues, such as illegal insider trading, eligibility for public health benefits, and tax evasion. In this regard, the Open Government organization [3] aims to bring transparency to the data and the algorithms used by governments, aiming for people and society to supervise governments' actions and decisions. However, opening a government blackbox can be very dangerous and can bring catastrophic results to society if

[3]Our scheme may not be a good fit for some application domains where it may not be possible to represent the decision process using a finite number of decision mappings, e.g., applications of natural language processing, such as speech/music recognition, speech/text understanding, and text/intent classification, or applications of image and video processing, such as text recognition, item detection and alert, and image classification.

the released information is not carefully treated, and hence it is crucial to have a provable privacy-preserving scheme for any planned-to-disclose information to protect people's privacy and secret information (tax data, health/medical records, banking information, business processes, and so on).

- *Online advertising*: ML algorithms can be biased [12, 86], and it has been shown that bias also appears in online advertising, one of the ML applications that we probably experience daily. In [78], authors indicate a bias and a privacy issue associated with online advertising settings, particularly when users select the "Rather Not Say" category of gender. In addition, [24] also found that setting the gender to female results in receiving fewer instances of advertisements related to high paying jobs as compared to setting it to male. With concerns about ML bias and the purpose of the collected data, GDPR and CCPA stipulate rights to explanation and non-discrimination for data subjects; ad providers are required to respond to data subjects' requests regarding what data has been collected, how data is used, and if the applied ML models treat them fairly. Thus, all the disclosed information in an ATR may need to be further processed to protect data subjects' privacy.

### 2.2 Ethicality

When our proposed privacy protection scheme is applied to an ATR, the announced information regarding the opaque decision process may be more or less distorted, and the announced measured fairness/bias may also deviate from the true one. This may raise concerns about the *ethicality* (manner of being ethical) of the process, i.e., whether the perturbed information could mislead the public into trusting or believing that a biased decision process is fair, and vice versa.

Similarly to **privacy preservation in data-mining (PPDM)** and **data-publishing (PPDP)**, a common theme is to find an optimal trade-off between utility and privacy, subject to a certain degree of privacy guarantee for data subjects. In the context of ATRs - although both transparency and privacy are major principles in data ethics [60, 82] - we believe that data subjects' privacy should have higher priority [6]. Similarly to PPDM or PPDP, an auxiliary note could be appended with the announced information indicating that some listed information might be anonymized or perturbed for data subjects' privacy, which could help the public understand how to interpret the disclosed information appropriately. Moreover, in light of this, in this work, we propose a fidelity measure (Section 5.2) for the announced decision mappings and characterize the influence of privacy perturbation on the measured fairness (Section 5.3). This information can also be disclosed with the announced ATR in order to further assist the information recipients in understanding the range of true measures.

## 3 DEMONSTRATING PRIVACY LEAKAGE VIA AN ATR

As a necessary and important step, we first motivate our research by comprehensively demonstrating via an example consumer database of how a data subject's (i.e., consumer's) private information can be leaked via an announced **algorithmic transparency report (ATR)**. *In this work we only focus on reports that provide a rationale on the use of ML models to process individual records.* As section structure, we start by briefly reviewing transparency approaches on which privacy leakages can be induced, and follow it up with a specific example of privacy leakage on each transparency approach.

### 3.1 Algorithmic Transparency Report (ATR) in a Nutshell

ML models used to make decisions on consumer individuals are often opaque to the latter, and act as blackboxes. A survey of popularity-gaining transparency schemes to explain ML blackboxes is provided in [46]. A common representative (from the survey) transparency approach collects both

input data and labeled outputs (decision outcomes) as a training dataset, to train an ML *surrogate model* (e.g., linear model, logistic regression, decision tree, decision rules) to mimic the behavior of the blackbox. Popular methods include Anchors [74] and PALM [57]. The output of such learned behavior must be interpretable (understandable) by humans. Another common approach (e.g., [18, 38]) extracts certain important "properties" from blackbox models, such as contributions of input features, to model outputs. Specifically, these transparency schemes measure feature importance (based on the underlying *decision mapping*, see Definition 1), using both amplitude and sign to represent importance/influence of input features, where larger amplitude represents greater influence, and the sign indicates positive or negative effect on the corresponding output. Popular methods include LIME [73], FIRM [93], QII [23], Shapley Value [56], PDP [41], ICE [43], and ALEPlot [10]. In addition to transparency schemes, an ATR may also provide information regarding whether a decision algorithm or ML model is biased against certain groups or individuals - in other words, an ATR may provide measured individual- or group-*fairness* of ML-based decisions based on the different desired metrics discussed in existing literature [15, 22, 30, 35, 54, 55, 92]. We refer readers to Appendix A for detailed definitions of various individual and group fairness measures. *In what follows, we investigate and demonstrate privacy leakage instances via various kinds of transparency schemes and fairness measures, given* honest-but-curious *adversaries.*

### 3.2 Privacy Leakage via Interpretable Surrogate Models

As noted, transparency schemes can interpret a blackbox's rules in a human-understandable manner, such as decision rules or decision trees. Here, we explain how such transparent information can hurt a data subject's privacy. Without loss of representativity, here we set up a synthetic scenario, in which we consider the existence of a perfect interpretable surrogate model,[4] to illustrate the possibility of causing a catastrophic privacy leak.

Consider the following synthetic credit card application scenario (summarized in Table 1). A credit card application takes several input attributes from applicants, while the bank's decision process only depends on two input attributes: the applicants' annual income and their gender (which, depending on the country, may be illegal and in those cases should not be used in any decision process). Due to the suspicious differences in approval rates between male and female applicants, a third-party regulatory agency actively takes action. It collects all applicants' data and their received decisions, and trains an (assumed perfect) interpretable surrogate model, disclosing the decision rules used in the credit card application to all past applicants, as follows

$$d(\{Income\} > 200\text{k}) = 1,$$
$$d(\{Income\} \in 100\text{k}{\sim}200\text{k}, \text{Male}) = 0.5,$$

where $d(\cdot)$ is *decision rule* representing the probability of receiving a positive decision given the condition. An equivalent *if-then* decision rule form is the following

$$\begin{cases} \text{if } Income > 200k, \text{ then } Positive\ Decision; \\ \text{if } 100k \leq Income \leq 200k \wedge Male, \text{ then } Random; \\ \text{otherwise, then } Negative\ Decision. \end{cases}$$

Note that other interpretable surrogate models such as a decision tree or logistic regression can also be equivalently expressed by decision rule $d(\cdot)$.

Next, we demonstrate how the data subjects' sensitive information (annual income in this scenario) could be leaked. Revisit Table 1 in which the key input attributes, population, and decision

[4]We consider the most privacy-catastrophic case, a perfect interpretation, which has the most accurate information in an ATR.

Table 1. A Synthetic Credit Card Application Scenario

| | | Adversaries' Knowledge | | |
|---|---|---|---|---|
| | Input Attributes | | ATR | Side-Info |
| Population | Annual Income | Gender | Decision Rule | Census Statistics |
| 139 | <100k | F | 0 | 93.1% |
| 9 | 100k~200k | F | 0 | 5.7% |
| 2 | >200k | F | 1 | 1.2% |
| 117 | <100k | M | 0 | 84.2% |
| 18 | 100k~200k | M | 0.5 | 12.3% |
| 5 | >200k | M | 1 | 3.5% |

rule of the credit card application are listed. Population of applicants are aggregated according to *decision regions*, i.e., the regions of input attributes partitioned by decision rules. Here the population proportion among decision regions refers to the U.S. census data [4], and adversaries assumed blind to population of applicants utilize the U.S. census data as side-information to estimate, for each decision region, the percentage of the total number of male/female applicants (listed in the "Census Statistics" attribute; for instance, the value 93.1% in Table 1 represents the following: given that the decision region is {Annual Income < 100k; Gender=Female}, 93.1% of female applicants belong to this region). Adversaries know public information of targeted applicants and also know decision rules from an announced ATR.

When an ATR containing such a decision rule is negligently announced, as it reveals strong dependencies between annual income and decisions, any female using such a credit card in public instantly tells anyone who has ever seen the report that her annual income is above 200k, which not only results in a privacy hazard to her, but may also result in unexpected safety concerns. In such a case, an adversary does not even require auxiliary information to be able to infer someone's secret.

Male credit card owners are also at risk, although not as much. For a male credit card owner, the confidence of an adversary believing that his income is above 200k is only around 36%, compared with 100% in the case of a female owner, while based on census statistics, the confidence of an adversary believing that his income is above 200k is merely 3.5%. In other words, once such a negligent algorithmic transparency report is announced to the public, a high-income (>200k) male credit card owner's risk of exposing annual income information is increased ten fold.

In summary, releasing precise information of interpretable surrogate models (that can be equivalently expressed by decision rules) can be harmful to the data subjects' privacy, as such information gives adversaries a clear mapping between input records and received decisions. With assistance from public information and/or side-information, adversaries can abuse algorithmic transparency to undermine people's privacy. The same privacy leakage concern applies when precise information of transparency scheme is released in the form of feature importance/interaction, which, however, in the interest of space, is explicitly demonstrated in Appendix B, using a real dataset.

### 3.3 Privacy Leakage via Fairness Measures

Recall that one of the main motivations for algorithmic transparency is to understand if a decision-making algorithm is fair and complies with regulations/law, e.g., the U.S. **Equal Employment Opportunity Commission (EEOC)** regulates the ratio of the hiring rates between women and men, which should not be lower than 80% (80%-rule). In an algorithmic transparency report, such fairness measures may be required upon data subjects' demands (e.g., GDPR, Article 22).

Table 2. Fairness Measures for Table 1 in an ATR

| |
|---|
| $\mathcal{Y}_1$ = {F}, $\mathcal{Y}_2$ = {M} |
| $\mathcal{W}_1$ = {Annual Income ≤ 100k} |
| $\mathcal{W}_2$ = {100k ≤ Annual Income ≤ 200k} |
| $\mathcal{W}_3$ = {Annual Income ≥ 200k} |
| Overall approval rate for female ($\mathcal{Y}_1$) = 1.33%; |
| Overall approval rate for male ($\mathcal{Y}_2$) = 10%; |
| Bias in SP for $\mathcal{Y}_1$ and $\mathcal{Y}_2$ = 0.0866; |
| Bias in CSP for $\{\mathcal{Y}_1, \mathcal{W}_1\}$ and $\{\mathcal{Y}_2, \mathcal{W}_1\}$ = 0; |
| Bias in CSP for $\{\mathcal{Y}_1, \mathcal{W}_2\}$ and $\{\mathcal{Y}_2, \mathcal{W}_2\}$ = 0.5; |
| Bias in CSP for $\{\mathcal{Y}_1, \mathcal{W}_3\}$ and $\{\mathcal{Y}_2, \mathcal{W}_3\}$ = 0. |

To this end, consider again the credit card application in Table 1, in which the bank is under suspicion of discriminating against female applicants. Upon female applicants' demands, a regulation agency gets involved and discloses the following fairness measures for gender: (i) bias in **statistical parity (SP)** (Definition 9) for male and female applicants, and (ii) bias in **conditional statistical parity (CSP)** (Definition 10) for male and female applicants who have the same level of income. An ATR listing all the above fairness measures w.r.t. the credit card application is shown in Table 2 (see Remark 1 for details), which can be announced to the public in an electronic form, e.g., through a website (e.g., GDPR, Recital 58, information related to the public's concerns).

Moreover, a data subject, which is a credit card applicant in our scenario, has the right to inquire about the decision principle w.r.t. his or her personal data. Mary, a low-income (<100k) female who would like to know why her applications are always denied, demands information regarding the decision processing for her record. The response indicates that the approval rate for a low-income female is 0. If we let $d_{i,j}$ be the decision rule for people in $\{\mathcal{Y}_i, \mathcal{W}_j\}$ in Table 2, by utilizing the census statistics as shown in Table 1, and based on the definitions of SP and CSP for binary decisions in (30) and (31), respectively, the information provided in Table 2 tells us the following:

$$\text{Overall approval rate for female}(\mathcal{Y}_1) = 0.0133 \approx 0.931d_{1,1} + 0.057d_{1,2} + 0.012d_{1,3} \quad (1)$$

$$\text{Overall approval rate for male}(\mathcal{Y}_2) = 0.1 \approx 0.842d_{2,1} + 0.123d_{2,2} + 0.035d_{2,3} \quad (2)$$

$$\text{Bias in CSP for } \{\mathcal{Y}_1, \mathcal{W}_1\} \text{ and } \{\mathcal{Y}_2, \mathcal{W}_1\} = 0 = |d_{1,1} - d_{2,1}| \quad (3)$$

$$\text{Bias in CSP for } \{\mathcal{Y}_1, \mathcal{W}_2\} \text{ and } \{\mathcal{Y}_2, \mathcal{W}_2\} = 0.5 = |d_{1,2} - d_{2,2}| \quad (4)$$

$$\text{Bias in CSP for } \{\mathcal{Y}_1, \mathcal{W}_3\} \text{ and } \{\mathcal{Y}_2, \mathcal{W}_3\} = 0 = |d_{1,3} - d_{2,3}|. \quad (5)$$

Since Mary just got a reply indicating $d_{1,1} = 0$, from (3) and (5), Mary then knows that $d_{1,1} = d_{2,1} = 0$, $d_{1,3} = d_{2,3}$, and from (4), either $d_{1,2} = d_{2,2} + 0.5$ or $d_{1,2} = d_{2,2} - 0.5$. She can first assume $d_{1,2} = d_{2,2} + 0.5$; by plugging the values of $d_{1,1}$ into (1) and $d_{2,1}$ into (2), and replacing $d_{2,2}$ and $d_{2,3}$ by $d_{1,2} - 0.5$ and $d_{1,3}$ in (1) and (2), respectively, she gets $0.057d_{1,2} + 0.012d_{1,3} = 0.0133$ from (1) and $0.123d_{1,2} + 0.035d_{1,3} = 0.1615$ from (2). Since $d_{i,j}$ are probabilities, $\forall i, j$, $d_{1,2}$ and $d_{1,3}$ can not be greater than 1, and thus the obtained equation from (2) is infeasible, which implies the assumption is wrong. She then knows $d_{1,2} = d_{2,2} - 0.5$. Repeat the same steps and she will obtain $d_{1,2} = 0.0088$ and $d_{1,3} = 1.0692$. By understanding that any $d_{i,j}$ cannot be greater than 1 and this is probably caused by the mismatch between the census statistics and the true distribution, she would thus update $d_{1,3} = 1$ and thus obtain $d_{1,2} = 0.0013 \approx 0$; these estimates are very close to the true values. In addition, Mary can use the obtained $d_{1,2}$ and $d_{1,3}$ to further acquire $d_{2,2}$ and $d_{2,3}$. Therefore, by utilizing the decision processing rule for her record and the publicly announced fairness measures, she can obtain accurate decision rules for the credit card application. As in Section 3.2,

we know that a privacy disaster can happen when accurate decision rules are released or hacked. The adversary Mary now can utilize her obtained decision rules to infer other applicants' income.

From the above demonstrations, we have seen that a negligent ATR can result in a serious hazard to data subjects' privacy. In the following sections, we formalize the privacy leakage problem, and propose the corresponding properties and solutions. *We will revisit the examples demonstrated above again in Section 8, with our proposed solutions applied.*

*Remark 1.* Here we demonstrate how the numbers in Table 2 are calculated based on Table 1. The definitions of SP and CSP for binary decisions can be found in (30) and (31), respectively.

Overall approval rate for female ($\mathcal{Y}_1$) $= (2 \times 1)/(139 + 9 + 2) = 1.33\%$;
Overall approval rate for male ($\mathcal{Y}_2$) $= (18 \times 0.5 + 5 \times 1)/(117 + 18 + 5) = 10\%$;
Bias in SP for $\mathcal{Y}_1$ and $\mathcal{Y}_2$ $= |1.33\% - 10\%| = 0.0866$;
Bias in CSP for $\{\mathcal{Y}_1, \mathcal{W}_1\}$ and $\{\mathcal{Y}_2, \mathcal{W}_1\}$ $= |0 - 0| = 0$;
Bias in CSP for $\{\mathcal{Y}_1, \mathcal{W}_2\}$ and $\{\mathcal{Y}_2, \mathcal{W}_2\}$ $= |0 - 0.5| = 0.5$;
Bias in CSP for $\{\mathcal{Y}_1, \mathcal{W}_3\}$ and $\{\mathcal{Y}_2, \mathcal{W}_3\}$ $= |1 - 1| = 0$.

## 4 PROBLEM SETUP

In the following sections, we formalize and analyze the privacy leakage problem in ATR. To begin with, in this section, we provide essential notations listed in Table 3 and useful definitions for problem setup, followed by adversarial settings and definition of privacy violation in ATRs formally.

### 4.1 Decision Mapping

Figure 2 illustrates an opaque decision-making blackbox, which is essentially an unknown *decision mapping* function defined as follows.

*Definition 1 (Decision Mapping [30]).* Consider a decision process as illustrated in Figure 2, where $X = \{X_k \mid k = 1, \ldots, K\}$ is a set of input attributes, $A$ the output attribute (decision outcomes), and $\mathcal{A}$ the range of $A$. Recall that $\Delta(S)$ is a set of probability distributions over $S$. A decision mapping $D_{\mathcal{A}} : \mathcal{R}_X \to \Delta(\mathcal{A})$ is a function mapping from the range of input attributes to a set of probability distributions over the range of decision outcomes. Formally,

$$D_{\mathcal{A}}(X) = \{P_{A|X}(A = a|X) \mid \forall a \in \mathcal{A}\} = \{D_a(X) \mid \forall a \in \mathcal{A}\}. \tag{6}$$

Particularly, for binary decisions (0 ='negative' and 1 ='positive'), we let

$$D_{\mathcal{A}}(X) = \begin{cases} D_1(X) = d(X), & \text{for } a = 1 \\ D_0(X) = 1 - d(X), & \text{for } a = 0, \end{cases} \tag{7}$$

where $d(X)$ is decision rule [22] representing probabilities of mapping from input space to the positive decision outcome.

Clearly, decision mapping is more comprehensive, while decision rule is more concise and convenient for an ATR, e.g., decision rule in Table 1.

As noted in Section 3.1, an ATR opens an opaque decision blackbox via transparency schemes such as an interpretable surrogate model (a surrogate of $D_{\mathcal{A}}$) or feature importance/interaction (a function of $D_{\mathcal{A}}$). In addition, an ATR may also contain fairness measures (functions of $D_{\mathcal{A}}$, see Appendix A). Clearly, information provided in an ATR is in general a function of decision mapping $D_{\mathcal{A}}$ (when there is no confusion, we omit the subscript and simply write $D$ in the rest of the paper for conciseness); while released, the mapping from decision inputs to outputs are made public, and thus it is very crucial to ensure the reverse inference is not possible, or limited with

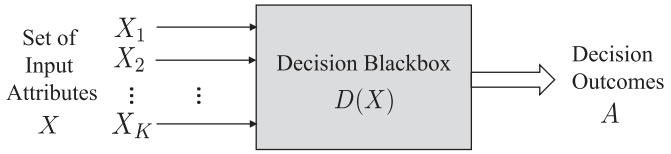


Fig. 2. A representative illustration of a decision blackbox.

low confidence. To explicitly characterize the reverse inference, we first need to understand the capability of the adversaries.

### 4.2 Adversarial Settings

For the privacy leakage problem brought on by releasing ATRs, we consider *honest-but-curious* (or *curious-but-not-malicious*) adversaries, i.e., adversaries who only perform legitimate actions and will not deviate from the defined protocol but would like to learn as much as possible (including others' secrets); in our ATR setting, this implies that an adversary will not hack into the system and steal information but only acquires as much as possible information that is made public or is widely-available. For example, the adversaries may know public information about his friends, e.g., gender, race, ZIP code, and age; the adversaries may also have knowledge about the census data [4] providing side-information (with *weak inference*) between public and private attributes, e.g., joint distributions between age, race, marriage status, household size, and income. Such kind of adversaries are ubiquitous, making privacy leakage via released ATRs omnipresent.

It's worth noting that the adversaries do not have access to *all* the input features and *all* the output responses (decision outcomes), and thus are not able to extract any information about the blackbox from the limited knowledge. Instead, the adversaries may just know the public information and the received decisions of the targeted several individuals. The reason that we particularly focus on honest-but-curious adversaries in this paper is that we would like to convey an important message that *candidly releasing an ATR could result in privacy hazards even for weak adversaries who are not able to probe or hack into the system but possess some public information and/or widely-available side-information.* Moreover, more powerful adversaries who may have access to all input features and output responses can train a more powerful surrogate model (as it need not be an interpretable model) to mimic the original model, and thus can obtain more accurate information w.r.t. the decision blackbox, as compared to what is provided in an announced ATR. In such a case, the privacy hazard is not due to the ATR, as adversaries have already obtained something more powerful (resulting in stronger inference), and thus such an adversarial setting is not meaningful for ATR.

In practice, since honest-but-curious adversaries can be ubiquitous, the background knowledge that adversaries may possess could be diverse and unknown to agencies in charge of ATRs. Therefore, it is important that agencies should consider *the worse-case scenario*, i.e., the most information that an honest-but-curious adversary can possess (which is the worst-case weak adversary). Hence, the agencies should assume that an adversary could possess *precise and full* knowledge of

- the range $\mathcal{R}_X$ and the joint distribution $P_X(\mathbf{x})$ of all inputs $\mathbf{x}$;
- all public records (a.k.a. *quasi-identifier* (QID) [5, 83, 84]) $\mathbf{x}_{\mathcal{U}}$ of specific individuals;
- the received decisions $a$ of the targeted individuals;
- the internal privacy parameters (e.g., the predefined required privacy level) of the privacy protection scheme $\mathcal{M}$ used for an ATR, if any.

The following information is assumed in general unknown (or known with little confidence) by adversaries before seeing an ATR: (i) data subjects' private records $\mathbf{x}_{\mathcal{S}}$ and (ii) the decision mapping

Table 3. Notation

| | |
|---|---|
| $\mathcal{U}$ | Set of all public attributes |
| $\mathcal{S}$ | Set of all private attributes |
| $X_k$ | Random variable (r.v.) of attribute $k$ |
| $X_{\mathcal{U}}$ | $= \{X_k \mid \forall k \in \mathcal{U}\}$; collection of r.v.'s of all public attributes |
| $X_{\mathcal{S}}$ | $= \{X_k \mid \forall k \in \mathcal{S}\}$; collection of r.v.'s of all private attributes |
| $X$ | $= (X_{\mathcal{U}}, X_{\mathcal{S}})$; collection of r.v.'s of all attributes |
| $\mathcal{R}_X$ | Range of $X$; the universe of inputs; $\mathcal{R}_X = \mathcal{R}_{X_{\mathcal{U}}} \times \mathcal{R}_{X_{\mathcal{S}}}$ |
| $\mathbf{x}_{\mathcal{U}}$ | An instance of $X_{\mathcal{U}}$ |
| $\mathbf{x}_{\mathcal{S}}$ | An instance of $X_{\mathcal{S}}$ |
| $\mathbf{x}$ | $= (\mathbf{x}_{\mathcal{U}}, \mathbf{x}_{\mathcal{S}})$, an instance of $X$ |
| $T_{\mathbf{x}_{\mathcal{U}}}$ | $= \{\mathbf{x}' \in \mathcal{R}_X \mid \mathbf{x}'_{\mathcal{U}} = \mathbf{x}_{\mathcal{U}}\}$ = range of $(\mathbf{x}_{\mathcal{U}}, X_{\mathcal{S}})$ |
| $A$ | The r.v. of decision outcome |
| $\mathcal{A}$ | Range of $A$ |
| $P(\cdot)$ | Aleatory probability; *chance* |
| $\tilde{P}(\cdot)$ | Epistemic probability; *credence* or *belief* |
| $D(X)$ | $= \{P(A = a\|X) \mid \forall a \in \mathcal{A}\}$; decision mapping (Definition 1) |
| $\tilde{D}(X)$ | $= \{\tilde{P}(A = a\|X) \mid \forall a \in \mathcal{A}\}$; announced decision mapping |
| $d(X)$ | $= P(A = 1\|X)$; decision rules (Definition 1) |
| $\tilde{d}(X)$ | $= \tilde{P}(A = 1\|X)$; announced decision rules |
| $\mathcal{M}$ | A privacy protection scheme for an ATR |

$D$ of the blackbox. Given the above adversarial settings, we clearly define privacy violation in releasing ATRs in the following.

*Definition 2.* The release of an ATR is privacy violating if any private or confidential information of any data subject to whom decision algorithms, disclosed in the ATR, have been applied can be (unintentionally) inferred by any honest-but-curious unauthorized individual or entity to whom the ATR is released, with confidence exceeding a tolerable threshold, due to the release of the ATR.

*Remark 2.* Given Definition 2, inferring attribute values due to high correlations between attributes, e.g., knowing people who have ovarian cancer are female, should not be mistaken as privacy breach (not private information; not via an ATR). Similarly, releasing ATRs to a doctor for the ML-assist diagnoses of his patients should not be considered as privacy violation (an authorized personnel).

### 4.3 Comparison with PPDM and PPDP

The main differences between privacy preservation in ATRs and **privacy preservation in data-mining (PPDM)** and **data-publishing (PPDP)** are their *adversarial settings*.

More specifically, in the PPDM setting, a dataset is not published; instead, users or data analysts send queries (a set of pre-defined/allowed deterministic functions, e.g., average, count, median, max, and min) to the curator, and the curator generates the corresponding query outputs based on the dataset. In such a setting, if the pre-defined queries are carefully designed, an adversary (a malicious user), in general, *is not able to determine the direct mappings between public and private attribute values of a record* nor *any private information of any individual* from any single query output. However, since query functions are known in advance, an adversary *can send multiple queries and compare the obtained results* to extract data subjects' private information from the outputs. In

this regard, **differential privacy (DP)** [21, 31] is usually adopted to preserve privacy in PPDM. In summary, the main differences between the settings in PPDM and ATRs are (i) in PPDM, mappings between public and private attributes are in general not available, or may be known only partially, while these could be known *statistically* in the ATRs setting; (ii) an adversary can send multiple (deterministic) queries and/or collude with other adversaries to extract data subjects' private information, while an ATR is a one-shot announcement, and the announced decision mappings from the inputs to the outputs could be probabilistic (i.e., random decisions).

In the PPDP setting, a dataset is published. Therefore, the mappings between public and private attributes are *clearly known* to an adversary (much stronger than auxiliary or side-information). When the published dataset shows *uniqueness of a public record* or *unique relationship between certain public and private attributes*, an adversary can utilize such uniqueness to identify data subjects' private information. Therefore, several techniques ($k$-anonymity [75, 76], $l$-diversity [61, 62], etc.) are proposed in the literature to obfuscate such uniqueness in order to preserve data subjects' privacy. In summary, the main differences between the settings in PPDP and ATRs are (i) unlike in PPDP where privacy is leaked due to strong inference between attributes, in ATR, as we have emphasized in Section 4.2, we only consider the case of weak correlation/inference between public and private attributes, i.e., an adversary is not able to identify any private information with high confidence before seeing an ATR, while the confidence could be dramatically enhanced after an ATR is released; (ii) in PPDP, depending on the application, there may or may not exist output attributes. When there exist output attributes, as the dataset is published, *all* output attribute values are available to an adversary, which could provide strong inference between some sensitive attributes and the output attributes (for learning purposes), and thus we need to guarantee that any output attribute value is not directly associated with any individual; while in ATRs, we consider the case that a decision outcome could be directly associated with an individual (credit card applications, university admissions, etc.), but an adversary knows a few decision outcomes only.

## 5 PRIVACY, UTILITY, AND MEASURED FAIRNESS

Given clear context of adversarial settings and the definition of privacy violation in releasing ATRs, we next formulate privacy leakage caused by inference attacks, and propose a privacy-preserving mechanism for ATRs. To this end, in this section, we provide a privacy measure to mathematically characterize and formulate the degree of privacy leakage. Based on the proposed privacy measure, we formulate the requirements for a privacy-preserving mechanism for ATRs, and introduce a utility measure to characterize the influences caused by the proposed privacy-preserving mechanism; similarly, we address the influence of the proposed privacy mechanism on fairness measures.

### 5.1 Privacy Measure and Privacy-Preserving Mechanism

Recall in Section 3 we have seen privacy leakage disasters when decision rules were divulged. The fundamental problem is that transparency schemes as well as fairness measures are closely related to, or functions of, decision mapping $D$, and more importantly, if $D$ provides strong inference from public knowledge to sensitive records, once it is utilized in an ATR and obtained by an adversary, the adversary can utilize it to further acquire data subjects' secrets with high confidence.

In light of this, here we propose the following: a carefully processed $D$, denoted by $\tilde{D}$, should be adopted as a substitute for $D$ in an ATR for preserving data subjects' privacy. $\tilde{D}$ should satisfy certain privacy requirements and can be safely announced (if an ATR chooses to release an interpretable surrogate model) or utilized by transparency schemes and fairness measures provided in an ATR.

In such a case, when an ATR is released, an adversary acquires information about $\tilde{D}$, and could further utilize it in an inference channel $\langle X_{\mathcal{U}}, A \xrightarrow{P_X, \tilde{D}} X_{\mathcal{S}} \rangle$ which maps from inference source $X_{\mathcal{U}}$ and $A$ to sensitive attribute values $X_{\mathcal{S}}$. (When the context is clear, we will omit $P_X$ and $\tilde{D}$ above the arrow for simplicity.) Based on the adversarial settings in Section 4.2, one reasonable privacy measure to characterize the above inference (caused by $\tilde{D}$) is the *maximum confidence* of an adversary in inferring *any* data subject's sensitive value $X_{\mathcal{S}}$, which is also known as the *worst-case posterior vulnerability* [34]. In other words, even if an adversary knows $\tilde{D}$ and further utilizes it to perform inference attacks, the maximal confidence that the adversary can have is carefully controlled in advance to prevent privacy violation. In this regard, privacy measures of the announced version of decision mapping $\tilde{D}$ used in an ATR should reflect the maximal degree of an adversary's confidence in inferring any data subject's secret via $\tilde{D}$. Consider the case in which $\mathcal{S}$ is a singleton set, we define maximum confidence formally in the following.

*Definition 3 (Maximum Confidence).* Given the adversarial settings in Section 4.2 and an inference channel $\langle X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}} \rangle$, the confidence of inferring a certain sensitive attribute value $x_{\mathcal{S}}$ from a certain inference source $(\mathbf{x}_{\mathcal{U}}, a)$, denoted by $conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}})$, is the posterior epistemic probability of $x_{\mathcal{S}}$ given $\mathbf{x}_{\mathcal{U}}$ and $a$ as follows

$$conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}}) = \tilde{P}_{X_{\mathcal{S}}|X_{\mathcal{U}},A}(x_{\mathcal{S}}|\mathbf{x}_{\mathcal{U}}, a).$$

The maximum confidence of inferring a specific sensitive attribute value $x_{\mathcal{S}}$ from any inference sources, denoted by $Conf(X_{\mathcal{U}}, A \rightarrow x_{\mathcal{S}})$, is defined as

$$Conf(X_{\mathcal{U}}, A \rightarrow x_{\mathcal{S}}) \triangleq \max_{\mathbf{x}_{\mathcal{U}}, a} \{conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}})\}.$$

Accordingly, the maximum confidence of inferring any sensitive attribute value from any inference channel is

$$Conf(X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}}) \triangleq \max_{\mathbf{x}_{\mathcal{U}}, a, x_{\mathcal{S}}} \{conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}})\}.$$

The privacy requirement, similar to confidence bounding [88, 89], $\beta$-likeness [19], and privacy enforcement in [58], restricts the maximum confidence on inferring any sensitive attribute by a confidence threshold, a pre-determined privacy parameter $\beta$.

*Definition 4 ($\beta$-Maximum Confidence).* In an algorithmic transparency report, $\tilde{D}$ satisfies the privacy requirement *$\beta$-Maximum Confidence* if $Conf(X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}}) \leq \beta$.

Lemma 1. *The privacy requirement $\beta$-Maximum Confidence imposes the following constraints to the announced decision mapping $\tilde{D}$, $\forall \mathbf{x} \in \mathcal{R}_X$, $\forall a \in \mathcal{A}$,*

$$\frac{\tilde{D}_a(\mathbf{x}) P_X(\mathbf{x})}{\sum_{\mathbf{x}' \in T_{\mathbf{x}_{\mathcal{U}}}} \tilde{D}_a(\mathbf{x}') P_X(\mathbf{x}')} \leq \beta. \tag{8}$$

Proof. Please refer to Appendix D for detailed proof. □

*Remark 3.* Note that a privacy requirement which only prevents an adversary from *correctly* inferring any sensitive attribute value of any data subject is insufficient. The reason is that an adversary can possess the knowledge of the privacy protection scheme and its internal privacy parameters. If the privacy requirement allows an adversary to incorrectly infer wrong sensitive values with arbitrary high confidence, since an adversary may know the privacy requirement, he/she may perceive that *any sensitive attribute value which can be inferred with confidence higher than the threshold is an incorrect one*; this could become additional side-information for the adversary. An adversary can further utilize such extra side-information to narrow down the range of

conjectures, which enhances the confidence of correctly guessing the right sensitive value. The enhanced confidence could result in exceeding the privacy threshold, and thus cause a privacy hazard.

The advantage of using maximum confidence as a privacy measure is that it results in intuitive understanding of $\beta$. This could be important when a privacy scheme is used for an ATR, the regulation may require a plain explanation for the adopted privacy scheme as well as the corresponding settings and meanings of its parameters. Alternatively, one can use other privacy measures, e.g., *minimum uncertainty* (Appendix C), which is essentially conveying the same concept as maximum confidence, but the privacy parameter $\gamma$ grows with the strength of privacy.

A privacy protection scheme $\mathcal{M}$ takes the original/true decision mapping $D$ as the input and generates a privacy-preserving decision mapping $\tilde{D}$ safe for announcement with careful processing based on privacy requirements. Inevitably, the original $D$ would differ from the generated $\tilde{D}$, which is a distorted/perturbed but private version of $D$. In the next section, we introduce a utility measure to characterize the distortion.

### 5.2 Utility Measure: Fidelity

In this section, we introduce an appropriate utility measure for our problem. Given proposed $D \xrightarrow{\mathcal{M}} \tilde{D}$, an appropriate utility measure should characterize the distortion from $\tilde{D}$ to $D$, or quantify the quality of faithfulness of $\tilde{D}$ (compared with $D$), and hence, particularly, is named *fidelity* measure hereafter. By imposing fidelity constraints to $\mathcal{M}$, the maximal distortion between $\tilde{D}$ and $D$ is guaranteed to be bounded accordingly.

*Definition 5 ($\delta$-Fidelity).* A privacy perturbation method $\mathcal{M} : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})$ satisfies $\delta$-fidelity, $\delta \in [0, 1]$, if $\forall \mathbf{x} \in \mathcal{R}_X$ and $\forall a \in \mathcal{A}$, we have

$$|\tilde{D}_a(\mathbf{x}) - D_a(\mathbf{x})| \leq 1 - \delta. \tag{9}$$

*Definition 6 ($\alpha$-Fidelity).* A privacy perturbation method $\mathcal{M} : \Delta(\mathcal{A}) \to \Delta(\mathcal{A})$ satisfies $\alpha$-fidelity, $\alpha \in [0, 1]$, if $\forall \mathbf{x} \in \mathcal{R}_X$ and $\forall a \in \mathcal{A}$, we have

$$\alpha \leq \tilde{D}_a(\mathbf{x})/D_a(\mathbf{x}) \leq 1/\alpha. \tag{10}$$

In the most general form, definition of fidelity can be

$$\tilde{D}_a(\mathbf{x})_{\min} \leq \tilde{D}_a(\mathbf{x}) \leq \tilde{D}_a(\mathbf{x})_{\max}, \tag{11}$$

which describes the restriction (the allowed range) of distortion of $\tilde{D}$ in a very general manner. The corresponding equivalent representations for $\delta$- and $\alpha$-fidelity are

$$D_a(\mathbf{x}) - (1 - \delta) \leq \tilde{D}_a(\mathbf{x}) \leq D_a(\mathbf{x}) + (1 - \delta), \tag{12}$$

$$\alpha D_a(\mathbf{x}) \leq \tilde{D}_a(\mathbf{x}) \leq \frac{1}{\alpha} D_a(\mathbf{x}), \tag{13}$$

in which the upper and the lower bounds $\tilde{D}_a(\mathbf{x})_{\max}$ and $\tilde{D}_a(\mathbf{x})_{\min}$ are functions of $D$ and $\delta$, or $\alpha$. In practice, $\delta$ and $\alpha$ should not be far from 1.

### 5.3 Influence of Privacy on Measured Fairness

As demonstrated in Section 3.3, since fairness measures are functions of decision mapping/rule, releasing precise fairness measures could bring privacy hazards. On account of this, the released fairness measures should be computed based on privacy-preserving $\tilde{D}$, which, however, would influence and distort the measured bias $\varepsilon$. In this section, we show that by knowing the fidelity constraints to $\mathcal{M}$, we are able to characterize and bound the distortion of the measured fairness/bias.

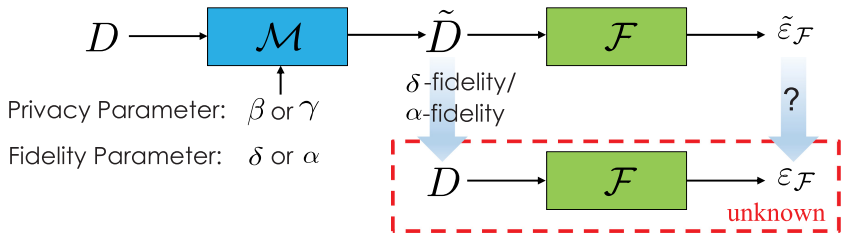


Fig. 3. A depiction of how fidelity of $\tilde{D}$ can aid in characterizing the difference between the measured bias $\tilde{\varepsilon}_{\mathcal{F}}$ and the true bias $\varepsilon_{\mathcal{F}}$, where $\mathcal{F}$ denotes a (general or specific) set of fairness definitions on which bias is computed. In Section 3, we have seen that releasing $D$ and $\varepsilon_{\mathcal{F}}$ can cause privacy leakage and thus should be prohibited, so that $\varepsilon_{\mathcal{F}}$ is unknown to the public. However, if the fidelity parameter used in $\mathcal{M}$ is known, we are able to characterize $\varepsilon_{\mathcal{F}}$ by $\tilde{\varepsilon}_{\mathcal{F}}$ based on Lemma 2.

Figure 3 is a representative illustration of the true bias $\varepsilon_{\mathcal{F}}$ and the measured bias $\tilde{\varepsilon}_{\mathcal{F}}$ influenced by $\mathcal{M}$, where $\mathcal{F}$ denotes a (general or specific) set of fairness definitions on which bias is computed. Since the true decision mapping $D$ should not be released and utilized to compute fairness measures, the true bias computed based on $D$ is generally unknown. A natural question may arise: by knowing $\tilde{\varepsilon}_{\mathcal{F}}$, and the degree of fidelity of $\tilde{D}$ ($\delta$ or $\alpha$), what can we know about $\varepsilon_{\mathcal{F}}$? The following lemma answers the question: if the maximum distortion from $\tilde{D}$ to $D$ is known, the maximum distortion from $\tilde{\varepsilon}_{\mathcal{F}}$ to $\varepsilon_{\mathcal{F}}$ can be known, and thus the range of $\varepsilon_{\mathcal{F}}$ can be known.

Lemma 2. *Let $\mathcal{F}_{tv}$ denote the set of all total-variation-based fairness definitions, and $\mathcal{F}_{rm}$ the set of all relative-metric-based fairness definitions (see Appendix A). Given $D \xrightarrow{\mathcal{M}} \tilde{D}$, if $\mathcal{M}$ satisfies $\delta$-fidelity, we can guarantee the measured bias $\tilde{\varepsilon}_{\mathcal{F}_{tv}}$ satisfies*

$$|\tilde{\varepsilon}_{\mathcal{F}_{tv}} - \varepsilon_{\mathcal{F}_{tv}}| \leq \min\{2(1-\delta), 1\}. \tag{14}$$

*If $\mathcal{M}$ satisfies $\alpha$-fidelity, we can guarantee the measured bias $\tilde{\varepsilon}_{\mathcal{F}_{rm}}$ satisfies*

$$|\tilde{\varepsilon}_{\mathcal{F}_{rm}} - \varepsilon_{\mathcal{F}_{rm}}| \leq \min\{-2\log\alpha, 1\}. \tag{15}$$

Proof. By applying the *reverse triangle inequality* [85], the results trivially follow. □

## 6 PRIVACY-FIDELITY TRADE-OFF

Strong privacy perturbation could cause serious distortion on the announced information including decision rules and fairness measures, and thus a privacy protection scheme should preserve privacy while guaranteeing a certain degree of fidelity to the announced information; this turns out a privacy-fidelity trade-off problem. In this section, we mathematically formulate the trade-off problem, and revisit existing algorithms that can efficiently solve this problem.

### 6.1 Optimization Formulation

We describe the privacy-fidelity trade-off problem in the following: given fidelity constraints, we would like to find the greatest privacy (the smallest $\beta$) that we can achieve. The problem is mathematically formulated as follow. For conciseness, we omit the subscript of all probability measures and simply write, e.g., $P(\mathbf{x})$ instead of $P_X(\mathbf{x})$.

$$\underline{\textbf{OPT}}(\mathcal{R}_X \times \mathcal{A}): \tag{OPT}$$

$$\min_{\tilde{D}} \quad \beta \tag{16a}$$

$$\text{s.t.} \quad \frac{P(\mathbf{x})\tilde{D}_a(\mathbf{x})}{\sum_{\mathbf{x}' \in T_{x_\mathcal{U}}} P(\mathbf{x}')\tilde{D}_a(\mathbf{x}')} \leq \beta\,, \quad \forall \mathbf{x} \in \mathcal{R}_X, \forall a \in \mathcal{A} \tag{16b}$$

$$\tilde{D}_a(\mathbf{x}) \leq \tilde{D}_a(\mathbf{x})_{\max}\,, \ \forall \mathbf{x} \in \mathcal{R}_X, \forall a \in \mathcal{A} \tag{16c}$$

$$\tilde{D}_a(\mathbf{x}) \geq \tilde{D}_a(\mathbf{x})_{\min}\,, \ \forall \mathbf{x} \in \mathcal{R}_X, \forall a \in \mathcal{A} \tag{16d}$$

$$\tilde{D}_a(\mathbf{x}) \geq 0\,, \forall \mathbf{x} \in \mathcal{R}_X, \forall a \in \mathcal{A} \tag{16e}$$

$$\sum_{a \in \mathcal{A}} \tilde{D}_a(\mathbf{x}) = 1\,, \ \forall \mathbf{x} \in \mathcal{R}_X\,. \tag{16f}$$

The first constraint in (16b) is the privacy constraint $\beta$-Maximum Confidence defined in Definition 4 and Lemma 1, and the last two constraints in (16e) and (16f) are probability distribution conditions. The second and the third constraints in (16c) and (16d) are fidelity constraints introduced in (11). Its corresponding representations for $\delta$- or $\alpha$-fidelity can be found in (12) and (13), respectively. The objective in (16a) is to find the minimal $\beta$ subject to the feasibility of $\tilde{D}$ based on the above-mentioned constraints. *A careful observation of the* ***optimization problem*** (OPT) *will reveal that it is an equivalent formulation of a generalized* ***linear fractional programming (LFP)*** *problem [94].*

### 6.2 Drawbacks of Existing Methods to Solve Generalized LFP Problems

It has been known that a generalized LFP is quasi-convex and not reducible to a **linear programming (LP)** problem; however, it can be solved as a sequence of LP feasibility problems [16], i.e., solving numerous sub-level LP problems iteratively according to the bisection method. By efficient algorithms such as interior point method, the solution of an LP problem can be obtained in pseudo-polynomial time $O(\frac{n^3}{\log n}L)$ [9], where $n$ is the number of variables, $L$ the input size, i.e., the length of the binary coding of the input data to represent the problem, which is roughly proportional to the number of constraints. For our problem, based on (16b)–(16f), it is clear that the number of constraints is proportional to $|\mathcal{R}_X \times \mathcal{A}|$, which is the number of variables $n = |\tilde{D}_a(\mathbf{x})|$, exponential in the number of input attributes $K$. For example, suppose the cardinality for each input attribute is consistent, e.g., $|X_k| = l$, $\forall k = 1, \ldots, K$, we have $n = |\mathcal{R}_X \times \mathcal{A}| = 2 \cdot l^K$ and roughly $4n = 8 \cdot l^K$ constraints. Even for a relatively small example, e.g., a binary decision process with $K = 10$ input attributes, each with $l = 5$ possible values, we have $n \approx 2 \cdot 10^7$ variables and $\approx 8 \cdot 10^7$ constraints for each sub-level LP problem, with time complexity $O(\frac{n^4}{\log n})$, i.e., not tractable on typical machines (e.g., as reported in [64], 'spal_004' with 10203 rows (w.r.t. constrains) and 321696 columns (w.r.t. variables) can encounter out of memory or timeout (>25,000 seconds) issue on a Linux-PC with a 4GHz i7-4790K CPU and 32GB RAM). To solve a generalized LFP problem, we need to solve such a huge sub-level LP problem iteratively. In practice, ML algorithms may require nontrivial amounts of attributes to aid in decision-making; therefore, without an efficient solver, the privacy-fidelity trade-off problem could be intractable and the feasibility of the associated privacy protection scheme could be dramatically reduced. *In this regard, it is crucial and essential to propose a more efficient method to solve the proposed privacy-fidelity trade-off optimization problem.*

## 7 LINEAR-TIME OPTIMAL PRIVACY SOLUTION

In this section, we analyze the **optimization problem** (OPT), reveal its important properties, and propose efficient methods to solve it. We first investigate the *decomposability* of the optimization problem, i.e, whether the problem can be decomposed into several smaller sub-problems for efficient solving. We find (OPT) decomposable and can be solved using a divide-and-conquer approach. In addition, we propose *closed-form solutions* for each optimization sub-problem. The optimization problem can thus be solved very efficiently by solving multiple independent sub-problems in *linear time*. Moreover, analysis insights into the optimal solutions are also provided in this section.

### 7.1 Decomposability

In the following, we show that the optimization problem can actually be decomposed into numerous small sub-problems and thus can be solved more efficiently. An optimization problem is *separable* or *trivially parallelizable* if the variables can be partitioned into disjoint subvectors and each constraint involves only variables from one of the subvectors [17]. By observing (i) each constraint in (16c), (16d), and (16e) involves only a single variable $\tilde{D}_a(\mathbf{x})$, (ii) each constraint in (16f) involves a set of variables $\{\tilde{D}_a(\mathbf{x}) \mid \forall a \in \mathcal{A}\}$, and (iii) each constraint in (16b) involves a set of variables $\{\tilde{D}_a(\mathbf{x}) \mid \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}\}$, we notice that any variable $\tilde{D}_a(\mathbf{x})$ is a *complicating variable* in $T_{\mathbf{x}_\mathcal{U}} \times \mathcal{A}$ but is irrelevant to any other variables outside the QID group $T_{\mathbf{x}_\mathcal{U}}$. Hence, (16b)–(16f) are *complicating constraints* within a tuple but *separable constraints* among tuples. (OPT) can thus be decomposed into multiple smaller sub-problems; each focuses on a particular QID group only. Let $h(\tilde{D}_a(\mathbf{x}), \beta) \geq 0$ be the affine function representing all linear inequality constraints (16b)–(16e). An optimization sub-problem can thus be expressed as follows.

$$\underline{\textbf{OPT-SUB}}(T_{\mathbf{x}_\mathcal{U}} \times \mathcal{A}) : \tag{OPT-Sub}$$

$$\min_{\tilde{D}} \quad \beta \tag{OBJ-Sub}$$

$$\text{s.t.} \quad h(\tilde{D}_a(\mathbf{x}), \beta) \geq 0\,, \ \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}, \forall a \in \mathcal{A} \tag{INEQ-Sub}$$

$$\sum_{a \in \mathcal{A}} \tilde{D}_a(\mathbf{x}) = 1\,, \ \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}\,. \tag{EQ-Sub}$$

(OPT) is then equivalent to the *master problem* below.

$$\underline{\textbf{OPT-MASTER}}(\mathcal{R}_X \times \mathcal{A}) : \tag{OPT-MS}$$

$$\min_{\tilde{D}} \quad \beta \tag{OBJ-MS}$$

$$\text{s.t.} \quad (\text{INEQ-Sub}(T_{\mathbf{x}_\mathcal{U}} \times \mathcal{A}))\,, \ \forall T_{\mathbf{x}_\mathcal{U}} \subseteq \mathcal{R}_X \tag{INEQ-MS}$$

$$(\text{EQ-Sub}(T_{\mathbf{x}_\mathcal{U}} \times \mathcal{A}))\,, \ \forall T_{\mathbf{x}_\mathcal{U}} \subseteq \mathcal{R}_X\,. \tag{EQ-MS}$$

Lemma 3. *Let $\beta^*_{T_{\mathbf{x}_\mathcal{U}}}$ denote the optimal value of a sub-problem* (OPT-Sub)*, $\beta^*$ the optimal value of* (OPT)*. We have $\beta^* = \max_{T_{\mathbf{x}_\mathcal{U}} \subseteq \mathcal{R}_X} \beta^*_{T_{\mathbf{x}_\mathcal{U}}}$.*

Proof. Since (OPT) is a generalized LFP (in an equivalent formulation), according to OPT-MS, the result trivially follows. □

The Lemma above basically tells us that given the same fidelity constraints, the *overall* highest privacy guarantee $\beta^*$ is the largest $\beta^*_{T_{\mathbf{x}_\mathcal{U}}}$ among all sub-problems, i.e., the *weakest* optimal privacy guarantee among all QID groups.

### 7.2 Solution Properties

According to the decomposability of the optimization problem, in the following, we only need to focus on solving an **optimization sub-problem** (OPT-Sub). To characterize the privacy-fidelity trade-off, we are particularly interested in where the trade-off starts and ends. In this section, we propose lemmas addressing the above question.

Before introducing the lemmas, we first define a useful quantity which will be further utilized to characterize the trade-off.

*Definition 7 (Maximum Posterior Confidence).* Given an optimization sub-problem (OPT-Sub) and 1-fidelity (100% faithfulness) requirement, i.e., $\alpha = \delta = 1$ and $\tilde{D} = D$, the highest confidence that an adversary can have on inferring any sensitive information from any decision outcome, denoted by $C^*$, is $C^* \triangleq Conf(X_{\mathcal{U}} = \mathbf{x}_{\mathcal{U}}, A \xrightarrow{P_X, D} X_{\mathcal{S}}) = \max_{a, x_{\mathcal{S}}}\{conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}})\}$.

Lemma 4. *An* (OPT-Sub) *has the 1-fidelity solution* $\tilde{D}_a(\mathbf{x}) = D_a(\mathbf{x}), \forall \mathbf{x} \in T_{\mathbf{x}_{\mathcal{U}}}, \forall a \in \mathcal{A}$, *iff* $\beta \geq C^*$.

Proof. Please refer to Appendix E for detailed proof. We provide intuitive explanation as proof sketch here. Since the highest confidence that an adversary can have ($C^*$) is lower than the privacy requirement (the confidence threshold $\beta$), it is safe to release $D$ directly, i.e., $\tilde{D} = D$ with perfect fidelity. On the other hand, as long as $C^*$ is greater than $\beta$, releasing $D$ violates privacy requirement and cannot be a feasible solution. □

*Lemma Insight* - Lemma 4 tells us when $\beta \geq C^*$, there is no trade-off between privacy and fidelity: as long as $\beta$ is greater than $C^*$, increasing the strength of privacy (decreasing $\beta$) would not cause degradation in fidelity. In other words, alone the strength of privacy from low to high (i.e., $\beta$ from 1 to 0), the trade-off between privacy and fidelity starts when $\beta$ is right below $C^*$. The next lemma will tell us the end of this trade-off region.

Lemma 5. *For* $\alpha = \delta = 0$, *i.e., fidelity constraints are trivialized or not presented, an* (OPT-Sub) *has feasible solutions if and only if (iff)* $\beta \geq \beta_{min} \triangleq \max_{\mathbf{x} \in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x}|T_{\mathbf{x}_{\mathcal{U}}})$. *In other words, there exists privacy limit, the strongest privacy that we can have, based on the adversarial settings in Section 4.2.*

Proof. Please refer to Appendix F for detailed proof. We provide intuition behind this lemma as proof sketch here. The privacy limit $\max_{\mathbf{x} \in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x}|T_{\mathbf{x}_{\mathcal{U}}})$ is the greatest conditional probability over the tuple,[5] which is actually the highest possible inference confidence of an adversary *before seeing an ATR*. It is the *baseline confidence*, which merely utilizes knowledge of public record $\mathbf{x}_{\mathcal{U}}$ and side-information $P(\mathbf{x})$ on an inference channel $\langle \mathbf{x}_{\mathcal{U}} \xrightarrow{P_X} x_{\mathcal{S}} \rangle$. Since an ATR does not contribute to such an inference channel, an associated privacy protection scheme is not able to help further reduce this baseline confidence. While achieving such a privacy limit, an ATR basically reveals zero useful information to the public. □

*Lemma Insight* - Lemma 4 and 5 tell us the start and the end of the privacy-fidelity trade-off region along $\beta$. Next, we show that the end point can never happen before the starting point.

Lemma 6. $C^* \geq \beta_{min}$.

Proof. Please refer to Appendix G for detailed proof. We first provide the intuition of the lemma as a proof sketch. The intuition here is very straightforward: the maximum posterior confidence can never be lower than the maximum prior confidence (*prior vulnerability cannot exceed posterior vulnerability* in [66]). Equality holds when the revealed information is completely useless. □

[5] Since $\forall \mathbf{x} \in T_{\mathbf{x}_{\mathcal{U}}}$, $\mathbf{x}_{\mathcal{U}}$ is the same, $P(\mathbf{x}|T_{\mathbf{x}_{\mathcal{U}}})$ is also $P(x_{\mathcal{S}}|T_{\mathbf{x}_{\mathcal{U}}})$, the conditional distribution over all sensitive records.

*Lemma Insight* - According to Lemma 4, when $\beta \in [C^*, 1]$, the true decision mapping $D$ can be safely released without perturbation (1-fidelity). Lemma 5 tells us when fidelity constraints are not imposed (0-fidelity), the feasible privacy region is $\beta \in [\beta_{\min}, 1]$. Moreover, based on Lemma 6, the region $[\beta_{\min}, C^*]$ is always non-empty. Clearly, this is the region where we trade off fidelity for privacy. Next, we propose our main theorem to characterize the trade-off in this region.

### 7.3 Optimal Privacy and Solutions

In the following, we propose our main theorem, which provides the optimal-privacy solutions to the optimization sub-problem (OPT-Sub) in a closed-form expression, in terms of fidelity. Given fidelity constraints, the proposed closed-form expression yields the optimal privacy value, and thus can be utilized to analytically characterize privacy-fidelity trade-off.

Theorem 1. *Consider an optimization sub-problem (OPT-Sub) for a QID group $T_{\mathbf{x}_\mathcal{U}}$, in which we seek for the strongest privacy guarantee given fidelity constraints. For a decision outcome $a$, define*

$$\mathbf{x}^a \triangleq \arg\max_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}) \tilde{D}_a(\mathbf{x})_{min},$$
$$b(\mathbf{x}) = P(\mathbf{x}) - \beta \textstyle\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}'),$$
$$\tilde{D}_a(\mathbf{x})_{max'} \triangleq \tfrac{1}{P(\mathbf{x})} \min\{P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{max}, P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{min}\},$$
$$\tilde{D}_a(\mathbf{x})_{min'} \triangleq \tfrac{1}{P(\mathbf{x})} \max\{P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{min}, P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{min} + b(\mathbf{x})\}.$$

*For binary decisions, i.e., $a \in \mathcal{A} = \{0, 1\}$, the optimal privacy $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \max\{\beta_0, \beta_1, \beta_p\}$, where*

$$\beta_0 = \frac{P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{min}}{P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{min} + \sum_{\mathbf{x} \neq \mathbf{x}^0, \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{max'}},$$
$$\beta_1 = \frac{P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{min}}{P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{min} + \sum_{\mathbf{x} \neq \mathbf{x}^1, \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{max'}},$$
$$\beta_p = \frac{P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{min} + P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{min}}{\sum_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x})},$$

*and the corresponding optimal privacy solutions are*

$$\textit{When } \beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_0: \quad \tilde{D}_0(\mathbf{x}) = \tilde{D}_0(\mathbf{x})_{max'}, \ \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$$
$$\textit{When } \beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_1: \quad \tilde{D}_1(\mathbf{x}) = \tilde{D}_1(\mathbf{x})_{max'}, \ \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$$
$$\textit{When } \beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p: \quad \tilde{D}_a(\mathbf{x}^a) = \tilde{D}_a(\mathbf{x}^a)_{min}, \ \forall a \in \mathcal{A}$$
$$\textstyle\sum_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x})\tilde{D}_a(\mathbf{x}) = \frac{1}{\beta_p} P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{min}, \ \forall a \in \mathcal{A}$$
$$\tilde{D}_a(\mathbf{x})_{min'} \leq \tilde{D}_a(\mathbf{x}) \leq \tilde{D}_a(\mathbf{x})_{max'}, \ \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}, \forall a \in \mathcal{A}.$$

*When $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p$ and $|T_{\mathbf{x}_\mathcal{U}}| > 3$, we have multiple solutions.*

Proof. We refer readers to Appendix H for the detailed proof. □

*Practical Implications of Theorem* - Based on Theorem 1, the optimal privacy guarantee $\beta^*_{T_{\mathbf{x}_\mathcal{U}}}$ for each QID group can be computed analytically (in closed form), and based on Lemma 3, the overall strongest privacy guarantee is the largest $\beta^*_{T_{\mathbf{x}_\mathcal{U}}}$ among all QID groups. This is particularly useful and practical in releasing ATRs - given any value of tolerable distortion, we can now easily obtain

the optimal privacy value without the need of solving an optimization problem, which can then be applied to aid in determining the desired trade-off between privacy and fidelity.

*Linear Time Justification of Algorithm 1* - The net time to achieve a solution to (OPT) is a function of the number of sub-problems - each of which is solved via Algorithm 1 in *linear-time* in the number of records $n$. Given an optimization sub-problem, the number of records $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$, i.e., $|T_{\mathbf{x}_\mathcal{U}}| \triangleq n$ is equal to the number of sensitive attribute values, as all records in a QID group $T_{\mathbf{x}_\mathcal{U}}$ have the same public record $\mathbf{x}_\mathcal{U}$. To see that Algorithm 1 is in $O(n)$, it is first worth noting that, based on Theorem 1, the time complexity of computing $\mathbf{x}^a$ and $b(\mathbf{x})$ are in $O(n)$; consequently, the time complexity of computing $\tilde{D}_a(\mathbf{x})_{\max'}$ and $\tilde{D}_a(\mathbf{x})_{\min'}$ are in $O(1)$. Given these complexities, it is clear that lines 1 to 4 in Algorithm 1 are in $O(n)$; all lines from line 5 to line 15, except line 13, are in $O(1)$; function Allocation called in line 13 is in $O(n)$ - since lines 18 to 19, as well as line 20, are in $O(n)$, line 21 and 22 are in $O(1)$, and lines 23 to 27 are in $O(n)$). Therefore, the time complexity of Algorithm 1 is in $O(n)$. *By using multi-threaded coding structures solving "parallelizable" sub-problems via Algorithm 1*, (OPT) can be solved in $O(n)$.

### 7.4 Theorem Insights on Achieving Solution Optimality

In this section, we provide some insights into the optimal-privacy solutions subject to fidelity constraints in Theorem 1.

An important observation is that the optimal-privacy candidate values ($\beta_0$, $\beta_1$, and $\beta_p$) and the inference confidence (left-hand-side of (16b)) are fully characterized by $P(\mathbf{x})\tilde{D}_a(\mathbf{x})$ pairs of product, which are the announced *joint probabilities* $\tilde{P}_{X,A}(\mathbf{x}, a)$ representing the portion of population with input record $\mathbf{x}$ receiving decision $a$, bounded within ranges $[P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\min}, P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\max}]$ due to fidelity constraints. Solving (OPT-Sub) to obtain optimal-privacy solution is thus equivalent to "tune" those pairs of product within the allowed range, for all inputs $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$ and outputs $a \in \mathcal{A}$, to minimize the maximal possible inference confidence $\beta$ of an adversary. Particularly, from Theorem 1, it turns out that for each decision outcome instance $a$, the term $P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{\min} = \max_{\mathbf{x}\in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\min}$, the maximum of the lower bounds of the allowed ranges over $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$, plays a crucial role in solving (OPT-Sub). Next, we show that the optimal-solution for $\mathbf{x} = \mathbf{x}^a$ can only be the minimum of its allowed range.

Corollary 1. $\tilde{D}_a(\mathbf{x}^a)_{min} = \tilde{D}_a(\mathbf{x}^a)_{min'} = \tilde{D}_a(\mathbf{x}^a)_{max'}, \forall a \in \{0, 1\}$.

Proof. By definitions of $\mathbf{x}^a$ and $\tilde{D}_a(\mathbf{x})_{\max'}$, the result $\tilde{D}_a(\mathbf{x}^a)_{\min} = \tilde{D}_a(\mathbf{x}^a)_{\max'}$ trivially follows. Based on Lemma 5, we have $b(\mathbf{x}) = P(\mathbf{x}) - \beta_p \sum_{\mathbf{x}'} P(\mathbf{x}') \leq 0$, and thus by plugging $\mathbf{x} = \mathbf{x}^a$ into $\tilde{D}_a(\mathbf{x})_{\min'}$, we obtain $\tilde{D}_a(\mathbf{x}^a)_{\min} = \tilde{D}_a(\mathbf{x}^a)_{\min'}$. □

The effective lower limits $P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\min'}$ and the effective upper limits $P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\max'}$ represent the feasible region where the fidelity constraints and privacy constraints intersect. Based on Corollary 1, the effective upper and lower limits of $\mathbf{x}^a$ are equal, which implies that when $\mathbf{x} = \mathbf{x}^a$, the only possible value for the optimal-privacy solution is $\tilde{D}_a(\mathbf{x}^a)_{\min}$. From Theorem 1, we can see that this is true *for all the three cases*. It is worth noting that the cases $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_0$ and $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_1$ are symmetric cases (by swapping 0's and 1's), so *we only have two representative cases:* $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_a$, $a \in \{0, 1\}$, and $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p$.

*7.4.1 Representative Case 1.* When $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_a$, $a \in \{0, 1\}$, the allowed ranges of all $P(\mathbf{x})\tilde{D}_a(\mathbf{x})$ pairs are imposed by an *additional upper limit* $P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{\min}$ caused by privacy constraints. In these cases, effective upper limits are the minimum of the original upper limits $P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\max}$ and the threshold, formally, $P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\max'} = \min\{P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\max}, P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{\min}\}$. According to

**ALGORITHM 1:** Optimal Privacy Protection Scheme

**Input:** $P(\mathbf{x})$, $T_{\mathbf{x}_\mathcal{U}}$, $\tilde{D}_a(\mathbf{x})_{\min}$, $\tilde{D}_a(\mathbf{x})_{\max}$
**Output:** $\tilde{D}_a(\mathbf{x})$, $\forall a, \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$

**for** $a \in \{0, 1\}$ **do**
  find $\mathbf{x}^a$
  **for all** $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$ **do**
    compute $\tilde{D}_a(\mathbf{x})_{\max'}$
compute $\beta_0$, $\beta_1$, $\beta_p$, and $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} \leftarrow \max\{\beta_0, \beta_1, \beta_p\}$
**if** $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_0$ **then**
  $\tilde{D}_0(\mathbf{x}) \leftarrow \tilde{D}_0(\mathbf{x})_{\max'}$
  $\tilde{D}_1(\mathbf{x}) \leftarrow 1 - \tilde{D}_0(\mathbf{x})_{\max'}$
**else if** $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_1$ **then**
  $\tilde{D}_1(\mathbf{x}) \leftarrow \tilde{D}_1(\mathbf{x})_{\max'}$
  $\tilde{D}_0(\mathbf{x}) \leftarrow 1 - \tilde{D}_1(\mathbf{x})_{\max'}$
**else if** $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p$ **then**
  $\tilde{D}_1(\mathbf{x}) \leftarrow$ Allocation()
  $\tilde{D}_0(\mathbf{x}) \leftarrow 1 - \tilde{D}_1(\mathbf{x})$
**return** $\tilde{D}_a(\mathbf{x})$, $\forall a, \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$

**function** Allocation( )
  **for all** $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}, \mathbf{x} \neq \mathbf{x}^1, \mathbf{x}^0$ **do**
    compute $\tilde{D}_1(\mathbf{x})_{\min'}$
  resid $\leftarrow$ RHS of (22) $- \sum_{\mathbf{x} \neq \mathbf{x}^1, \mathbf{x}^0} P(\mathbf{x}) \tilde{D}_1(\mathbf{x})_{\min'}$
  $\tilde{D}_1(\mathbf{x}^1) \leftarrow \tilde{D}_1(\mathbf{x}^1)_{\min}$
  $\tilde{D}_1(\mathbf{x}^0) \leftarrow 1 - \tilde{D}_0(\mathbf{x}^0)_{\min}$
  **for all** $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}, \mathbf{x} \neq \mathbf{x}^1, \mathbf{x}^0$ **do**
    capacity $\leftarrow \tilde{D}_1(\mathbf{x})_{\max'} - \tilde{D}_1(\mathbf{x})_{\min'}$
    allocation $\leftarrow \min\{\frac{\text{resid}}{P(\mathbf{x})}, \text{capacity}\}$
    $\tilde{D}_1(\mathbf{x}) \leftarrow \tilde{D}_1(\mathbf{x})_{\min'} +$ allocation
    resid $\leftarrow$ resid $- P(\mathbf{x}) \cdot$ allocation
  **return** $\tilde{D}_1(\mathbf{x})$, $\forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$

Theorem 1, when $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_a$, the corresponding optimal-privacy solution is simply the effective upper limit $\tilde{D}_a(\mathbf{x}) = \tilde{D}_a(\mathbf{x})_{\max'}$, $\forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$.

An illustration that aids in understanding the intuition behind Theorem 1 is shown in Figure 4, in which joint probabilities for $a = 1$ and $\forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$ are depicted for the case $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_1$. For conciseness, let $m$ denote $|T_{\mathbf{x}_\mathcal{U}}|$,[6] $p_i = P(\mathbf{x}_i)$, and $\tilde{d}_i = \tilde{d}(\mathbf{x}_i) = \tilde{D}_1(\mathbf{x}_i)$, $\forall i = 1, \ldots, m$. The yellow spots in Figure 4 denote the true joint probabilities $p_i d_i$, and the regions indicated by grey arrows interpret the allowed perturbation ranges $[p_i \tilde{d}_{i\min}, p_i \tilde{d}_{i\max}]$. In this example, the maximum of the lower limits is $p_6 \tilde{d}_{6\min}$, i.e., $\mathbf{x}^1 = \mathbf{x}_6$. The value $p_6 \tilde{d}_{6\min}$ serves as a threshold (the blue dash line), imposing an upper limit on all perturbation ranges. The output of the optimal-privacy solution is

[6]Since $T_{\mathbf{x}_\mathcal{U}}$ is the range of $(\mathbf{x}_\mathcal{U}, X_S)$, $m$ represents the number of distinct sensitive records in the tuple.

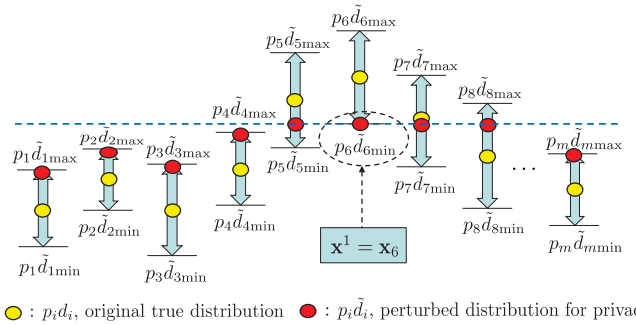


Fig. 4. A representative illustration for changes of joint probabilities caused by the optimal-privacy scheme.

denoted by the red spots, which take values from the effective upper bounds $\min\{p_i\tilde{d}_{i\max}, p_6\tilde{d}_{6\min}\}$, $\forall i$. Here we get a clear insight into the optimal-privacy solutions for a QID sub-group $T_{\{\mathbf{x}_\mathcal{U}, a\}}$: the optimality is achieved by *flattening the joint distribution* $P(\mathbf{x})\tilde{D}_a(\mathbf{x})$ *over all* $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$ *as much as possible* subject to the allowed perturbation range imposed by fidelity constraints. By flattening the joint distribution, the (Bayesian) posterior distribution over distinct sensitive values seen by an adversary becomes more uniform, and hence the maximal inference confidence is reduced.

Define $\overline{a}$ the complement of $a$, e.g., $\overline{a} = 0$ if $a = 1$. For binary decisions, the optimal-privacy scheme for a QID group $T_{\mathbf{x}_\mathcal{U}}$ is also privacy optimal for a sub-group $T_{\{\mathbf{x}_\mathcal{U}, a\}}$ when the joint distribution of $T_{\{\mathbf{x}_\mathcal{U}, \overline{a}\}}$ is *much* "flatter" (more private) than $T_{\{\mathbf{x}_\mathcal{U}, a\}}$. In other words, the optimal-privacy solution flattens the least private distribution as much as possible; although this might influence the other (the much more private) one and cause it to be less private,[7] as long as its maximal inference confidence is less than $\beta_a$, the optimal privacy for the entire QID group is dominated by $\beta_a$, and thus the optimal privacy scheme for the sub-group $T_{\{\mathbf{x}_\mathcal{U}, a\}}$ is the optimal privacy scheme for the entire QID group.

*7.4.2 Representative Case 2.* When neither distribution is *much* flatter than the other one, making one sub-group highly private causes the other one's privacy to degrade, i.e., none of the optimal schemes for any QID sub-group can be optimal for the entire QID group. In such a case, both sub-groups need to find a "balanced point" at which both sub-groups are equally private. Such a balanced privacy value for the maximal inference confidence for two sub-groups is denoted by $\beta_p$ in Theorem 1, representing the optimal privacy for the QID group. As shown in Theorem 1, in general, we have multiple solutions to achieve this balanced privacy value. This is because, in such a case, the optimality of privacy is guaranteed if (i) $\tilde{D}_a(\mathbf{x}^a) = \tilde{D}_a(\mathbf{x}^a)_{\min}$, $\forall a$, and (ii) the following two equalities hold

$$\sum_{\mathbf{x}} P(\mathbf{x})\tilde{D}_0(\mathbf{x}) = \tfrac{1}{\beta_p} P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{\min}, \tag{19}$$

$$\sum_{\mathbf{x}} P(\mathbf{x})\tilde{D}_1(\mathbf{x}) = \tfrac{1}{\beta_p} P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min}. \tag{20}$$

While in the following, we show that the above two equalities are equivalent, i.e., one implies the other.

Corollary 2. *When* $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p$, (19) *implies* (20)*, and vice versa.*

[7]Based on (EQ-Sub), any changes made to $\tilde{D}_a(\mathbf{x})$ will also change $\tilde{D}_{\overline{a}}(\mathbf{x})$.

Proof. Recall $\beta_p$ from Theorem 1, we have

$$\sum_{\mathbf{x}} P(\mathbf{x}) = \frac{1}{\beta_p} P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min} + \frac{1}{\beta_p} P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{\min}. \tag{21}$$

Since $\tilde{D}_0(\mathbf{x}) + \tilde{D}_1(\mathbf{x}) = 1$, by subtracting (19) from (21), we obtain (20). Similarly, by subtracting (20) from (21), we obtain (19). □

Therefore, to compute an optimal solution when $\beta^*_{T_{\mathbf{x}\mathcal{U}}} = \beta_p$, we only need to solve (20). Since $\tilde{D}_0(\mathbf{x}) + \tilde{D}_1(\mathbf{x}) = 1$, and $\tilde{D}_a(\mathbf{x}^a) = \tilde{D}_a(\mathbf{x}^a)_{\min}$, $\forall a$, in general we only have $m - 2$ variables (see Remark 4), and based on (21), equality (20) is equivalent to

$$\sum_{\mathbf{x}\neq\mathbf{x}^0,\mathbf{x}^1} P(\mathbf{x})\tilde{D}_1(\mathbf{x}) = \left(\frac{1-2\beta_p}{\beta_p}\right) P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min} - b(\mathbf{x}^0). \tag{22}$$

When $\beta^*_{T_{\mathbf{x}\mathcal{U}}} = \beta_p$, the right-hand-side (RHS) of (22) is strictly bounded by $[\sum_{\mathbf{x}\neq\mathbf{x}^0,\mathbf{x}^1} P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\min'}$, $\sum_{\mathbf{x}\neq\mathbf{x}^0,\mathbf{x}^1} P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max'}]$, which implies there always exists a feasible solution for (20). When $m > 3$, since the number of variables to solve $(m-2)$ is greater than the number of equation (one, which is (22)), an optimal solution, in general, is not unique.

*Remark 4.* For the special case $\mathbf{x}^0 = \mathbf{x}^1$, we have $m-1$ variables to solve. Such a case can happen when the population of a certain record dominates its corresponding QID group. When this is the case, the prior (distribution) knowledge provides very high (baseline) confidence on inferring this record. In particular, for such a case, we must have $\beta_p = \beta_{\min}$. If $\beta_p > \beta_a$, $\forall a$, i.e., $\beta^*_{T_{\mathbf{x}\mathcal{U}}} = \beta_p$, this becomes trivial: according to Lemma 5 and its following discussion, the announced ATR can only provide trivial information to achieve this lowest-possible baseline confidence.

## 8 NUMERICAL EXAMPLES

In the previous section, we proposed lemmas characterizing important properties about privacy-fidelity trade-off, a theorem providing closed-form optimal solutions for the trade-off problem, and insights into the optimal solutions for both the representative cases. In this section, we provide numerical examples to demonstrate *privacy-fidelity trade-off regions*, and aid in understanding the properties of the trade-off regions and the insights into the optimal solutions for both the representative cases. Without loss of generality, we reuse the same examples demonstrated in Section 3 showcasing how the proposed linear-time optimal-privacy scheme (Algorithm 1) can be applied in practice to solve the problem, as long as there is no privacy preference among sensitive attribute values.

Consider Table 1 again but for a smaller size population {12, 5, 3, 9, 7, 4} (first column of the table) for ease of demonstration, and let $\mathbf{x}_i$ denote the record of the $i$-th row, $i = 1, \ldots, 6$. Suppose an announced ATR needs to satisfy a pre-determined fidelity constraint $\delta = 0.9$ (90%-fidelity), and we would like to preserve the data subjects' privacy as much as possible subject to the fidelity constraint.

First, consider the QID group of female $T_{\mathbf{x}\mathcal{U}=\{\text{F}\}}$, i.e., the tuple of records $T_{\{\text{F}\}} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$. Based on lines 1 to 4 in Algorithm 1, we first need to determine $\mathbf{x}^a$ and $\tilde{D}_a(\mathbf{x})_{\max'}$, $\forall a \in \{0, 1\}$, $\forall \mathbf{x} \in T_{\{\text{F}\}}$. Detailed computations are demonstrated in Remark 5, and the computed results are presented in Table 4; from which, we observe that $\mathbf{x}^1 = \mathbf{x}_3$ and $\mathbf{x}^0 = \mathbf{x}_1$ (see Remark 5 for details as well).

Table 4. Detailed Inputs and Computations of the Provided Numerical Example

| | Inputs | | | | | | | Computations | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{x}$ | $P(\mathbf{x})$ | $D_1(\mathbf{x})$ | $D_0(\mathbf{x})$ | $\tilde{D}_1(\mathbf{x})_{\min}$ | $\tilde{D}_1(\mathbf{x})_{\max}$ | $\tilde{D}_0(\mathbf{x})_{\min}$ | $\tilde{D}_0(\mathbf{x})_{\max}$ | $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\min}$ | $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max}$ | $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max'}$ | $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\min}$ | $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\max}$ | $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\max'}$ |
| $\mathbf{x}_1$ | 0.3 | 0 | 1 | 0 | 0.1 | 0.9 | 1 | 0 | 0.03 | 0.03 | 0.27 | 0.3 | 0.27 |
| $\mathbf{x}_2$ | 0.125 | 0 | 1 | 0 | 0.1 | 0.9 | 1 | 0 | 0.0125 | 0.0125 | 0.1125 | 0.125 | 0.125 |
| $\mathbf{x}_3$ | 0.075 | 1 | 0 | 0.9 | 1 | 0 | 0.1 | 0.0675 | 0.075 | 0.0675 | 0 | 0.0075 | 0.0075 |
| $\mathbf{x}_4$ | 0.225 | 0 | 1 | 0 | 0.1 | 0.9 | 1 | 0 | 0.0225 | 0.0225 | 0.2025 | 0.225 | 0.2025 |
| $\mathbf{x}_5$ | 0.175 | 0.5 | 0.5 | 0.4 | 0.6 | 0.4 | 0.6 | 0.07 | 0.105 | 0.09 | 0.07 | 0.105 | 0.105 |
| $\mathbf{x}_6$ | 0.1 | 1 | 0 | 0.9 | 1 | 0 | 0.1 | 0.09 | 0.1 | 0.09 | 0 | 0.01 | 0.01 |

Proceeding to line 5, we compute $\beta_0$, $\beta_1$, and $\beta_p$ as follows:

$$\beta_1 = \frac{P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min}}{P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min} + \sum_{\mathbf{x}\neq\mathbf{x}^1, \mathbf{x}\in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max'}} = \frac{0.0675}{0.0675 + 0.03 + 0.0125} \approx 0.6136,$$

$$\beta_0 = \frac{P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{\min}}{P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{\min} + \sum_{\mathbf{x}\neq\mathbf{x}^0, \mathbf{x}\in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\max'}} = \frac{0.27}{0.27 + 0.125 + 0.0075} \approx 0.6708,$$

$$\beta_p = \frac{P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min} + P(\mathbf{x}^0)\tilde{D}_0(\mathbf{x}^0)_{\min}}{\sum_{\mathbf{x}\in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x})} = \frac{0.0675 + 0.27}{0.5} = 0.675,$$

and obtain $\beta^*_{T_{\{\mathbf{F}\}}} \triangleq \max\{\beta_0, \beta_1, \beta_p\} = \beta_p = 0.675$. Proceeding to lines 12 and 13, in this case we need to call function Allocation in line 17. Based on lines 18 and 19, we first need to compute

$$\begin{aligned}\tilde{D}_1(\mathbf{x}_2)_{\min'} &= \frac{1}{P(\mathbf{x}_2)} \max\{P(\mathbf{x}_2)\tilde{D}_1(\mathbf{x}_2)_{\min}, P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min} + b(\mathbf{x}_2)\} \\ &= \frac{1}{0.125} \max\{0, 0.0675 + 0.125 - (0.675)(0.5)\} = 0.\end{aligned}$$

Proceeding to line 20, since $\tilde{D}_1(\mathbf{x}_2)_{\min'} = 0$, we have

$$\begin{aligned}\texttt{resid} = \text{RHS of (22)} &= \left(\frac{1-2\beta_p}{\beta_p}\right) P(\mathbf{x}^1)\tilde{D}_1(\mathbf{x}^1)_{\min} - b(\mathbf{x}^0) \\ &= \left(\frac{-0.35}{0.675}\right)(0.075)(0.9) + (0.675)(0.5) - 0.3 = 0.0025.\end{aligned}$$

Based on lines 21 and 22, we obtain

$$\begin{aligned}\tilde{D}_1(\mathbf{x}_3) &= \tilde{D}_1(\mathbf{x}^1) = \tilde{D}_1(\mathbf{x}^1)_{\min} = \tilde{D}_1(\mathbf{x}_3)_{\min} = 0.9, \\ \tilde{D}_1(\mathbf{x}_1) &= \tilde{D}_1(\mathbf{x}^0) = 1 - \tilde{D}_0(\mathbf{x}^0)_{\min} = 1 - \tilde{D}_0(\mathbf{x}_1)_{\min} = 0.1.\end{aligned}$$

Moreover, proceeding to lines 23 to 27, we obtain

$$\begin{aligned}\texttt{capacity} &= \tilde{D}_1(\mathbf{x}_2)_{\max'} - \tilde{D}_1(\mathbf{x}_2)_{\min'} = \frac{0.0125}{0.125} - 0 = 0.1, \\ \texttt{allocation} &= \min\left\{\frac{0.0025}{0.125}, 0.1\right\} = 0.02, \\ \tilde{D}_1(\mathbf{x}_2) &= \tilde{D}_1(\mathbf{x}_2)_{\min'} + \texttt{allocation} = 0 + 0.02 = 0.02.\end{aligned}$$

We therefore obtain the optimal solution $\tilde{D}_1(\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}) = [0.1, 0.02, 0.9]$ for the QID group of female, which yields maximum confidence of 67.5% for an adversary inferring any sensitive information from any female data subject.

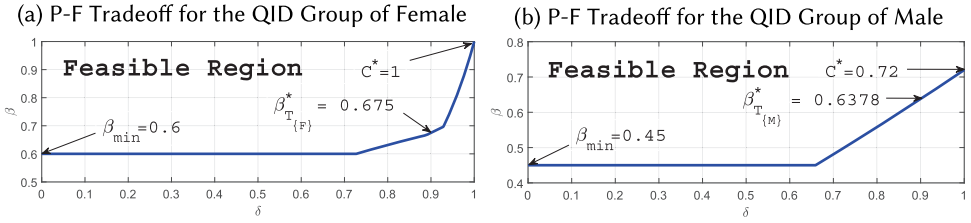


Fig. 5. Privacy-Fidelity (P-F) Tradeoffs for QID Groups of Female and Male.

We then consider the QID group for male $T_{\mathbf{x}_\mathcal{U}=\{\mathrm{M}\}}$, i.e., the tuple of records $T_{\{\mathrm{M}\}} = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$. Similarly, based on Table 4, we obtain $\mathbf{x}^1 = \mathbf{x}_6$, $\mathbf{x}^0 = \mathbf{x}_4$, and

$$\beta_1 = \frac{0.09}{0.0225 + 0.09 + 0.09} \approx 0.4444,$$
$$\beta_0 = \frac{0.2025}{0.2025 + 0.105 + 0.01} \approx 0.6378,$$
$$\beta_p = \frac{0.09 + 0.2025}{0.5} = 0.585,$$

and we get $\beta^*_{T_{\{\mathrm{M}\}}} = \max\{\beta_0, \beta_1, \beta_p\} = \beta_0 \approx 0.6378$. Based on lines 6 to 8, we obtain the optimal solution for this group $\tilde{D}_1(\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}) = 1 - \tilde{D}_0(\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\})_{\max'} = 1 - [\frac{0.2025}{0.225}, \frac{0.105}{0.175}, \frac{0.01}{0.1}] = [0.1, 0.4, 0.9]$ for QID group of male, which yields maximum confidence of 63.78% for an adversary inferring any sensitive information from any male data subject. Based on Lemma 3, the optimal-privacy $\beta^*$ for the entire dataset is $\max\{\beta^*_{T_{\{\mathrm{F}\}}}, \beta^*_{T_{\{\mathrm{M}\}}}\} = \max\{0.675, 0.6378\} = 0.675$, which is the maximum confidence for an adversary inferring any sensitive information from any data subject from this dataset, based on the announced ATR.

The optimal solution for the QID group of female is a "balanced point" between inferring the annual income of $\mathbf{x}_1$ and $\mathbf{x}_3$ correctly, i.e., $Conf(\mathrm{F}, A \rightarrow \text{Annual Income}) = conf(\mathrm{F}, A = 0 \rightarrow < 100\mathrm{k}) = conf(\mathrm{F}, A = 1 \rightarrow > 200\mathrm{k})$, where $A$ is defined in Table 3, the random variable of decision outcome (0: negative; 1: positive), and

$$conf(\mathrm{F}, 0 \rightarrow < 100\mathrm{k}) = \frac{P(\mathbf{x_1})\tilde{D}_0(\mathbf{x_1})}{P(\mathbf{x_1})\tilde{D}_0(\mathbf{x_1}) + P(\mathbf{x_2})\tilde{D}_0(\mathbf{x_2}) + P(\mathbf{x_3})\tilde{D}_0(\mathbf{x_3})} = \frac{0.3 \times 0.9}{0.3 \times 0.9 + 0.125 \times 0.98 + 0.075 \times 0.1} = 0.675,$$
$$conf(\mathrm{F}, 1 \rightarrow > 200\mathrm{k}) = \frac{P(\mathbf{x_3})\tilde{D}_1(\mathbf{x_3})}{P(\mathbf{x_1})\tilde{D}_1(\mathbf{x_1}) + P(\mathbf{x_2})\tilde{D}_1(\mathbf{x_2}) + P(\mathbf{x_3})\tilde{D}_1(\mathbf{x_3})} = \frac{0.075 \times 0.9}{0.3 \times 0.1 + 0.125 \times 0.02 + 0.075 \times 0.9} = 0.675.$$

Making either inference more private will cause the other one to be less private and hence degrades the overall privacy guarantee as discussed in Section 7.4. In contrast, the optimal solution for the male group tries to minimize the confidence of correctly inferring the annual income of $\mathbf{x}_4$, i.e., $Conf(\mathrm{M}, A \rightarrow \text{Annual Income}) = conf(\mathrm{M}, A = 0 \rightarrow < 100\mathrm{k})$, and

$$conf(\mathrm{M}, 0 \rightarrow < 100\mathrm{k}) = \frac{P(\mathbf{x_4})\tilde{D}_0(\mathbf{x_4})}{P(\mathbf{x_4})\tilde{D}_0(\mathbf{x_4}) + P(\mathbf{x_5})\tilde{D}_0(\mathbf{x_5}) + P(\mathbf{x_6})\tilde{D}_0(\mathbf{x_6})} = \frac{0.225 \times 0.9}{0.225 \times 0.9 + 0.175 \times 0.6 + 0.1 \times 0.1} \approx 0.6378.$$

From the above equation, it is not hard to see that the optimal solution maximizes the denominator while minimizing the numerator in order to minimize the ratio for optimal privacy.

From the above example, we demonstrated that subject to fidelity constraints, how an optimal-privacy ATR can be obtained efficiently using Algorithm 1. The maximum confidence of an adversary, which is $conf(\mathrm{F}, 1 \rightarrow > 200\mathrm{k})$, drops from 100% to 67.5% by setting a 10%-distortion tolerance for the announced ATR.

Figure 5 depicts the privacy-fidelity tradeoffs for both QID groups in this numerical example. Given any fidelity requirement $\delta$, the optimal privacy (i.e., the smallest possible $\beta$) that we can

achieve is the boundary of the trade-off region (the blue curve). The trade-off region for $\beta$, as discussed in Section 7.2, should be within the range $[\beta_{min}, C^*]$, which can be easily computed based on Definition 7 and Lemma 5:

For the QID group of female: $[\beta_{min}, C^*] = [\frac{P(\mathbf{x}_1)}{P(\mathbf{x}_1)+P(\mathbf{x}_2)+P(\mathbf{x}_3)}, \frac{P(\mathbf{x}_1)D_1(\mathbf{x}_1)}{P(\mathbf{x}_1)D_1(\mathbf{x}_1)+P(\mathbf{x}_2)D_1(\mathbf{x}_2)+P(\mathbf{x}_3)D_1(\mathbf{x}_3)}] = [\frac{0.3}{0.5}, \frac{0.075\times 1}{0.075\times 1}] = [0.6, 1]$.

For the QID group of male: $[\beta_{min}, C^*] = [\frac{P(\mathbf{x}_4)}{P(\mathbf{x}_4)+P(\mathbf{x}_5)+P(\mathbf{x}_6)}, \frac{P(\mathbf{x}_4)D_1(\mathbf{x}_4)}{P(\mathbf{x}_4)D_1(\mathbf{x}_4)+P(\mathbf{x}_5)D_1(\mathbf{x}_5)+P(\mathbf{x}_6)D_1(\mathbf{x}_6)}] = [\frac{0.225}{0.5}, \frac{0.225\times 1}{0.225\times 1+0.175\times 0.5}] = [0.45, 0.72]$.

Both results show consistency with Figure 5 : in Figure 5(a), the range of $\beta$ is within [0.6, 1]; in Figure 5(b), the range of $\beta$ is within [0.45, 0.72]. Note that based on Lemma 5, any privacy requirement with $\beta < \max\{0.6, 0.45\} = 0.6$ is not feasible for this dataset, and based on Lemma 4, any privacy requirement with $\beta > 0.72$ can have 1-fidelity solution for the QID group of male, i.e., no perturbation is needed. How much fidelity should be sacrificed in order to achieve a certain level of privacy can thus be known based on the trade-off curves.

*Remark 5.* Here we demonstrate how the values presented in Table 4 are computed. Note that, we only demonstrate the computation of values in the first row (i.e., for $\mathbf{x_1}$), as computations for values in all other rows (for all other $\mathbf{x_i}$'s) follow similar steps. In the following, we start from the left-most value and then move to the right.

For $\mathbf{x} = \mathbf{x_1}$, $\mathbf{x}_{\mathcal{U}} = \{\text{F}\}$ (Female), and since the total population is $12 + 5 + 3 + 9 + 7 + 4 = 40$, $P(\mathbf{x}) = 12/40 = 0.3$. Based on Table 1, since the decision rule represents the probability of receiving a positive decision, $D_1(\mathbf{x})$ is basically the decision rule in Table 1, and $D_0(\mathbf{x})$ is simply $1 - D_1(\mathbf{x})$. The pre-defined fidelity parameter $\delta$ is 0.9, i.e., the announced decision mapping $\tilde{D}_a(\mathbf{x})$ can differ from the true decision mapping $D_a(\mathbf{x})$ by at most 10%, $\forall a = 0, 1$. Therefore, $|\tilde{D}_1(\mathbf{x}) - 0| \leq 0.1$, and we get $\tilde{D}_1(\mathbf{x})_{\min} = 0$ and $\tilde{D}_1(\mathbf{x})_{\max} = 0.1$. Similarly, $|\tilde{D}_0(\mathbf{x}) - 1| \leq 0.1$, and we get $\tilde{D}_0(\mathbf{x})_{\min} = 0.9$ and $\tilde{D}_0(\mathbf{x})_{\max} = 1$. The values of the above terms are based on their definitions (refer to Theorem 1) and the input parameters. Since now we have values for $P(\mathbf{x})$, $\tilde{D}_1(\mathbf{x})_{\min}$, $\tilde{D}_1(\mathbf{x})_{\max}$, $\tilde{D}_0(\mathbf{x})_{\min}$, and $\tilde{D}_0(\mathbf{x})_{\max}$, the values for the terms $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\min}$, $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max}$, $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\min}$, and $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\max}$ are just simple multiplications.

Next, we show how the values for the terms $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max'}$ (third column in the "Computations" category) and $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\max'}$ (the last column) are obtained. Based on Theorem 1, $\tilde{D}_a(\mathbf{x})_{\max'} \triangleq \frac{1}{P(\mathbf{x})}\min\{P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\max}, P(\mathbf{x}^a)\tilde{D}_a(\mathbf{x}^a)_{\min}\}$, where $\mathbf{x}^a \triangleq \arg\max_{\mathbf{x}\in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x})\tilde{D}_a(\mathbf{x})_{\min}$, $\forall a = 0, 1$, we thus have $\mathbf{x}^0 \triangleq \arg\max_{\mathbf{x}\in T_{\{\text{F}\}}} P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\min} = \arg\max_{\mathbf{x}\in\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}}\{0.27, 0.1125, 0\} = \mathbf{x_1}$ and $\mathbf{x}^1 = \arg\max_{\mathbf{x}\in\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3\}}\{0, 0, 0.0675\} = \mathbf{x}_3$. Hence, for $\mathbf{x} = \mathbf{x_1}$, $\tilde{D}_1(\mathbf{x})_{\max'} = \frac{1}{0.3}\min\{0.03, 0.0675\} = 0.03/0.3$ and $\tilde{D}_0(\mathbf{x})_{\max'} = \frac{1}{0.3}\min\{0.3, 0.27\} = 0.27/0.3$. Therefore, we obtain $P(\mathbf{x})\tilde{D}_1(\mathbf{x})_{\max'} = 0.03$ and $P(\mathbf{x})\tilde{D}_0(\mathbf{x})_{\max'} = 0.27$.

## 9 RELATED WORK

There is a huge amount of literature on transparency [28, 46] and fairness [63] for ML. [59] provides a detailed survey on techniques proposed for enhancing transparency and fairness for ML models. However, the perspectives of transparency and fairness in ML models may not be completely in sync with those in algorithmic transparency, e.g., the philosophy of fairness in ML is to train fair ML models or algorithms, while the philosophy of fairness in accountable algorithmic transparency is to verify or to demonstrate whether the examined ML algorithms comply with certain fairness requirements.

There are a number of studies on transparency and fairness and several addressing privacy in data transparency, e.g., [72, 82, 91]; however, *there is little effort in considering the potential impact on privacy brought on by algorithmic transparency schemes and/or fairness measures.* [24] provides transparency in the interaction between Google Ads, users' ad privacy settings, and user behaviors, showing the disparate impact that female gender setting has (vs. male gender setting) on results, e.g., with fewer instances of ads related to high paying jobs; while whether users' privacy could be leaked from Google Ads or the associated transparency report is unclear without further investigation. [8] investigates the limitations of transparency and its impact on society and notes that transparency can threaten privacy, *but it is yet to be made clear what possible aspects of transparency can hurt privacy, and by what privacy-preserving techniques could remedy the situation.* Here, we show that data subjects' privacy can be leaked via various kinds of transparency schemes and fairness measures in an announced ATR and propose a privacy protection scheme yielding privacy preserving information on an ATR. Motivated by transparency and fairness, [33] raises questions regarding fair privacy for all participating users, as it is considered discriminatory when different users are protected by different levels of privacy; however, *on what notion of privacy should be fair, by what methodology to protect such privacy and to achieve it fairly are still unclear.* However, in contrast, numerical examples in Section 8 show that *the optimal privacy for different QID groups, subject to the same fidelity constraints, is in general different* due to the disparity of prior distributions, prior vulnerabilities [66], side-information, and associated decision mapping between groups. [81] studies the problem of providing transparency to consumers while preserving information privacy for them, and proposes *informational norms* to constrain the collection, use, and distribution of transparent information in role-appropriate manners to fulfill the goal. *However, the definition of role-appropriate manners is yet to be more specific, and it is also unclear how fairness measures, which compare decision rules between two individuals or groups (likely in different roles), should be announced under such a norm.* In contrast, our privacy protection scheme does not make any assumptions on informational norm and does not rely on any norm (potentially hard to accomplish) to protect users' privacy, and thus can be applied generally. In addition, informational norm may still not be adequate to protect users' privacy: individuals belonging to the same role may still be able to infer private information of others, e.g., in Table 1, any female credit card owner can infer other female credit card owners' income range.

There exist a couple of works using differential privacy (DP) to remedy the privacy leakage/attack issue in algorithmic transparency or model explanations. A recent work [79] demonstrates membership inference attacks [80] on training datasets of ML models by utilizing information from the corresponding featured-based model explanations (i.e., feature importance/interaction transparency schemes). To address this issue, in [70], DP is applied to the gradient descent algorithm for generating feature-based model explanations. [23], arguably the only previous work that addresses transparency, fairness, and privacy in an accountable ATR, proposes a feature-based measure, named ***quantitative input influence*** **(QII)**; based on which, the authors propose public and personalized transparency reports, as well as a fairness measure, named *group disparity*, to measure potential disparate impacts on different groups of people. DP is adopted to the above measures in order to prevent potential privacy leaks caused by the provided QII and group disparity in an announced ATR. *However, applying DP solely does not result in prevention of inference attacks, in particular, attribute inference attacks: once strong correlations between attribute values are known, sensitive attribute values can be inferred no matter whether a privacy victim belongs to a specific dataset or not, and thus DP cannot help in such a scenario* (see [32], Section 2.3.2, the *smoking-causes-cancer* example). In light of this, here, we propose a privacy-preserving scheme to prevent attribute inference attacks by limiting the attribute

inference confidence from public/known attribute values, via an announced ATR with assistance of side-information, to any data subjects' private attribute values.

## 10 SUMMARY

In this work, we demonstrated how an honest-but-curious adversary can utilize widely-available information together with information provided in an algorithmic transparency report to obtain data subjects' private information. From this we glean which potential aspects of transparency and fairness measures can hurt privacy. We then propose a privacy scheme that perturbs the information to be announced, to remedy the potential privacy leaks. We systematically study the impact of such perturbation on fairness measures and the fidelity of the announced information, formulated as an optimization problem for optimal privacy subject to fidelity constraints. To efficiently solve the optimization problem, we identify important properties and provide closed-form solutions, based on which, we propose a privacy protection scheme. Given fidelity requirements, the proposed scheme can efficiently produce optimal-privacy ATRs in linear time. In addition, we provide insight into our proposed optimal privacy scheme. We believe that our proposed methodology is suited for more general problems beyond algorithmic transparency, where the release of the model information is controlled and the input data cannot be modified - for instance, one example is the setting of model inversion attacks [40] where the model owner has no authority to modify the input data (patients' clinical history and genomic data) but has the control of the amount of information about the (dose-suggesting) model to be released. In such a scenario, our scheme can help privately release information of a model to pharmacists for better understanding of suggesting personalized dosage.

## APPENDICES

## A FAIRNESS MEASURES

Another important motivation of providing algorithmic transparency is to understand if a decision-making algorithm is fair. GDPR Article 5 regulation indicates that personal data should be processed fairly and in a transparent manner. Many researchers are committed to providing proper measures for fairness and making ML algorithms fair [15, 22, 30, 35, 54, 55, 92]. In general, there are two main categories of fairness: (i) individual fairness, and (ii) group fairness. Popular definitions of group fairness includes **statistical parity (SP)**, **conditional statistical parity (CSP)**, and **p%-rule (PR)**.

### A.1 Measures for Individual Fairness

*Definition 8 ($(\mathfrak{D},\mathcal{D})$-Individual Fairness [30]).* Given a distance measure $\mathcal{D}:\mathcal{R}_X\times\mathcal{R}_X\to\mathbb{R}^+\triangleq[0,\infty)$ on individuals' records, a decision mapping $D:\mathcal{R}_X\to\Delta(\mathcal{A})$ satisfies *individual fairness* if it complies with the $(\mathfrak{D},\mathcal{D})$-Lipschitz property for every two individuals' records $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{R}_X$, i.e.,

$$\mathfrak{D}(D(\mathbf{x}_1),D(\mathbf{x}_2))\le\mathcal{D}(\mathbf{x}_1,\mathbf{x}_2), \tag{23}$$

where $\mathfrak{D}:\Delta(\mathcal{A})\times\Delta(\mathcal{A})\to\mathbb{R}^+$ is a distance measure on distributions over $\mathcal{A}$. Moreover, we define $D$ satisfying individual fairness up to bias $\varepsilon$ if for all $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{R}_X$, we have

$$\mathfrak{D}(D(\mathbf{x}_1),D(\mathbf{x}_2))\le\mathcal{D}(\mathbf{x}_1,\mathbf{x}_2)+\varepsilon. \tag{24}$$

Individual fairness ensures a decision mapping maps similar people similarly. When two individuals' records $\mathbf{x}_1$ and $\mathbf{x}_2$ are similar, i.e., $\mathcal{D}(\mathbf{x}_1,\mathbf{x}_2)\cong 0$, the Lipschitz condition in equation (23) ensures that both records map to similar distributions over $\mathcal{A}$. Candidates for distance measure $\mathfrak{D}$ include (but are not limited to) *statistical distance* and *relative $l_\infty$ metric*. The relative $l_\infty$ metric

(a.k.a. *relative infinity norm*) of two distributions $Z_1$ and $Z_2$, defined as follows

$$\mathfrak{D}_{\infty}(Z_1, Z_2) = \sup_{a \in \mathcal{A}} \log \left( \max \left\{ \frac{Z_1(a)}{Z_2(a)}, \frac{Z_2(a)}{Z_1(a)} \right\} \right), \tag{25}$$

is considered a potential better choice in the aspect that it does not require the distance measure $\mathcal{D}$ to be re-scaled within $[0, 1]$.[8] However, it has the shortcoming that it is sensitive to small probability values. The statistical distance, or the *total variation norm*, of two distributions $Z_1$ and $Z_2$, defined as follows

$$\mathfrak{D}_{tv}(Z_1, Z_2) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} |Z_1(a) - Z_2(a)|, \tag{26}$$

is a more stable measure in this aspect.

### A.2 Measures for Group Fairness

Popular measures for group fairness include (but are not limited to) **statistical parity (SP)** (a.k.a. demographic parity) [30, 35, 55, 92], **conditional statistical parity (CSP)** [22, 54], $p$-% rule (PR) [15, 35], accuracy parity (a.k.a. equalized odds) [47], and true positive parity (a.k.a. equal opportunity) [47]. However, the last two measures require knowledge of labeled outputs and are thus particularly used to train fair ML algorithms in supervised learning. For algorithmic transparency, we use the former three measures for group fairness.

Define $g(X)$ a projection function from input attributes $X$ onto a group in protected attributes, $v(X)$ a score/valuation function from $X$ onto a set scores, and $T_{\mathcal{Y}} \triangleq \{\mathbf{x} \in \mathcal{R}_X \mid g(\mathbf{x}) \in \mathcal{Y}\}$ the set/tuple in which records belong to a protected group $\mathcal{Y}$. We summarize definitions of measures for group fairness in the following:

*Definition 9 (*Statistical Parity (SP)*).* A decision mapping $D : \mathcal{R}_X \to \Delta(\mathcal{A})$ satisfies statistical parity for two groups $\mathcal{Y}_1$ and $\mathcal{Y}_2$ up to bias $\varepsilon$ if for every decision outcome $a \in \mathcal{A}$, we have the following property

$$\mathfrak{D}_{tv}(\mathbb{E}[D_a(X)|T_{\mathcal{Y}_1}], \mathbb{E}[D_a(X)|T_{\mathcal{Y}_2}]) \leq \varepsilon. \tag{27}$$

*Definition 10 (*Conditional Statistical Parity (CSP)*).* Given a score/valuation function $v(X)$ based on input attributes $X$, define $T_{\mathcal{Y},\mathcal{V}} \triangleq \{\mathbf{x} \in \mathcal{R}_X \mid g(\mathbf{x}) \in \mathcal{Y}, v(\mathbf{x}) \in \mathcal{V}\}$ the set/tuple in which records belong to a protected group $\mathcal{Y}$ having scores in a set $\mathcal{V}$. A decision mapping $D : \mathcal{R}_X \to \Delta(\mathcal{A})$ satisfies conditional statistical parity given the same score conditions $\mathcal{V}$ for two groups $\mathcal{Y}_1$ and $\mathcal{Y}_2$ up to bias $\varepsilon$ if for every decision outcome $a \in \mathcal{A}$, we have the following property

$$\mathfrak{D}_{tv}(\mathbb{E}[D_a(X)|T_{\mathcal{Y}_1,\mathcal{V}}], \mathbb{E}[D_a(X)|T_{\mathcal{Y}_2,\mathcal{V}}]) \leq \varepsilon. \tag{28}$$

*Definition 11 (*$p$-% Rule (PR)*).* A decision mapping $D : \mathcal{R}_X \to \Delta(\mathcal{A})$ satisfies $p$-% rule for two groups $\mathcal{Y}_1$ and $\mathcal{Y}_2$ if for every decision outcome $a \in \mathcal{A}$, we have the following property

$$\left| \log \left( \frac{\mathbb{E}\left[D_a(X)|T_{\mathcal{Y}_1}\right]}{\mathbb{E}\left[D_a(X)|T_{\mathcal{Y}_2}\right]} \right) \right| \leq -\log p. \tag{29}$$

[8] The normalization could bring non-trivial burden, especially when the maximal distance can be arbitrarily large.

In particular, for binary decisions, we say a decision rule $d$ satisfies SP, CSP, or PR for two groups $\mathcal{Y}_1$ and $\mathcal{Y}_2$ up to bias $\varepsilon$ (SP and CSP only) if

$$\text{SP: } |\mathbb{E}[d(X)|T_{\mathcal{Y}_1}] - \mathbb{E}[d(X)|T_{\mathcal{Y}_2}]| \leq \varepsilon \tag{30}$$

$$\text{CSP: } |\mathbb{E}[d(X)|T_{\mathcal{Y}_1,\mathcal{V}}] - \mathbb{E}[d(X)|T_{\mathcal{Y}_2,\mathcal{V}}]| \leq \varepsilon \tag{31}$$

$$\text{PR: } p \leq \frac{\mathbb{E}[d(X)|T_{\mathcal{Y}_1}]}{\mathbb{E}[d(X)|T_{\mathcal{Y}_2}]} \leq \frac{1}{p}. \tag{32}$$

Note that all fairness definitions are based on the distance between the decision of two groups,[9] specifically, *total variation* (26) and relative metric (25). Let $\mathcal{F}$ denote the set of all fairness definitions. Based on the use of distance metrics, $\mathcal{F}$ can be classified as follows:

- *Total-variation-based fairness definitions ($\mathcal{F}_{tv}$)*: Definitions include $(\mathfrak{D}_{tv}, \mathcal{D})$-individual fairness, statistical parity, and conditional statistical parity.
- *Relative-metric-based fairness definitions ($\mathcal{F}_{rm}$)*: Definitions include $(\mathfrak{D}_{\infty}, \mathcal{D})$- individual fairness and p%-rule.

## B PRIVACY LEAKAGE VIA FEATURE IMPORTANCE/INTERACTION

Feature (value) importance, or feature (value) interaction, measures the *importance* (or *influence*) of input attributes (or attribute values) to the decision outcomes. The importance of an input attribute (value) is measured based on *the corresponding change of output due to change of that certain input.* By changing an input, if the change of output is significant, it implies the input is important (has significant influence) to the output. On the other hand, if the output changes very little, the input contributes very little to the output.

Different works may propose different measures, but their philosophies are almost the same (as stated above). For example, the measures for change of an input can be (i) removing the presence of an input attribute, or (ii) permuting attribute values on an input attribute. The measures of outputs are many, e.g., (i) accuracy of the (predicted) outputs [18, 38], (ii) probability of receiving a certain outcome [23], (iii) statistics measures, such as partial dependence [41, 45], H-statistic [42], or variable interaction networks [51], or (iv) a self-defined quantity or a score/gain function. The measures for the change of outputs can be (i) difference (i.e., subtraction), (ii) ratio, or (iii) averaged difference/contribution, e.g., the Shapley value [56], of the measured outputs. In this regard, it is impractical for us to demonstrate the privacy leakage issue for all present methods. However, since the philosophies of all these methods are similar, it is reasonable for us to demonstrate the privacy hacking procedures via a representative one. The principles of hacking procedures can be transferred and applied to other methods.

We investigate potential privacy leakage via the ***quantitative input influence* (QII)** proposed in the most pioneering work [23] in accountable ATR. For QII, the measure for change of an input is permuting attribute values (called *intervention* in the paper) on an input attribute. The measure of output can be user-specified, called *quantity of interest*, denoted by $Q$. The measure for change of output is difference between (subtraction of) two measured outputs. Formally, the QII of an input attribute $k$ for a quantity of interest $Q$ is defined as

$$I^Q(k) = Q(X) - Q(X_{-k}U_k), \tag{33}$$

in which $X_{-k}U_k$, meaning that attribute $k$ is (removed from input $X$ and) replaced by a permuted version $U_k$, represents intervention on attribute $k$. In particular, for $Q(X) = P\{c(X) = 1|X \in T_{\mathcal{W}}\}$,

[9] More precisely, from (27), (28), and (29), the decision distribution of a group is the expected decision mapping among the group, over all decision outcomes $a \in \mathcal{A}$.

the fraction of records belonging to a set $T_{\mathcal{W}}$ (e.g., women) with positive classification, the QII of an input attribute $k$ is

$$I(k) = P\{c(X) = 1 | X \in T_{\mathcal{W}}\} - P\{c(X_{-k}U_k) = 1 | X \in T_{\mathcal{W}}\}, \tag{34}$$

where $c(\cdot)$ is a classifier (decision-maker). The QII of a set of input attributes $\mathcal{K}$ is defined similarly, using $\mathcal{K}$ instead of $k$.

In the following, we conduct an experiment to demonstrate the hacking of decision rules via provided QII's on an ATR for a real dataset, and utilize the hacked decision rules to further infer private records as what we did in Section 3.2. We use the Australian credit approval dataset from UCI machine learning repository [25] in our experiment . The dataset has 690 instances, with 15 input attributes and 1 output attribute. All attribute information can be found in Table 5. In order to protect confidentiality of the data, all attribute names and values have been changed to meaningless symbols by the dataset provider. Based on the dataset, with adequate data cleaning and pre-processing, we train a classifier based on a fully-connected neural network with one input layer (36 inputs, after one-hot encoding for categorical attribute values), two hidden layers (147 and 85 neurons, respectively), and one output layer (binary outputs), with dropout rate 0.5. The averaged testing accuracy of the trained classifier is 89.5%.

The trained classifier is served as the knowledge of a trust-worthy 3rd-party regulation agency which feeds both inputs and outputs of the dataset to an ML model in order to learn the unknown decision-making rules of this Australian credit card company. Since QII is a data-mining based approach [23], the regulation agency provides information regarding input influences (QII) in an ATR upon users' demand. Since the access control is still an open question, we assume a user is able to request such information in a reasonable manner.

Based on the above experimental settings, we first construct a scenario to demonstrate the hacking.

**Scenario**:

- Let $\mathcal{U}$ = {A4, A5, A6, A7} be public attributes and all other attributes are private and unknown to adversaries (See Remark 6).
- Alice has public record $\mathbf{x}_{\mathcal{U}}$ = {y, p, k, v}. She gets a positive decision (+) and receives a credit card.
- Tom also has the same public record $\mathbf{x}_{\mathcal{U}}$ = {y, p, k, v}. He gets a negative decision (-).
- An adversary is a friend of both, knowing their public records, knowing that Alice owns such a credit card but Tom doesn't. The adversary also has the knowledge of joint distribution of A4~A7, A9, and A11, e.g., demographic statistics of age, marriage status, race, and annual income.

A snapshot of the QID group $T_{\mathbf{x}_{\mathcal{U}}=\{\text{y, p, k, v}\}}$ is shown in Table 6, in which public attributes are marked in grey, class attribute (decision outcome) is marked in light blue, and for those attributes that an adversary has associated side-information (joint distribution) are marked in bold.

We next demonstrate privacy hacking procedures in the following. Let $\mathcal{W}_0$ = {A4 = y, A5 = p, A6 = k, A7 = v, A11 ∈ [0,1]}, and $\mathcal{W}_1$ = {A4 = y, A5 = p, A6 = k, A7 = mv, A11 ∈ [10,11]}.

**Privacy Hacking**:

(1) Since the input to QII can be a set of attributes, i.e., the joint influence of a set of input attributes. Let $\mathcal{S}$ be the collection of all private attributes as denoted in Table 3, which is {A1~A3, A8~A15} in our scenario. The adversary then sends the following QII query to the regulation agency:
   - Input Attribute: $\mathcal{S}$
   - Quantity of Interest: $Q(X) = P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\}$.

Table 5. Attribute Information of the Australian Credit Approval Dataset

| | | | |
|---|---|---|---|
| A1: | b, a. | A9: | t, f. |
| A2: | continuous. | A10: | t, f. |
| A3: | continuous. | A11: | continuous. |
| A4: | u, y, l, t. | A12: | t, f. |
| A5: | g, p, gg. | A13: | g, p, s. |
| A6: | c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff. | A14: | continuous. |
| A7: | v, h, bb, j, n, z, dd, ff, o. | A15: | continuous. |
| A8: | continuous. | A16: | +,- (class attribute) |

Table 6. A Snapshot of the QID Group $T_{\mathbf{x}_{\mathcal{U}}=\{y, p, k, v\}}$ in the Australian Credit Approval Dataset After Data Cleaning, where Public Attributes are Marked in Grey, Class Attribute (Decision Outcome) is Marked in Light Blue, and Attributes where an Adversary has Associated Side-Information (Joint Distribution) are Marked in Bold

| A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | A10 | A11 | A12 | A13 | A14 | A15 | A16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| b | 38.25 | 10.125 | **y** | **p** | **k** | **v** | 0.125 | **f** | f | **0** | f | g | 160 | 0 | - |
| b | 29.83 | 1.25 | **y** | **p** | **k** | **v** | 0.25 | **f** | f | **0** | f | g | 224 | 0 | - |
| b | 16.92 | 0.335 | **y** | **p** | **k** | **v** | 0.29 | **f** | f | **0** | f | s | 200 | 0 | - |
| a | 29.85 | 1.75 | **y** | **p** | **k** | **v** | 1.25 | **f** | f | **0** | t | g | 280 | 0 | - |
| b | 20 | 1.25 | **y** | **p** | **k** | **v** | 0.125 | **f** | f | **0** | f | g | 140 | 4 | - |
| b | 22.5 | 0.125 | **y** | **p** | **k** | **v** | 0.125 | **f** | f | **0** | f | g | 200 | 70 | - |
| b | 28.17 | 0.125 | **y** | **p** | **k** | **v** | 0.085 | **f** | f | **0** | f | g | 216 | 2100 | - |
| b | 23.5 | 3.165 | **y** | **p** | **k** | **v** | 0.415 | **f** | t | **1** | t | g | 280 | 80 | - |
| b | 21.67 | 1.165 | **y** | **p** | **k** | **v** | 2.5 | **t** | t | **1** | f | g | 180 | 20 | - |
| b | 36.67 | 4.415 | **y** | **p** | **k** | **v** | 0.25 | **t** | t | **10** | t | g | 320 | 0 | + |
| b | 48.58 | 0.205 | **y** | **p** | **k** | **v** | 0.25 | **t** | t | **11** | f | g | 380 | 2732 | + |

(2) The adversary gets a response $I(\mathcal{S}) = 0.66475333$, which indicates the degree of influence of all private input attributes $\mathcal{S}$ to the group $\mathcal{W}_1$.
(3) The adversary sends the following QII query to the regulation agency:
- Input Attribute: $\mathcal{S}$
- Quantity of Interest: $Q(X) = P\{c(X) = 1 | X \in T_{\mathcal{W}_0}\}$.

(4) The adversary gets a response $I(\mathcal{S}) = -0.33524666$, which indicates the degree of influence of all private input attributes $\mathcal{S}$ to the group $\mathcal{W}_0$. Note that negative sign stands for negative impact as mentioned in Section 3.1.
(5) From the above two query responses, the adversary has

$$0.66475333 = P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} - P\{c(X_{-\mathcal{S}}U_{\mathcal{S}}) = 1 | X \in T_{\mathcal{W}_1}\}$$
$$-0.33524666 = P\{c(X) = 1 | X \in T_{\mathcal{W}_0}\} - P\{c(X_{-\mathcal{S}}U_{\mathcal{S}}) = 1 | X \in T_{\mathcal{W}_0}\}.$$

(6) Since $\mathcal{W}_1$ and $\mathcal{W}_0$ have the same public record $\mathbf{x}_{\mathcal{U}} = \{y, p, k, v\}$, for the same classifier, we must have

$$P\{c(X_{-\mathcal{S}}U_{\mathcal{S}}) = 1 | X \in T_{\mathcal{W}_1}\} = P\{c(X_{-\mathcal{S}}U_{\mathcal{S}}) = 1 | X \in T_{\mathcal{W}_0}\}.$$

(7) Utilize the above equality, the adversary obtains

$$P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} - P\{c(X) = 1 | X \in T_{\mathcal{W}_0}\} = 1.$$

Since probabilities are always within $[0, 1]$, the adversary thus obtains decision rules

$$P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} = 1,$$
$$P\{c(X) = 1 | X \in T_{\mathcal{W}_0}\} = 0.$$

It is worth mentioning that the attack may not be unique. As shown in the following, there could exist many ways to obtain decision rules, and thus it seems hopeless to cease the attack simply by access control.

**Privacy Hacking (Method 2)**:

(1) The adversary sends the following QII query to the regulation agency:
- Input Attribute: A9
- Quantity of Interest: $Q(X) = P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\}$.

(2) The adversary gets a response $I(\text{A9}) = 0.45142778$, which indicates the degree of influence of input attribute A9 to the group $\mathcal{W}_1$.

(3) The adversary analyzes the response $I(\text{A9})$. Define $P_t = P\{c(X_{-\text{A9}}U_{\text{A9}}) = 1 | X \in T_{\mathcal{W}_1}, U_{\text{A9}} = \text{t}\}$ and $P_f = P\{c(X_{-\text{A9}}U_{\text{A9}}) = 1 | X \in T_{\mathcal{W}_1}, U_{\text{A9}} = \text{f}\}$. He gets

$$\begin{aligned} 0.45142778 &= P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} - P\{c(X_{-\text{A9}}U_{\text{A9}}) = 1 | X \in T_{\mathcal{W}_1}\} \\ &= P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} - P\{U_{\text{A9}} = \text{t}\}P_t - P\{U_{\text{A9}} = \text{f}\}P_f. \end{aligned}$$

(4) The adversary realizes the fact that, for the same classifier, we must have

$$\begin{aligned} P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} &= P\{c(X) = 1 | X \in T_{\mathcal{W}_1}, A9 = \text{t}\} \\ &= P\{c(X_{-A9}U_{A9}) = 1 | X \in T_{\mathcal{W}_1}, U_{A9} = \text{t}\} = P_t. \end{aligned}$$

(5) Since the adversary has joint distribution knowledge as mentioned in the scenario, he knows the marginal distribution $P\{U_{A9} = \text{f}\} = 1 - P\{U_{A9} = \text{t}\} = 0.45142857$, he then gets

$$P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} - P_f = \frac{0.45142778}{0.45142857} \approx 1.$$

(6) Since probabilities are always within $[0, 1]$, the adversary knows $P_f \approx 0$, and

$$P\{c(X) = 1 | X \in T_{\mathcal{W}_1}\} \approx 1.$$

The adversary obtains very accurate information regarding decision rule for $\mathcal{W}_1$.

Based on the hacked decision rules above, the adversary has 100% confidence that Alice's record belongs to $T_{\mathcal{W}_1}$ and Tom's record belongs to $T_{\mathcal{W}_0}$. Based on Table 6, he then knows that Alice's A11 attribute value is either 10 or 11, and Tom's is either 0 or 1. If the adversary has richer side-information, e.g., joint distribution including A8 and A14, then the adversary has 100% confidence that Alice's A8 attribute value is 0.25, her A14 attribute value is in the range between 300 and 400, and Tom's A14 attribute value is in the range between 100 and 300.

It is worth mentioning that, based on our investigation, we do not find a general attack method that can be applied to all datasets and decision rules. However, this does not mean the attacks demonstrated above are cherry-picked. As we have shown, there could exist many feasible attack approaches. Adversaries can simply try multiple different attempts and/or collude their test results so that eventually they acquire a successful attack result. Moreover, similar to the privacy incidents of AOL search data leak [11] and de-anonymization of the Netflix Price dataset [67], although there is no guarantee that the attacks can always succeed in all the cases, *as long as the attack **can** succeed, there exists a privacy breach which can result in a catastrophic disaster.*

In fact, the authors of the pioneering work, i.e., [23], had already noticed the potential privacy issue in algorithmic transparency and added noise to make the measures differentially private. Unfortunately, adding differentially private noise [31] solely cannot mitigate the demonstrated

privacy leakage issue. The fundamental reason is that differential privacy only guarantees a small amount of information leakage when an individual participates the survey or opts into a database. Differential privacy itself does not guarantee information leakage due to strong statistical inference between attributes; this has been noted in many previous works such as [7, 13, 29], and Section 2.3.2 in [32]. The most classic example is the study of "smoking causes cancer", in which no matter whether a person opts into the survey or not, once we know that he is a smoker, we know he has a certain high chance of getting lung cancer. What can be guaranteed in the proposed differentially private perturbation for an ATR is that an adversary can only gain very little information by comparing two ATRs of which the training data to train the classifiers differ in only one data subject's record. When the size of the dataset is very large, the required variance of DP noise is very small. This is why they claimed only very little noise needs to be added.

*Remark 6.* Although all attribute names in the dataset are removed, we are still able to reasonably conjecture public and private attributes based on their influences to the decision outcome. Attributes with high influences are more likely to be private attributes such as income or credit score, and attributes with low influences are likely to be public ones. Observe that attribute A9, A11, and A15 are the most influential ones and others are less significant (from experiments). For ease of demonstration, we choose four adjacent categorical attributes from insignificant ones, A4 to A7, to serve as public attributes.

## C MINIMUM UNCERTAINTY

*Definition 12 (Minimum Uncertainty).* Given an inference channel $\langle X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}} \rangle$, the uncertainty of inferring a certain sensitive attribute value $x_{\mathcal{S}}$ from a certain inference source $\{\mathbf{x}_{\mathcal{U}}, a\}$ is defined as $ucrt(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}}) = -\log(conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}}))$. The minimum uncertainty of inferring any sensitive value from any inference channel is

$$\begin{aligned} Ucrt(X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}}) &= \min_{\mathbf{x}_{\mathcal{U}}, a, x_{\mathcal{S}}} \{-\log(conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}}))\} \\ &= -\log\left(\max_{\mathbf{x}_{\mathcal{U}}, a, x_{\mathcal{S}}} \{conf(\mathbf{x}_{\mathcal{U}}, a \rightarrow x_{\mathcal{S}})\}\right) \\ &= -\log(Conf(X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}})). \end{aligned}$$

Similarly, the corresponding privacy requirement for minimal uncertainty is the following.

*Definition 13 ($\gamma$-Minimum Uncertainty).* In an algorithmic transparency report, $\tilde{D}$ satisfies *$\gamma$-Minimum Uncertainty* if $Ucrt(X_{\mathcal{U}}, A \rightarrow X_{\mathcal{S}}) \geq \gamma$.

The above privacy requirement is basically saying that an adversary's uncertainty on inferring any sensitive value from any inference channel cannot be too low and should be lower-bounded by a threshold $\gamma$; the larger the $\gamma$, the higher the minimum uncertainty, and thus the stronger the privacy. From definition (12), it is clear that $\gamma$-Minimum Uncertainty implies $e^{-\gamma}$-Maximum Confidence, and $\beta$-Maximum Confidence implies $-\log\beta$-Minimum Uncertainty.

Lemma 7. *The privacy requirement $\gamma$-Minimum Uncertainty imposes the following constraints to the announced decision mapping $\tilde{D}$, $\forall \mathbf{x} \in \mathcal{R}_X$, $\forall a \in \mathcal{A}$,*

$$\log\left(\sum_{\mathbf{x}'} \tilde{D}_a(\mathbf{x}') P_X(\mathbf{x}')\right) - \log(\tilde{D}_a(\mathbf{x}) P_X(\mathbf{x})) \geq \gamma. \tag{35}$$

## D PROOF OF LEMMA 1

Proof. Recall that $conf(\mathbf{x}_{\mathcal{U}}, a \to x_{\mathcal{S}})$, the confidence of inferring a sensitive attribute value $x_{\mathcal{S}}$, is a posterior epistemic probability which can be expressed as

$$conf(\mathbf{x}_{\mathcal{U}}, a \to x_{\mathcal{S}}) = \tilde{P}_{X_{\mathcal{S}}|X_{\mathcal{U}},A}(x_{\mathcal{S}}|\mathbf{x}_{\mathcal{U}}, a) = \frac{\tilde{P}_{A|X_{\mathcal{U}},X_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{U}}, x_{\mathcal{S}})P_{X_{\mathcal{U}},X_{\mathcal{S}}}(\mathbf{x}_{\mathcal{U}}, x_{\mathcal{S}})}{\sum\limits_{x'_{\mathcal{S}} \in \mathcal{R}_{X_{\mathcal{S}}}} \tilde{P}_{A|X_{\mathcal{U}},X_{\mathcal{S}}}(a|\mathbf{x}_{\mathcal{U}}, x'_{\mathcal{S}})P_{X_{\mathcal{U}},X_{\mathcal{S}}}(\mathbf{x}_{\mathcal{U}}, x'_{\mathcal{S}})}. \quad (36)$$

Let $\mathbf{x} = (\mathbf{x}_{\mathcal{U}}, x_{\mathcal{S}})$ and define $T_{\mathbf{x}_{\mathcal{U}}} \triangleq \{\mathbf{x}' \in \mathcal{R}_X \mid \mathbf{x}'_{\mathcal{U}} = \mathbf{x}_{\mathcal{U}}\}$ to denote the tuple in which records having the same QID $\mathbf{x}_{\mathcal{U}}$. We have a more comprehensive expression

$$conf(\mathbf{x}_{\mathcal{U}}, a \to x_{\mathcal{S}}) = \frac{\tilde{P}_{A|X}(a|\mathbf{x})P_X(\mathbf{x})}{\sum\limits_{\mathbf{x}' \in T_{\mathbf{x}_{\mathcal{U}}}} \tilde{P}_{A|X}(a|\mathbf{x}')P_X(\mathbf{x}')} = \frac{\tilde{D}_a(\mathbf{x})P_X(\mathbf{x})}{\sum\limits_{\mathbf{x}' \in T_{\mathbf{x}_{\mathcal{U}}}} \tilde{D}_a(\mathbf{x}')P_X(\mathbf{x}')}. \quad (37)$$

Therefore, based on Definitions 3 and 4, the privacy requirement $\beta$-Maximum Confidence imposes the following constraints for all $\mathbf{x} = (\mathbf{x}_{\mathcal{U}}, x_{\mathcal{S}}) \in \mathcal{R}_X, \forall a \in \mathcal{A}$.

$$\frac{\tilde{D}_a(\mathbf{x})P_X(\mathbf{x})}{\sum\limits_{\mathbf{x}' \in T_{\mathbf{x}_{\mathcal{U}}}} \tilde{D}_a(\mathbf{x}')P_X(\mathbf{x}')} \le \beta. \quad (38)$$

## E PROOF OF LEMMA 4

Proof. We first prove that if $\beta \ge C^*$, $\tilde{D} = D$ is a feasible solution. We then prove its converse: if $\tilde{D} = D$ is a feasible solution, we must have $\beta \ge C^*$.

We first prove that if $\beta \ge C^*$, the 1-fidelity solution $\tilde{D} = D$ is a feasible solution, i.e., it satisfies all constraints. Obviously, the solution $\tilde{D} = D$ satisfies probability distribution conditions and fidelity constraints. Based on Definition 7, $\tilde{D} = D$ yields

$$\frac{P(\mathbf{x})\tilde{D}_a(\mathbf{x})}{\sum\limits_{\mathbf{x}' \in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x}')\tilde{D}_a(\mathbf{x}')} = C^* \le \beta\,, \ \forall \mathbf{x} \in T_{\mathbf{x}_{\mathcal{U}}}, \forall a \in \mathcal{A}.$$

Therefore, it also satisfies privacy constraints, and hence when $\beta \ge C^*$, the 1-fidelity solution is a feasible solution.

Next, we prove the converse by proving its contrapositive, i.e., if $\beta < C^*$, $\tilde{D} = D$ is not a feasible solution. Apparently when $\tilde{D} = D$, the highest confidence that an adversary can have exceeds $\beta$, and hence it violates privacy requirements and cannot be a feasible solution. We therefore prove the converse. □

## F PROOF OF LEMMA 5

Proof. We first prove that if an (OPT-Sub) has feasible solutions, $\beta \ge \beta_{\min}$. We then prove its converse: if $\beta \ge \beta_{\min}$, an (OPT-Sub) must have feasible solutions.

We first prove the conditional statement by proving its contrapositive, i.e., if $\beta < \beta_{\min}$, there exists no feasible solution for an (OPT-Sub). Since $\tilde{D}$ is non-negative, we can rewrite the privacy constraints as follows

$$P(\mathbf{x})\tilde{D}_a(\mathbf{x}) - \beta \sum_{\mathbf{x}' \in T_{\mathbf{x}_{\mathcal{U}}}} P(\mathbf{x}')\tilde{D}_a(\mathbf{x}') \le 0\,, \quad (39)$$

which has to be satisfied $\forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$ and $\forall a \in \mathcal{A}$. Sum (39) over all $a \in \mathcal{A}$, by (EQ-Sub), we have

$$P(\mathbf{x}) - \beta \sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}') \leq 0\,, \forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}, \tag{40}$$

which is equivalent to $\beta \geq \max_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}|T_{\mathbf{x}_\mathcal{U}})$. Therefore, if there exists *any* $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$ such that $\beta < P(\mathbf{x}|T_{\mathbf{x}_\mathcal{U}})$, then (39) cannot be satisfied for *all* $\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$, and hence no feasible solution exists.

We then prove the converse. If $\beta \geq \max_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}|T_{\mathbf{x}_\mathcal{U}})$, there always exists a feasible solution $\tilde{D}_a(\mathbf{x}') = 1/|\mathcal{A}|$, $\forall \mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}$, $\forall a \in \mathcal{A}$. To see this, we only need to verify if it satisfies all constraints. It is very obvious that the solution satisfies probability distribution conditions. Since fidelity constraints are trivialized, we then only need to verify if the solution satisfies privacy constraints. Since $\tilde{D}_a(\mathbf{x}')$ is a constant for all $a$ and $\mathbf{x}$, the left hand side of (39) becomes $P(\mathbf{x}|T_{\mathbf{x}_\mathcal{U}})$, and thus the privacy constraints are also satisfied. Hence $\tilde{D}_a(\mathbf{x}') = 1/|\mathcal{A}|$ is a feasible solution and we proved the converse. □

## G PROOF OF LEMMA 6

Proof. We prove this by contradiction. Assume that $C^* < \beta_{\min}$. By their definitions in Lemma 4 and 5, it follows that

$$\max_{\substack{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}, \\ a \in \mathcal{A}}} \frac{P(\mathbf{x}) D_a(\mathbf{x})}{\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}') D_a(\mathbf{x}')} < \max_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} \frac{P(\mathbf{x})}{\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}')}. \tag{41}$$

Let $\mathbf{x}^\dagger = \arg\max_{\mathbf{x} \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}|T_{\mathbf{x}_\mathcal{U}})$. The right hand side of (41) is equivalent to $P(\mathbf{x}^\dagger)/\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}')$. If inequality (41) holds, the following inequalities must hold

$$\frac{P(\mathbf{x}^\dagger) D_a(\mathbf{x}^\dagger)}{\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}') D_a(\mathbf{x}')} < \frac{P(\mathbf{x}^\dagger)}{\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}')}\,, \ \forall a \in \mathcal{A}, \tag{42}$$

since the maximum of the left hand side of (42) over all $a \in \mathcal{A}$ is not greater than the left hand side of (41). Therefore, if there exists any $a \in \mathcal{A}$ for which the corresponding inequality in (42) does not hold, it implies our assumption $C^* < \beta_{\min}$ is not true, and, if so, we are done with the proof.

If there exists no such an $a$ and (42) holds, by eliminating $P(\mathbf{x}^\dagger)$ from both sides of (42) and cross-multiplying (as all terms are non-negative), (42) is equivalent to the following

$$D_a(\mathbf{x}^\dagger) \sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}') < \sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}') D_a(\mathbf{x}')\,, \ \forall a \in \mathcal{A}. \tag{43}$$

Sum (43) over $a \in \mathcal{A}$ for both sides, based on (EQ-Sub), we obtain $\sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}') < \sum_{\mathbf{x}' \in T_{\mathbf{x}_\mathcal{U}}} P(\mathbf{x}')$, which is obviously not true. Therefore, it implies the inequality (43) (and (42), equivalently) cannot be true for all $a \in \mathcal{A}$, i.e., there must exist some $a$ for which the left hand side is not smaller than the right hand side of (42), so that both sides are equal when summed over all $a$. Therefore, the initial assumption is incorrect and the lemma is proved. □

## H PROOF OF THEOREM 1

For the convenience and conciseness of the proof, as long as there is no confusion, we abuse some notation in this and the following Appendix sections. *All notation in the following Appendix sections only follow their definitions in this section.*

Recall that an optimization subproblem in (OPT-Sub) is formulated over a quasi-identifier (QID) group $T_{\mathbf{x}_\mathcal{U}}$ in which all public records are equal to $\mathbf{x}_\mathcal{U}$. Let $m = |T_{\mathbf{x}_\mathcal{U}}|$ be the cardinality of the QID group, or equivalently, the number of rows of this tuple. Let $\mathbf{x}_k$ be the unique record of row $k$ in

the tuple, $k = 1, \ldots, m$, and define $p_k \triangleq P(\mathbf{x}_k)$, $x_k \triangleq \tilde{D}_1(\mathbf{x}_k)$, and $y_k \triangleq \tilde{D}_0(\mathbf{x}_k) = 1 - x_k$. The privacy constraints can thus be re-written as

$$\frac{p_k x_k}{\sum_{i=1}^m p_i x_i} \leq \beta, \quad \forall k = 1, \ldots, m,$$
$$\frac{p_k y_k}{\sum_{i=1}^m p_i y_i} \leq \beta, \quad \forall k = 1, \ldots, m,$$

which can be combined as

$$p_k - \beta \sum_{i=1}^m p_i \leq p_k x_k - \beta \sum_{i=1}^m p_i x_i \leq 0, \tag{44}$$

$\forall k = 1, \ldots, m$. Moreover, let $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$, where $T$ represents the transpose operator. Define $\mathbf{A}$ as

$$\mathbf{A} = \begin{pmatrix} (1-\beta)p_1 & -\beta p_2 & \cdots & -\beta p_m \\ -\beta p_1 & (1-\beta)p_2 & \cdots & -\beta p_m \\ \vdots & \vdots & \ddots & \vdots \\ -\beta p_1 & -\beta p_2 & \cdots & (1-\beta)p_m \end{pmatrix}, \tag{45}$$

and let $\mathbf{b} = [b_1, b_2, \ldots, b_m]^T$, in which $b_k = p_k - \beta \sum_{i=1}^m p_i$. We can further simplify the privacy constraints as

$$\mathbf{b} \preceq \mathbf{A}\mathbf{x} \preceq \mathbf{0}, \tag{46}$$

where $\mathbf{0}$ is an $m \times 1$ zero vector.

*Remark 7.* Note that $b_k = p_k - \beta \sum_{i=1}^m p_i \leq 0$ due to Lemma 5, or (40), equivalently.

Similarly, the fidelity constraints can be re-written as

$$x_{k\min} \leq x_k \leq x_{k\max}, \quad \forall k = 1, \ldots, m,$$
$$y_{k\min} \leq y_k \leq y_{k\max}, \quad \forall k = 1, \ldots, m.$$

However, since for binary decision, $y_k = 1 - x_k$, the above two constraints are basically equivalent (to see this, simply let $y_{k\min} = 1 - x_{k\max}$ and $y_{k\max} = 1 - x_{k\min}$), so we obtain the following fidelity constraints

$$x_{k\min} \leq x_k \leq x_{k\max}, \quad \forall k = 1, \ldots, m. \tag{47}$$

Note that the $2m$ privacy constraints in (44) (or their equivalent vectorized form in (46)) form a parallelotope in the $m$-dimensional space, and the $2m$ fidelity constraints in (47) form a hypercube in the $m$-dimensional space. Let $\mathcal{P}$ denote the parallelotope and $\mathcal{H}$ denote the hypercube. Moreover, define $\mathcal{I} \triangleq \mathcal{P} \bigcap \mathcal{H}$ be the intersection of $\mathcal{P}$ and $\mathcal{H}$. $\mathcal{I} = \varnothing$ if and only if $\mathcal{P}$ and $\mathcal{H}$ are disjoint, where $\varnothing$ denotes the empty set. We have the following fact.

Fact 1. *An optimization subproblem has feasible solutions if and only if $\mathcal{I} \neq \varnothing$, i.e., $\mathcal{P}$ and $\mathcal{H}$ intersect/collide with each other.*

To prove Theorem 1, based on the above fact, it is hence equivalent to show that $\mathcal{P}$ and $\mathcal{H}$ collide with each other *if and only if* $\beta \geq \beta^*_{T_{\mathbf{x}_\mathcal{U}}} \triangleq \max\{\beta_0, \beta_1, \beta_p\}$. Let $\pi \triangleq \arg\max_k p_k x_{k\min}$ and

$\theta \triangleq \arg\max_k p_k y_{k\min}$, we can re-write $\beta_0$, $\beta_1$, and $\beta_p$ in the following

$$\beta_0 = \frac{p_\theta y_{\theta\min}}{p_\theta y_{\theta\min} + \sum_{\substack{i=1 \\ i\neq\theta}}^{m} p_i y_{i\max'}}, \tag{48}$$

$$\beta_1 = \frac{p_\pi x_{\pi\min}}{p_\pi x_{\pi\min} + \sum_{\substack{i=1 \\ i\neq\pi}}^{m} p_i x_{i\max'}}, \tag{49}$$

$$\beta_p = \frac{p_\pi x_{\pi\min} + p_\theta y_{\theta\min}}{\sum_{i=1}^{m} p_i}, \tag{50}$$

where

$$x_{i\max'} \triangleq \min\left\{x_{i\max}, \frac{p_\pi}{p_i} x_{\pi\min}\right\}, \tag{51}$$

$$y_{i\max'} \triangleq \min\left\{y_{i\max}, \frac{p_\theta}{p_i} y_{\theta\min}\right\}. \tag{52}$$

Consider the following two optimization problems for $x_j$, where $j$ is an arbitrary index, $1 \le j \le m$:

$$\begin{aligned} \text{minimize} \quad & x_j \\ \text{s.t.} \quad & \mathbf{b} \preceq \mathbf{A}\mathbf{x} \preceq \mathbf{0}, \\ & x_{k\min} \le x_k \le x_{k\max}, \text{for } k = 1, \ldots, m, k \neq j. \end{aligned} \tag{OPT-1}$$

$$\begin{aligned} \text{maximize} \quad & x_j \\ \text{s.t.} \quad & \mathbf{b} \preceq \mathbf{A}\mathbf{x} \preceq \mathbf{0}, \\ & x_{k\min} \le x_k \le x_{k\max}, \text{for } k = 1, \ldots, m, k \neq j. \end{aligned} \tag{OPT-2}$$

The above two problems have exactly the same constraints. The first line constraint forms the parallelotope $\mathcal{P}$, and let $\mathcal{H}_j'$ denote the hypercube formed by the second line constraints, i.e., $x_{k\min} \le x_k \le x_{k\max}$, for $k = 1, \ldots, m, k \neq j$. In addition, define $\mathcal{I}_j' \triangleq \mathcal{P} \bigcap \mathcal{H}_j'$ be the intersection of $\mathcal{P}$ and $\mathcal{H}_j'$, interpreting the geometric space formed by the constraints of the above two optimization problems. Moreover, if $\mathcal{I}_j' \neq \emptyset$, (i.e., there exist feasible solutions for (OPT-1) and (OPT-2)), we let $x_j^\dagger$ and $x_j^\ddagger$ denote the optimal objective values of (OPT-1) and (OPT-2), respectively. We have the following lemma.

Lemma 8. *If $\mathcal{I}_k' \neq \emptyset$ for all $k = 1, \ldots, m$, $\mathcal{P}$ and $\mathcal{H}$ are disjoint ($\mathcal{I} = \emptyset$) if and only if either $x_j^\dagger > x_{j\,max}$ or $x_j^\ddagger < x_{j\,min}$. In other words, $\mathcal{P}$ and $\mathcal{H}$ collide with each other if and only if $\mathcal{I}_k' \neq \emptyset$, $x_k^\dagger \le x_{k\,max}$, and $x_k^\ddagger \ge x_{k\,min}$, $\forall k = 1, \ldots, m$.*

Proof. Apparently, since $\mathcal{H} = \mathcal{H}_j' \bigcap \mathcal{H}_k'$ for any $k \neq j$, we have $\mathcal{I} \subseteq \mathcal{I}_j'$ true for any $j$, which implies if there exists any $j$ such that $\mathcal{I}_j' = \emptyset$, $\mathcal{I} = \emptyset$, and $\mathcal{P}$ and $\mathcal{H}$ must be disjoint. Since $\mathcal{I} \subseteq \mathcal{I}_j'$ for every $j$, if $\mathbf{x} \notin \mathcal{I}_j'$ for any $j$, then $\mathbf{x} \notin \mathcal{I}$. Moreover, for any point $\mathbf{x} \in \mathcal{I}_j'$, $x_j^\dagger \le x_j \le x_j^\ddagger$.

If $\mathcal{I}_k' \neq \emptyset$ for all $k = 1, \ldots, m$, and either $x_j^\dagger > x_{j\max}$ or $x_j^\ddagger < x_{j\min}$, since for any $\mathbf{x} \in \mathcal{I}_j'$, $x_j^\dagger \le x_j \le x_j^\ddagger$, which implies either $x_j < x_{j\min}$ or $x_j > x_{j\max}$, and thus either $\mathcal{I} = \emptyset$, or $\mathcal{I} \nsubseteq \mathcal{I}_j'$ (which violates the truth). Therefore, $\mathcal{P}$ and $\mathcal{H}$ are disjoint.

We next prove the converse. If $\mathcal{I}_k' \neq \emptyset$, $x_k^\dagger \le x_{k\max}$, and $x_k^\ddagger \ge x_{k\min}$ are true for all $k = 1, \ldots, m$, since for any $\mathbf{x} \in \mathcal{I}_k'$, $\forall k = 1, \ldots, m$, $x_k^\dagger \le x_k \le x_k^\ddagger$, we have $x_{k\min} \le x_k \le x_{k\max}$, $\forall k$, which implies

$\mathbf{x} \in \mathcal{I}$, so that $\mathcal{I} \neq \emptyset$, $\mathcal{P}$ and $\mathcal{H}$ collide with each other. We thus prove the converse and the proof is done. □

Based on Fact 1 and Lemma 8, the following statements are equivalent.

(S1) An optimization sub-problem has feasible solutions.
$\Longleftrightarrow$ (S2) $\mathcal{P}$ and $\mathcal{H}$ intersect/collide with each other.
$\Longleftrightarrow$ (S3) (OPT-1) and (OPT-2) have feasible solutions *for* ***all*** $j$.
$\Longleftrightarrow$ (S4) $\mathcal{I}_j' \neq \emptyset$, $x_j^{\dagger} \leq x_{j_{\max}}$ and $x_j^{\ddagger} \geq x_{j_{\min}}$, $\forall j = 1, \ldots, m$.

Our next goal is to show that (S1)~(S4) are true *if and only if* $\beta \geq \max\{\beta_0, \beta_1, \beta_p\}$. To show this, we need the following lemma.

Lemma 9. *Consider the optimization problem* (OPT-1) *for some (arbitrary) $j$. If (S1)~(S4) are true, we have $\beta \geq \max\{\beta_0, \beta_1, \beta_{p_j}\}$, where*

$$\beta_0 = \frac{p_\theta y_{\theta\,min}}{p_\theta y_{\theta\,min} + \sum_{\substack{k=1 \\ k\neq\theta}}^{m} p_k y_{k\,max'}}, \tag{53}$$

$$\beta_1 = \frac{p_\pi x_{\pi\,min}}{p_\pi x_{\pi\,min} + \sum_{\substack{k=1 \\ k\neq\pi}}^{m} p_k x_{k\,max'}}, \tag{54}$$

$$\beta_{p_j} = \frac{p_\pi x_{\pi\,min} + p_j y_{j\,min}}{\sum_{k=1}^{m} p_k}. \tag{55}$$

*For each of the above cases, i.e., $\beta = \beta_0$, $\beta_1$, or $\beta_{p_j}$, the corresponding optimal objective value $x_j^{\dagger}$ and its corresponding optimal solutions are*

$$\begin{aligned}
\beta = \beta_0 \Longleftrightarrow x_j^{\dagger} &= \frac{1}{p_j}\left\{p_j - \frac{\beta}{1-\beta}\sum_{\substack{k=1 \\ k\neq j}}^{m} p_k y_{k\,max'}\right\} \triangleq x_j^{\dagger 0} \\
&\Longleftrightarrow y_j = y_\theta = y_{\theta\,min} \\
&\quad\; y_k = y_{k\,max'}, \forall k = 1, \ldots, m,\ k \neq \theta, \\
\beta = \beta_1 \Longleftrightarrow x_j^{\dagger} &= \frac{1}{p_j}\left\{\frac{1-\beta}{\beta} p_\pi x_{\pi\,min} - \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_{k\,max'}\right\} \triangleq x_j^{\dagger 1} \\
&\Longleftrightarrow x_\pi = x_{\pi\,min} \\
&\quad\; x_k = x_{k\,max'}, \forall k = 1, \ldots, m,\ k \neq j, \pi, \\
\beta = \beta_{p_j} \Longleftrightarrow x_j^{\dagger} &= \frac{1}{p_j}\left\{p_\pi x_{\pi\,min} + p_j - \beta\sum_{k=1}^{m} p_k\right\} \triangleq x_j^{\dagger p} \\
&\Longleftrightarrow x_\pi = x_{\pi\,min} \\
&\quad\; \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_k = \frac{1-2\beta}{\beta} p_\pi x_{\pi\,min} - p_j + \beta\sum_{k=1}^{m} p_k.
\end{aligned}$$

Proof. Please refer to Appendix I for the proofs. □

When (S1)~(S4) are true, $\mathcal{I}_j' \neq \emptyset$ and $x_j^\dagger \leq x_{j_{\max}}$ need to be met *for all* $j = 1, \ldots, m$. Based on Lemma 9, it implies that $\beta \geq \beta_0$, $\beta \geq \beta_1$, and

$$\beta \geq \beta_{p_j} = \frac{p_\pi x_{\pi\min} + p_j y_{j_{\min}}}{\sum_{k=1}^m p_k}, \quad \forall j = 1, \ldots, m,$$

which is equivalent to

$$\beta \geq \max_j \frac{p_\pi x_{\pi\min} + p_j y_{j_{\min}}}{\sum_{k=1}^m p_k} = \frac{p_\pi x_{\pi\min} + p_\theta y_{\theta\min}}{\sum_{k=1}^m p_k} = \beta_p.$$

We then obtain $\beta \geq \max\{\beta_0, \beta_1, \beta_p\} = \beta^*_{T_{\mathbf{x}_\mathcal{U}}}$. Since when $\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_0$, based on Lemma 9, we have $y_j = y_\theta = y_{\theta\min}$. Based on (54) and (55), we have

$$\begin{aligned}
\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_1 &\iff x_j = x_{j_{\max'}}, \\
\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p &\iff y_j = y_\theta = y_{\theta\min}.
\end{aligned}$$

Combining the above results with Lemma 9, we thus have

$$\begin{aligned}
\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_0 &\iff y_\theta = y_{\theta\min} \\
&\quad\; y_k = y_{k_{\max'}}, \forall k = 1, \ldots, m,\ k \neq \theta \\
\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_1 &\iff x_\pi = x_{\pi\min} \\
&\quad\; x_k = x_{k_{\max'}}, \forall k = 1, \ldots, m,\ k \neq \pi \\
\beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \beta_p &\iff x_\pi = x_{\pi\min} \\
&\quad\; y_\theta = y_{\theta\min} \\
&\quad\; \sum_{\substack{k=1 \\ k\neq\theta,\pi}}^m p_k x_k = \frac{1-2\beta}{\beta} p_\pi x_{\pi\min} - p_j + \beta \sum_{k=1}^m p_k.
\end{aligned}$$

Similarly, if (S1)~(S4) are true, $\mathcal{I}_j' \neq \emptyset$ and $x_j^\ddagger \geq x_{j_{\min}}$ need to be met for all $j = 1, \ldots, m$. By letting $y_k = 1 - x_k$, $y_{k_{\min}} = 1 - x_{k_{\max}}$, and $y_{k_{\max}} = 1 - x_{k_{\min}}$, the optimization problem (OPT-2) is essentially equivalent to the following optimization problem:

$$\begin{aligned}
\text{minimize} \quad & y_j \qquad\qquad \text{(OPT-3)} \\
\text{s.t.} \quad & \mathbf{b} \leq \mathbf{A}\mathbf{y} \leq \mathbf{0}, \\
& y_{k_{\min}} \leq y_k \leq y_{k_{\max}}, \text{for } k = 1, \ldots, m, k \neq j.
\end{aligned}$$

Let $y_j^\dagger$ be the optimal objective value of the above optimization problem. Clearly, $y_j^\dagger = 1 - x_j^\ddagger$. Therefore, for all $j = 1, \ldots, m$, $x_j^\ddagger \geq x_{j_{\min}}$ is equivalent to $y_j^\dagger \leq y_{j_{\max}}$. By applying results from $x_j^\dagger$ in Lemma 9, we will obtain exactly the same conditions for $\beta$, i.e., $\beta \geq \beta^*_{T_{\mathbf{x}_\mathcal{U}}}$. Therefore, if (S1)~(S4) are true, we have $\beta \geq \max\{\beta_0, \beta_1, \beta_p\}$.

We next prove the converse. If $\beta \geq \max\{\beta_0, \beta_1, \beta_p\}$, which, based on (53), (54), and (55), implies $x_j^\dagger \leq x_{j_{\max}}$ and $y_j^\dagger \leq y_{j_{\max}}$, which is equivalent to $x_j^\ddagger \geq x_{j_{\min}}$, $\forall j = 1, \ldots, m$. Therefore, based on Lemma 9, $x_j$ is feasible, which implies $\mathcal{I}_j' \neq \emptyset$, $\forall j = 1, \ldots, m$, and thus (S4) is true. Since (S1)~(S4) are equivalent, an optimization sub-problem has feasible solutions *if and only if* $\beta \geq \beta^*_{T_{\mathbf{x}_\mathcal{U}}} = \max\{\beta_0, \beta_1, \beta_p\}$. We thus finish the proof.

## I PROOF OF LEMMA 9

Here we demonstrate the proof of Lemma 9, which shows the optimal objective value of the optimization problem (OPT-1).

If $\mathcal{I}_j' \neq \emptyset$, there exists (at least one or some) $\mathbf{x} \in \mathcal{I}_j'$, and for all $\mathbf{x}$, $x_j^\dagger \leq x_j \leq x_j^\ddagger$. Since $\mathcal{I}_j' = \mathcal{P} \bigcap \mathcal{H}_j'$, any $\mathbf{x} \in \mathcal{I}_j'$ also belongs to $\mathcal{P}$ and $\mathcal{H}'$. Since $\mathcal{P}$ is a $m$-dimensional parallelotope, and $\mathbf{0} \in \mathcal{P}$ is a vertex of $\mathcal{P}$, any point $\mathbf{x} \in \mathcal{P}$ can be uniquely represented by a linear combination of $m$ linear independent edge vectors emitted from $\mathbf{0}$, denoted by $\mathbf{L}_k, k = 1, \ldots, m$, and $\mathbf{x} = \sum_{k=1}^m \alpha_k \mathbf{L}_k$, $0 \leq \alpha_k \leq 1, \forall k = 1, \ldots, m$. Let $\mathbf{L}$ be the collection of these $m$ vectors; specifically, $\mathbf{L} \triangleq [\mathbf{L}_1 \mathbf{L}_2 \ldots \mathbf{L}_m]$, where $\mathbf{L}_k$ is an $m \times 1$ column vector and $\mathbf{L}$ is an $m \times m$ matrix. $\mathbf{L}$ can be obtained by

$$\mathbf{L} = \mathbf{A}^{-1}\mathbf{B}, \tag{56}$$

where $\mathbf{A}$ is defined in (45) and $\mathbf{B} = \mathrm{dg}(\mathbf{b})$ where $\mathrm{dg}(\mathbf{b})$ denotes a diagonal matrix with elements of $\mathbf{b} = (b_1, b_2, \ldots, b_m)$ along the diagonal. To find $\mathbf{A}^{-1}$, note that since $\mathbf{A}$ can be represented by

$$\mathbf{A} = \mathrm{dg}(\mathbf{p}) + (-\beta)\mathbf{1}_m \mathbf{p}^T, \tag{57}$$

where $\mathbf{p} = [p_1, p_2, \ldots, p_m]^T$ and $\mathbf{1}_m$ is an all-one vector with $m$ elements, we can thus apply the following matrix inversion formula [48]

$$(\mathbf{Z} + c\mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{Z}^{-1} - \frac{1}{1 + c\mathbf{v}^T\mathbf{Z}^{-1}\mathbf{u}}\mathbf{Z}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{Z}^{-1} \tag{58}$$

to compute $\mathbf{A}^{-1}$ and obtain $\mathbf{L}$ as follows

$$\mathbf{L} = \frac{1}{1 - m\beta}\begin{pmatrix} \frac{b_1}{p_1}[1-(m-1)\beta] & \frac{b_2}{p_1}\beta & \cdots & \frac{b_m}{p_1}\beta \\ \frac{b_1}{p_2}\beta & \frac{b_2}{p_2}[1-(m-1)\beta] & \cdots & \frac{b_m}{p_2}\beta \\ \vdots & \vdots & \ddots & \vdots \\ \frac{b_1}{p_m}\beta & \frac{b_2}{p_m}\beta & \cdots & \frac{b_m}{p_m}[1-(m-1)\beta] \end{pmatrix}. \tag{59}$$

Define $\frac{\mathbf{b}}{\mathbf{p}} \triangleq (\frac{b_1}{p_1}, \frac{b_2}{p_2}, \ldots, \frac{b_m}{p_m})$ as the element-wise division operation of two vectors. It is not hard to see that $\mathbf{L}$ can be represented as the following equivalent form

$$\mathbf{L} = \mathrm{dg}\left(\frac{\mathbf{b}}{\mathbf{p}}\right) + \frac{\beta}{1 - m\beta}\frac{\mathbf{1}}{\mathbf{p}}\mathbf{b}^T, \tag{60}$$

which implies that its inverse can also be found by applying the matrix inversion formula in (58). We will utilize this property in the later of the proof.

Recall that if $\mathbf{x} \in \mathcal{I}_j'$, $\mathbf{x} \in \mathcal{P}$ as well. Therefore, any $\mathbf{x} \in \mathcal{I}_j'$ can be uniquely represented by

$$\mathbf{x} = \sum_{k=1}^m \alpha_k \mathbf{L}_k, \tag{61}$$

in which $0 \leq \alpha_k \leq 1, \forall k = 1, \ldots, m$. Recall that we are solving the optimization problem (OPT-1) for some $j$, $1 \leq j \leq m$. We first take out the $j$-th row from (61),

$$x_j = \sum_{k=1}^m \alpha_k L_{k,j}, \tag{62}$$

and for the rest $m-1$ equalities, we move the term $\alpha_j \mathbf{L}_j$ from the right-hand-side (RHS) to the left-hand-side (LHS). We then obtain

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{j-1} \\ x_{j+1} \\ \vdots \\ x_m \end{bmatrix} - \alpha_j \begin{bmatrix} L_{1,j} \\ L_{2,j} \\ \vdots \\ L_{j-1,j} \\ L_{j+1,j} \\ \vdots \\ L_{m,j} \end{bmatrix} = \begin{bmatrix} L_{1,1} & L_{1,2} & \cdots & L_{1,j-1} & L_{1,j+1} & \cdots & L_{1,m} \\ L_{2,1} & L_{2,2} & \cdots & L_{2,j-1} & L_{2,j+1} & \cdots & L_{2,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ L_{j-1,1} & L_{j-1,2} & \cdots & L_{j-1,j-1} & L_{j-1,j+1} & \cdots & L_{j-1,m} \\ L_{j+1,1} & L_{j+1,2} & \cdots & L_{j+1,j-1} & L_{j+1,j+1} & \cdots & L_{j+1,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ L_{m,1} & L_{m,2} & \cdots & L_{m,j-1} & L_{m,j+1} & \cdots & L_{m,m} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{j-1} \\ \alpha_{j+1} \\ \vdots \\ \alpha_m \end{bmatrix}, \tag{63}$$

and let $\mathbf{x}' - \alpha_j \mathbf{L}_j^{'} = \mathbf{L}_{\text{sub}} \boldsymbol{\alpha}'$ be the corresponding vector form of (63), in which, based on (59), $L_{k,k} = \frac{1-(m-1)\beta}{1-m\beta} \frac{b_k}{p_k}$, $\forall k = 1, \ldots, m$, and $L_{k,i} = \frac{\beta}{1-m\beta} \frac{b_i}{p_k}$, $\forall k, i = 1, \ldots, m$, $k \neq i$. Note that $\mathbf{L}_{\text{sub}}$ is an $(m-1) \times (m-1)$ square sub-matrix of $\mathbf{L}$ by removing the $j$-th row and the $j$-th column. Therefore, it has the similar form as shown in (60) by removing the $j$-th row/element of $\mathbf{b}$ and $\mathbf{p}$, and thus its inverse $\mathbf{L}_{\text{sub}}^{-1}$ can also be found by (58). By applying $\mathbf{L}_{\text{sub}}^{-1}$ to both sides of (63), we have

$$\boldsymbol{\alpha}' = \mathbf{L}_{\text{sub}}^{-1} \mathbf{x}' - \alpha_j \mathbf{L}_{\text{sub}}^{-1} \mathbf{L}_j^{'}. \tag{64}$$

Let $\mathbf{u} \triangleq \mathbf{L}_{\text{sub}}^{-1} \mathbf{x}'$ and $\mathbf{v} \triangleq \alpha_j \mathbf{L}_{\text{sub}}^{-1} \mathbf{L}_j^{'}$, we obtain

$$\begin{aligned} u_k &= \frac{1}{1-\beta} \frac{1}{b_k} \left[ (1-\beta) p_k x_k - \beta \sum_{\substack{i=1 \\ i \neq j}}^{m} p_i x_i \right], \\ v_k &= \frac{\beta}{1-\beta} \frac{1}{b_k} \alpha_j b_j, \\ \alpha_k &= u_k - v_k = \frac{\beta}{1-\beta} \frac{1}{b_k} \left[ \frac{1-\beta}{\beta} p_k x_k - \sum_{\substack{i=1 \\ i \neq j}}^{m} p_i x_i - \alpha_j b_j \right], \end{aligned} \tag{65}$$

for all $k = 1, \ldots, m$, $k \neq j$. Similarly, if we define $\alpha_k{}' \triangleq 1 - \alpha_k$, we have

$$\alpha_k{}' = \frac{\beta}{1-\beta} \frac{1}{b_k} \left[ \frac{1-\beta}{\beta} p_k y_k - \sum_{\substack{i=1 \\ i \neq j}}^{m} p_i y_i - \alpha_j{}' b_j \right]. \tag{66}$$

Substituting the $\alpha_k$ in (65) into (62), we obtain

$$x_j = \sum_{k=1}^{m} \alpha_k L_{k,j} = \frac{1}{1-\beta} \frac{1}{p_j} \left[ \beta \sum_{\substack{k=1 \\ k \neq j}}^{m} p_k x_k + \alpha_j b_j \right]. \tag{67}$$

Based on (67), we are looking for the values of $\alpha_j$ and $x_k$'s ($k \neq j$) yielding the minimum of $x_j$.

Since $0 \leq \alpha_k \leq 1$, $\forall k = 1, \ldots, m$, which implies $u_k \geq v_k$, $\forall k = 1, \ldots, m$, $k \neq j$, and

$$\frac{u_k}{v_k} = \frac{1}{\beta \alpha_j b_j} \left[ (1-\beta) p_k x_k - \beta \sum_{\substack{i=1 \\ i \neq j}}^{m} p_i x_i \right] \geq 1. \tag{68}$$

We then have

$$\alpha_j \leq \frac{1}{\beta b_j}\left[(1-\beta)p_k x_k - \beta\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_i\right] \triangleq R_j^k, \tag{69}$$

for all $k = 1, \ldots, m, k \neq j$. Let $R_j^* \triangleq \min_k R_j^k$. Combining with the fact that $\alpha_j \leq 1$, we obtain

$$\alpha_j \leq \min(R_j^*, 1). \tag{70}$$

**Case 1**:
We first consider the case $R_j^* \leq 1$. Please refer to Case 2 for $R_j^* \geq 1$.

When $R_j^* \leq 1$, based on (70), we have $\alpha_j = R_j^*$. Therefore, based on (69), there must exist a $k$ (denoted by $\kappa$) such that

$$\alpha_j = \frac{1}{\beta b_j}\left[(1-\beta)p_\kappa x_\kappa - \beta\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_i\right] = R_j^* \leq 1. \tag{71}$$

Because $b_j \leq 0$ (see Remark 7), from the LHS of (71), we have

$$\beta\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_i + \alpha_j b_j = (1-\beta)\left[\frac{1}{\beta}p_\kappa x_\kappa - \sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_i\right]. \tag{72}$$

Substitute the LHS of (72) into (67). We obtain

$$p_j x_j = \frac{1}{\beta}p_\kappa x_\kappa - \sum_{\substack{k=1\\ k\neq j}}^{m} p_i x_i = \frac{1-\beta}{\beta}p_\kappa x_\kappa - \sum_{\substack{k=1\\ k\neq j,\kappa}}^{m} p_i x_i, \tag{73}$$

or equivalently,

$$\beta = \frac{p_\kappa x_\kappa}{\sum_{k=1}^{m} p_k x_k} = \frac{p_\kappa x_\kappa}{p_\kappa x_\kappa + \sum_{\substack{k=1\\ k\neq\kappa}}^{m} p_k x_k}. \tag{74}$$

Therefore, $x_j$ is minimized if the RHS of (73) is minimized. In addition, in Case 1, based on (74), $\beta$ achieves its minimum when (i) minimizing $p_\kappa x_\kappa$ and (ii) maximizing $\sum_{\substack{k=1\\ k\neq\kappa}}^{m} p_k x_k$.

We next find the minimum of the RHS of (73). Since from (71), we have

$$\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_i = \sum_{\substack{k=1\\ k\neq j}}^{m} p_k x_k = \frac{1-\beta}{\beta}p_\kappa x_\kappa - \alpha_j b_j, \tag{75}$$

and from (73), we have

$$\beta\sum_{\substack{k=1\\ k\neq j}}^{m} p_k x_k = p_\kappa x_\kappa - \beta p_j x_j. \tag{76}$$

Substituting the LHS of (75) into (65), we obtain

$$\alpha_k = \frac{1}{b_k}(p_k x_k - p_\kappa x_\kappa), \forall k = 1, \ldots, m, k \neq j, \tag{77}$$

and substituting the LHS of (75) into (76), we obtain

$$\alpha_j = \frac{1}{b_j}(p_j x_j - p_\kappa x_\kappa). \tag{78}$$

Combining (77) and (78), we have

$$\alpha_k = \frac{1}{b_k}(p_k x_k - p_\kappa x_\kappa), \forall k = 1, \ldots, m. \tag{79}$$

Since $\alpha_k \geq 0$ and $b_k \leq 0$, we have $p_\kappa x_\kappa \geq p_k x_k$, $\forall k$, i.e., $p_\kappa x_\kappa = \max_k p_k x_k$. Moreover, from (73), since $\beta$ is non-negative and $x_k \geq 0$ for all $k$, in order to minimize $x_j$, we need to (i) minimize $p_\kappa x_\kappa$ and (ii) maximize $\sum_{\substack{k=1 \\ k\neq j,\kappa}}^{m} p_k x_k$. Note that both (i) and (ii) minimize $\beta$ as well, which implies that the optimal solutions that minimize $x_j$ also minimize $\beta$.

To minimize $p_\kappa x_\kappa$, since $p_\kappa x_\kappa = \max_k p_k x_k$, and $p_k x_k \geq p_k x_{k\min}$, $\forall k$ (including $\kappa$), the minimal $p_\kappa x_\kappa$, i.e., $p_\kappa x_{\kappa\min}$, is therefore the largest effective lower limit $p_k x_{k\min}$ over all $k$, i.e., $p_\kappa x_{\kappa\min} = \max_k p_k x_{k\min} = p_\pi x_{\pi\min}$ by definition, and thus we get $\kappa = \pi$ and $x_\pi = x_{\pi\min}$.

To maximize $\sum_{\substack{k=1 \\ k\neq j,\kappa}}^{m} p_k x_k$, we need to find the maximum of each $x_k$. Since $0 \leq \alpha_k \leq 1$, by substituting $p_\kappa x_\kappa = p_\pi x_{\pi\min}$ into (79), we have $p_\pi x_{\pi\min} + b_k \leq p_k x_k \leq p_\pi x_{\pi\min}$, $\forall k = 1, \ldots, m$ (including $j$). By definitions, $x_{k\max'} \triangleq \min\{x_{k\max}, \frac{p_\pi}{p_k} x_{\pi\min}\}$ and $x_{k\min'} \triangleq \max\{x_{k\min}, \frac{p_\pi}{p_k} x_{\pi\min} + b_k\}$. Combining with the constraints $x_{k\min} \leq x_k \leq x_{k\max}$, we obtain $x_{k\min'} \leq x_k \leq x_{k\max'}$, $\forall k = 1, \ldots, m$, and thus $\sum_{\substack{k=1 \\ k\neq j,\kappa}}^{m} p_k x_k$ is maximized when $x_k = x_{k\max'}$, $\forall k, k \neq j, \pi$.

By substituting the $x_k$'s we obtained above into (73), based on which, the optimal objective value $x_j^\dagger$, the minimum of $x_j$, is thus

$$x_j^\dagger = \frac{1}{p_j}\left\{\frac{1-\beta}{\beta} p_\pi x_{\pi\min} - \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_i x_{i\max'}\right\} \triangleq x_j^{\dagger 1}. \tag{80}$$

If (S1)~(S4) are true, $x_j^\dagger \leq x_{j\max}$. In addition, based on (78), we have $p_j x_j = p_\pi x_{\pi\min} + \alpha_j b_j \leq p_\pi x_{\pi\min}$, and thus from (51), we get $x_{j\max} = x_{j\max'}$. Therefore, based on (80), the minimum $\beta$ such that (OPT-1) has feasible solution, for Case 1, is

$$\begin{aligned} \beta &= \frac{p_\pi x_{\pi\min}}{p_\pi x_{\pi\min} + p_j x_j^\dagger + \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_{k\max'}} \\ &\geq \frac{p_\pi x_{\pi\min}}{p_\pi x_{\pi\min} + \sum_{\substack{k=1 \\ k\neq \pi}}^{m} p_k x_{k\max'}} \triangleq \beta_1. \end{aligned} \tag{81}$$

We next prove the converse. If (81) and (80) holds, i.e., $x_\kappa = x_\pi = x_{\pi\min}$, and $x_k = x_{k\max'}$, $\forall k = 1, \ldots, m, k \neq j, \pi$, since $x_{\pi\min} = x_{\pi\max'}$, we have $x_k = x_{k\max'}$, $\forall k = 1, \ldots, m, k \neq j$, and

$$\begin{aligned} p_j x_j^\dagger &= \frac{1-\beta}{\beta} p_\pi x_{\pi\min} - \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_{k\max'} \\ &= \frac{1}{\beta} p_\pi x_{\pi\min} - \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_{k\max'}. \end{aligned} \tag{82}$$

Given the above $x_k$'s and (82), since $x_{\pi\min} = x_{\pi\max'}$ and $b_j \le 0$, and by definition in (51), $p_\pi x_{\pi\min} \ge p_k x_{k\max'}$, $\forall k$, we have

$$\begin{aligned} R_j^* &\triangleq \min_k R_j^k \\ &= \min_k \frac{1}{\beta b_j}\left[(1-\beta)p_k x_{k\max'} - \beta\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_{i\max'}\right] \\ &= \frac{1}{\beta b_j}\max_k \left[(1-\beta)p_k x_{k\max'} - \beta\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_{i\max'}\right] \\ &= \frac{1}{\beta b_j}\left[(1-\beta)p_\pi x_{\pi\min} - \beta\sum_{\substack{i=1\\ i\neq j}}^{m} p_i x_{i\max'}\right] \\ &= \frac{1}{b_j}(p_j x_j^\dagger - p_\pi x_{\pi\min}). \end{aligned} \tag{83}$$

Besides, by substituting $x_k = x_{k\max'}$, $\forall k = 1,\ldots,m$, $k \neq j$, into (67), we obtain

$$p_j x_j^\dagger = \frac{1}{1-\beta}\left[\beta\sum_{\substack{k=1\\ k\neq j}}^{m} p_k x_{k\max'} + \alpha_j b_j\right]. \tag{84}$$

By substituting the RHS of (82) into (84), we get

$$\alpha_j = \frac{1}{b_j}\left(p_j x_j^\dagger - p_\pi x_{\pi\min}\right). \tag{85}$$

Since $\alpha_j \le 1$, and according to (83), we obtain $R_j^* = \alpha_j \le 1$, and thus finish the proof of the converse.

**We summarize Case 1 in the following:**

$$\begin{aligned} & R_j^* \le 1 \\ \iff & \frac{1-\beta}{\beta}p_\pi x_{\pi\min} - b_j \ge \sum_{\substack{k=1\\ k\neq j}}^{m} p_i x_{i\max'} \\ \iff & x_\pi = x_{\pi\min} \text{ and } x_k = x_{k\max'}, \forall k = 1,\ldots,m, k\neq j,\pi \\ \iff & x_j^\dagger = \frac{1}{p_j}\left\{\frac{1-\beta}{\beta}p_\pi x_{\pi\min} - \sum_{\substack{k=1\\ k\neq j,\pi}}^{m} p_k x_{k\max'}\right\} \triangleq x_j^{\dagger 1} \\ \iff & \beta \ge \frac{p_\pi x_{\pi\min}}{p_\pi x_{\pi\min} + \sum_{\substack{k=1\\ k\neq\pi}}^{m} p_k x_{k\max'}} \triangleq \beta_1. \end{aligned}$$

**Case 2**:

Now consider the case $R_j^* \ge 1$. In this case, according to (70), we have $\alpha_j = \min(R_j^*, 1) = 1$, and (67) thus becomes

$$p_j x_j = \frac{1}{1-\beta}\left[\beta \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_k + b_j\right]. \tag{86}$$

Moreover, based on (69), there must exist a $k$ (denoted by $\kappa$) such that

$$\alpha_j = 1 \leq \frac{1}{\beta b_j}\left[(1-\beta)p_\kappa x_\kappa - \beta \sum_{\substack{i=1 \\ i\neq j}}^{m} p_i x_i\right] = R_j^*, \tag{87}$$

which, since $x_i \leq x_{i\max'}$, $\forall i$, yields

$$\sum_{\substack{i=1 \\ i\neq j}}^{m} p_i x_{i\max'} \geq \sum_{\substack{i=1 \\ i\neq j}}^{m} p_i x_i \geq \frac{1-\beta}{\beta} p_\kappa x_\kappa - b_j. \tag{88}$$

Since $\sum_{\substack{i=1 \\ i\neq j}}^{m} p_i x_i \geq \sum_{\substack{i=1 \\ i\neq j}}^{m} p_i x_{i\min'}$ as well, we need to consider different cases in the following in order to proceed the proof.

**Case 2.1**:
First, we consider the case that the RHS of (88) is greater or equal to the sum of the equivalent lower limits, i.e.,

$$\frac{1-\beta}{\beta} p_\kappa x_\kappa - b_j \geq \sum_{\substack{i=1 \\ i\neq j}}^{m} p_i x_{i\min'}. \tag{89}$$

In such a case, the equality of the RHS of (88) can hold. Since, based on (86), in order to minimize $x_j$, we need to minimize $\sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_k$, from the RHS of (88), which is

$$\sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_k = \frac{1-\beta}{\beta} p_\kappa x_\kappa - b_j, \tag{90}$$

and thus we need to minimize $p_\kappa x_\kappa$. By substituting the LHS of (90) into (86) and (65), we obtain

$$p_j x_j = p_\kappa x_\kappa + b_j \tag{91}$$

and

$$\alpha_k = \frac{1}{b_k}(p_k x_k - p_\kappa x_\kappa), \forall k = 1, \ldots, m, k \neq j, \tag{92}$$

respectively. Based on (92), similar to Case 1, since $\alpha_k \geq 0$ and $b_k \leq 0$, we have $p_\kappa x_\kappa \geq p_k x_k$, $\forall k$, i.e., $p_\kappa x_\kappa = \max_k p_k x_k$. Moreover, in order to minimize $x_j$, we need to minimize $p_\kappa x_\kappa$. To minimize $p_\kappa x_\kappa$, since $p_\kappa x_\kappa = \max_k p_k x_k$, and $p_k x_k \geq p_k x_{k\min}$, $\forall k$ (including $\kappa$), the minimal $p_\kappa x_\kappa$, i.e., $p_\kappa x_{\kappa\min}$, is therefore the largest effective lower limit $p_k x_{k\min}$ over all $k$, i.e., $p_\kappa x_{\kappa\min} = \max_k p_k x_{k\min} = p_\pi x_{\pi\min}$ by definition, and thus we get $\kappa = \pi$ and $x_\pi = x_{\pi\min}$. Substituting which into (91), we thus obtain

the minimum of $x_j$

$$\begin{aligned} x_j^{\dagger} &= \frac{1}{p_j}\left\{p_\pi x_{\pi\min} + b_j\right\} \\ &= \frac{1}{p_j}\left\{p_\pi x_{\pi\min} + p_j - \beta\sum_{k=1}^{m} p_k\right\} \triangleq x_j^{\dagger p}. \end{aligned} \tag{93}$$

For the rest of $k$, $k \neq j, \pi$, according to (90), we have

$$\begin{aligned} \sum_{\substack{k=1\\k\neq j,\pi}}^{m} p_k x_k &= \frac{1-2\beta}{\beta} p_\pi x_{\pi\min} - b_j \\ &= \frac{1-2\beta}{\beta} p_\pi x_{\pi\min} - p_j + \beta\sum_{k=1}^{m} p_k. \end{aligned} \tag{94}$$

If (S1)~(S4) are true, $x_j^{\dagger} \leq x_{j_{\max}}$. In addition, based on (91), we have $p_j x_j = p_\pi x_{\pi\min} + b_j \leq p_\pi x_{\pi\min}$, and thus from (51), we get $x_{j_{\max}} = x_{j_{\max'}}$. Therefore, based on (93), the minimum $\beta$ such that (OPT-1) has feasible solution, for Case 2.1, is

$$\begin{aligned} \beta &= \frac{p_\pi x_{\pi\min} + p_j\left(1 - x_j^{\dagger}\right)}{\sum_{k=1}^{m} p_k} \\ &\geq \frac{p_\pi x_{\pi\min} + p_j(1 - x_{j_{\max}})}{\sum_{k=1}^{m} p_k} \\ &= \frac{p_\pi x_{\pi\min} + p_j y_{j_{\min}}}{\sum_{k=1}^{m} p_k} \triangleq \beta_{p_j}. \end{aligned} \tag{95}$$

We next prove the converse. If $x_\kappa = x_\pi = x_{\pi\min}$, and (93), (94), and (95) hold, from (67), we have

$$p_j x_j = \frac{1}{1-\beta}\left[\beta\sum_{\substack{k=1\\k\neq j}}^{m} p_k x_k + \alpha_j b_j\right], \tag{96}$$

and, since $\sum_{\substack{k=1\\k\neq j}}^{m} p_k x_k \geq \sum_{\substack{k=1\\k\neq j}}^{m} p_k x_{k\min'}$, from (94), we have

$$\sum_{\substack{k=1\\k\neq j}}^{m} p_k x_k = \frac{1-\beta}{\beta} p_\pi x_{\pi\min} - b_j \geq \sum_{\substack{k=1\\k\neq j}}^{m} p_k x_{k\min'}. \tag{97}$$

Since (94) and (97) hold when $x_j = x_j^{\dagger}$, i.e., (93) holds, by substituting the LHS of (97) into (96), we get

$$p_j x_j^{\dagger} = p_\pi x_{\pi\min} + \left(\frac{\alpha_j - \beta}{1-\beta}\right) b_j. \tag{98}$$

Comparing (98) with (93), we have $\frac{\alpha_j - \beta}{1-\beta} = 1$, and therefore $\alpha_j = 1$. Based on (70), we thus obtain $R_j^* \geq 1$ and finish the proof of the converse.

**We summarize Case 2.1 in the following:**

$$
\begin{aligned}
& R_j^* \geq 1 \text{ and } \frac{1-\beta}{\beta} p_\pi x_{\pi\min} - b_j \geq \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_{k\min'} \\
\Longleftrightarrow & \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_{k\max'} \geq \frac{1-\beta}{\beta} p_\pi x_{\pi\min} - b_j \geq \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_{k\min'} \\
\Longleftrightarrow & x_\pi = x_{\pi\min} \text{ and } \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_k = \frac{1-2\beta}{\beta} p_\pi x_{\pi\min} - p_j + \beta \sum_{k=1}^{m} p_k \\
\Longleftrightarrow & x_j^\dagger = \frac{1}{p_j} \left\{ p_\pi x_{\pi\min} + p_j - \beta \sum_{k=1}^{m} p_k \right\} \triangleq x_j^{\dagger p} \\
\Longleftrightarrow & \beta \geq \frac{p_\pi x_{\pi\min} + p_j y_{j\min}}{\sum_{k=1}^{m} p_k} \triangleq \beta_{p_j}.
\end{aligned}
$$

**Case 2.2**:
Next, we consider the case that

$$
\frac{1-\beta}{\beta} p_\pi x_{\pi\min} - b_j \leq \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k x_{k\min'}. \tag{99}
$$

Recall that $y_k \triangleq 1 - x_k$, $y_{k\max} \triangleq 1 - x_{k\min}$, $y_{k\min} \triangleq 1 - x_{k\max}$, and ${\alpha_k}' \triangleq 1 - \alpha_k$, $\forall k$. Since from (87), ${\alpha_j}' = 1 - \alpha_j = 0$, based on (66), we have

$$
{\alpha_k}' = \frac{\beta}{1-\beta} \frac{1}{b_k} \left[ \frac{1-\beta}{\beta} p_k y_k - \sum_{\substack{i=1 \\ i\neq j}}^{m} p_i y_i \right], \forall k = 1, \ldots, m, k \neq j. \tag{100}
$$

Moreover, by substituting $x_k = 1 - y_k$, $\forall k$, into (86), we get

$$
p_j x_j = p_j - \frac{\beta}{1-\beta} \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k y_k, \tag{101}
$$

or equivalently,

$$
p_j y_j = \frac{\beta}{1-\beta} \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k y_k, \tag{102}
$$

and

$$
\beta = \frac{p_j y_j}{\sum_{k=1}^{m} p_k y_k} = \frac{p_j y_j}{p_j y_j + \sum_{\substack{k=1 \\ k\neq j}}^{m} p_k y_k}. \tag{103}
$$

Substituting the RHS of (102) into (100), we obtain

$$
{\alpha_k}' = \frac{1}{b_k} (p_k y_k - p_j y_j), \forall k = 1, \ldots, m, k \neq j, \tag{104}
$$

Since $\alpha_k' \geq 0$ and $b_k \leq 0$, we have $p_j y_j \geq p_k y_k$, $\forall k = 1, \ldots, m$, $k \neq j$, i.e., $p_j y_j = \max_k p_k y_k$. Moreover, from (101), since $\beta$ is non-negative and $x_k \geq 0$ for all $k$, in order to minimize $x_j$, we need to maximize $\sum_{\substack{k=1 \\ k \neq j}}^{m} p_k y_k$. In addition, based on (103), in Case 2.2, $\beta$ achieves its minimum when (i) minimizing $p_j y_j$ and (ii) maximizing $\sum_{\substack{k=1 \\ k \neq j}}^{m} p_k y_k$.

To minimize $p_j y_j$, since $p_j y_j = \max_k p_k y_k$, and $p_k y_k \geq p_k y_{k\min}$, $\forall k$ (including $j$), the minimal $p_j y_j$, i.e., $p_j y_{j\min}$, is therefore the largest effective lower limit $p_k y_{k\min}$ over all $k$, i.e., $p_j y_{j\min} = \max_k p_k y_{k\min} = p_\theta y_{\theta\min}$ by definition, and thus we get $j = \theta$ and $y_\theta = y_{\theta\min}$.

To maximize $\sum_{\substack{k=1 \\ k \neq j}}^{m} p_k y_k$, we need to find the maximum of each $y_k$. Since $0 \leq \alpha_k' \leq 1$, by substituting $p_j y_j = p_\theta y_{\theta\min}$ into (104), we have $p_\theta y_{\theta\min} + b_k \leq p_k y_k \leq p_\theta y_{\theta\min}$, $\forall k = 1, \ldots, m$, $k \neq \theta$. By definitions, $y_{k\max'} \triangleq \min\{y_{k\max}, \frac{p_\theta}{p_k} y_{\theta\min}\}$ and $y_{k\min'} \triangleq \max\{y_{k\min}, \frac{p_\theta}{p_k} y_{\theta\min} + b_k\}$. Combining with the constraints $y_{k\min} \leq y_k \leq y_{k\max}$, we obtain $y_{k\min'} \leq y_k \leq y_{k\max'}$, $\forall k = 1, \ldots, m$, and thus $\sum_{\substack{k=1 \\ k \neq j}}^{m} p_k y_k$ is maximized when $y_k = y_{k\max'}$, $\forall k$, $k \neq j$.

By substituting the $y_k$'s we obtained above into (101), based on which, the optimal objective value $x_j^\dagger$, the minimum of $x_j$, is thus

$$x_j^\dagger = \frac{1}{p_j} \left\{ p_j - \frac{\beta}{1-\beta} \sum_{\substack{k=1 \\ k \neq j}}^{m} p_k y_{k\max'} \right\} \triangleq x_j^{\dagger 0}. \tag{105}$$

If (S1)~(S4) are true, $x_j^\dagger \leq x_{j\max}$. Based on (103), the minimum $\beta$ such that (OPT-1) has feasible solution, for Case 2.2, is

$$\beta \geq \frac{p_\theta y_{\theta\min}}{p_\theta y_{\theta\min} + \sum_{\substack{k=1 \\ k \neq \theta}}^{m} p_k y_{k\max'}} \triangleq \beta_0. \tag{106}$$

We next prove the converse. Before proving the converse, we first need the following lemma.

Lemma 10. *If $p_\theta y_{\theta\,min} \leq p_k y_k - b_k$, $\forall k = 1, \ldots, m$, $k \neq \theta$, we have $p_k x_{k\,min'} \leq p_k(1 - y_{k\,max'})$, $\forall k = 1, \ldots, m$, $k \neq \theta$.*

Proof. Recall that $\forall k = 1, \ldots, m$, by definitions we have

$$p_k x_{k\min'} \triangleq \max\{p_\pi x_{\pi\min} + b_k, p_k x_{k\min}\}, \tag{107}$$

$$\begin{aligned} p_k(1 - y_{k\max'}) &= p_k - p_k y_{k\max'} \\ &\triangleq \max\{p_k - p_\theta y_{\theta\min}, p_k - p_k y_{k\max}\} \\ &= \max\{p_k - p_\theta y_{\theta\min}, p_k x_{k\min}\}. \end{aligned} \tag{108}$$

Given the conditions that $p_\theta y_{\theta\min} \leq p_k y_k - b_k$, $\forall k = 1, \ldots, m$, $k \neq \theta$, from which we get

$$\begin{aligned} & p_\theta y_{\theta\min} \leq p_k y_k - b_k, \forall k = 1, \ldots, m, k \neq \theta, \\ \Longrightarrow\ & p_\theta y_{\theta\min} \leq p_\pi y_\pi - b_\pi \\ \Longrightarrow\ & p_\theta y_{\theta\min} \leq p_\pi(1 - x_\pi) - p_\pi + \beta \sum_{k=1}^{m} p_k \\ \Longrightarrow\ & p_\pi x_\pi = p_\pi x_{\pi\min} \leq -p_\theta y_{\theta\min} + \beta \sum_{k=1}^{m} p_k \end{aligned}$$

$$\begin{aligned}
&\Longrightarrow p_\pi x_{\pi\min} + b_k \le -p_\theta y_{\theta\min} + \beta \sum_{k=1}^{m} p_k + p_k - \beta \sum_{k=1}^{m} p_k \\
&\Longrightarrow p_\pi x_{\pi\min} + b_k \le p_k - p_\theta y_{\theta\min}.
\end{aligned} \tag{109}$$

Therefore, since the above inequalities hold for all $k$ except $k = \theta$, by comparing the RHS of (107) and (108), we obtain that $p_k x_{k\min'} \le p_k(1 - y_{k\max'})$, $\forall k = 1, \ldots, m$, $k \ne \theta$. We thus finish the proof. □

Now we start proving the converse. If $y_j \triangleq 1 - x_j = y_\theta = y_{\theta\min}$, $y_k \triangleq 1 - x_k = y_{k\max'}$, $\forall k, k \ne j$, (105) and the equality in (106) hold, since from (87) we have $\alpha_j = 1$, or equivalently, $\alpha_j{}' = 0$, by substituting the above $y_k$'s and $\alpha_j{}'$ into (66), we obtain

$$\alpha_k{}' = \frac{\beta}{1-\beta}\frac{1}{b_k}\left[\frac{1-\beta}{\beta}p_k y_k - \sum_{\substack{i=1 \\ i\ne j}}^{m} p_i y_{i\max'}\right], \forall k, k \ne j, \tag{110}$$

and by substituting the $x_k$'s transformed from the above $y_k$'s (including $k = j = \theta$) into (86), we have

$$\sum_{\substack{k=1 \\ k\ne\theta}}^{m} p_k x_k = \sum_{\substack{k=1 \\ k\ne\theta}}^{m} p_k(1 - y_{k\max'}) = \frac{1-\beta}{\beta}p_\theta x_\theta - \frac{1}{\beta}b_\theta. \tag{111}$$

From (110), since $j = \theta$, by substituting (106) into (110), we obtain

$$\alpha_k{}' = \frac{1}{b_k}(p_k y_k - p_\theta y_{\theta\min}), \forall k = 1, \ldots, m, k \ne \theta. \tag{112}$$

Since $0 \le \alpha_k{}' \le 1$, from (112), we obtain $p_\theta y_{\theta\min} + b_k \le p_k y_k \le p_\theta y_{\theta\min}$, $\forall k = 1, \ldots, m$, $k \ne \theta$. Therefore, based on Lemma 10, we have $p_k x_{k\min'} \le p_k(1 - y_{k\max'})$, $\forall k = 1, \ldots, m$, $k \ne \theta$.

Recall that based on (108), for each $k$, $k \ne \theta$, $p_k(1 - y_{k\max'})$ is the maximum of $p_k - p_\theta y_{\theta\min}$ and $p_k x_{k\min}$. Define $\Phi$ the set of $k$'s yielding $p_k(1 - y_{k\max'}) = p_k x_{k\min} \ge p_k - p_\theta y_{\theta\min}$, $k \in \Phi$, and define $\Omega$ the complement of $\Phi$, i.e., the set of $k$'s yielding $p_k(1 - y_{k\max'}) = p_k - p_\theta y_{\theta\min} > p_k x_{k\min}$, $k \in \Omega$. In addition, let $\phi \triangleq |\Phi|$ and $\omega \triangleq |\Omega|$. Note that $\phi + \omega = m - 1$. We have the following lemma.

Lemma 11. *If $y_\theta = y_{\theta\,min}$, $y_k = y_{k\,max'}$, $\forall k, k \ne \theta$, and the equality in (106) holds, we have $\omega \le \frac{1-\beta}{\beta}$.*

Proof. If $y_\theta = y_{\theta\min}$, $y_k = y_{k\max'}$, $\forall k, k \ne \theta$, and the equality in (106) holds, from (106) we have

$$\frac{\beta}{1-\beta} = \frac{p_\theta y_{\theta\min}}{\sum_{\substack{k=1 \\ k\ne\theta}}^{m} p_k y_{k\max'}}. \tag{113}$$

Note that since for those $k$'s in $\Omega$, we have $p_k(1 - y_{k\max'}) = p_k - p_\theta y_{\theta\min}$, or equivalently, $p_k y_{k\max'} = p_\theta y_{\theta\min}$, and for those $k$'s in $\Phi$, we have $p_k(1 - y_{k\max'}) = p_k x_{k\min}$, or equivalently, $p_k y_{k\max'} = p_k y_{k\max}$, (113) thus becomes

$$\frac{p_\theta y_{\theta\min}}{\sum_{\substack{k=1 \\ k\ne\theta}}^{m} p_k y_{k\max'}} = \frac{p_\theta y_{\theta\min}}{\omega p_\theta y_{\theta\min} + \sum_{k\in\Phi} p_k y_{k\max}} \le \frac{1}{\omega}. \tag{114}$$

We thus finish the proof. □

Define $\mathbf{X} \triangleq \sum_{\substack{k=1 \\ k\neq\theta}}^{m} p_k(1-y_{k_{\max'}})$ and $\mathbf{Y} \triangleq \sum_{\substack{k=1 \\ k\neq\theta}}^{m} p_k x_{k_{\min'}}$. Based on Lemma 10, we have $\mathbf{X} \geq \mathbf{Y}$. In addition, by definitions of $\Phi$ and $\Omega$, we have

$$\begin{aligned}\mathbf{X} &\triangleq \sum_{\substack{k=1 \\ k\neq\theta}}^{m} p_k(1-y_{k_{\max'}}) \\ &= \sum_{k\in\Phi} p_k x_{k_{\min}} + \sum_{k\in\Omega}(p_k - p_\theta y_{\theta_{\min}}) \\ &= \frac{1-\beta}{\beta} p_\theta x_\theta - \frac{1}{\beta} b_\theta. \text{ (Based on (111))}\end{aligned} \tag{115}$$

Similarly to the $\Phi$ and $\Omega$ in $\mathbf{X}$, define $\Phi$' the set of $k$'s yielding $p_k x_{k_{\min'}} = p_k x_{k_{\min}} \geq p_\pi x_{\pi_{\min}} + b_k$, $k \in \Phi$', for $\mathbf{Y}$. Since based on (I) in Lemma 10, we have $p_\pi x_{\pi_{\min}} + b_k \leq p_k - p_\theta y_{\theta_{\min}}$ for all $k$ except $\theta$. Therefore, for those $k$'s belonging to $\Phi$ (in $\mathbf{X}$), based on (108), we get $p_k x_{k_{\min}} \geq p_k - p_\theta y_{\theta_{\min}} \geq p_\pi x_{\pi_{\min}} + b_k$, and based on (107), we find that *those k's belonging to* $\Phi$ *(in* $\mathbf{X}$*) also belong to* $\Phi$*' (in* $\mathbf{Y}$*), i.e.,* $\Phi \subseteq \Phi$'. Therefore, $\mathbf{Y}$ can be interpreted by $\Phi$ as follows.

$$\begin{aligned}\mathbf{Y} &\triangleq \sum_{\substack{k=1 \\ k\neq\theta}}^{m} p_k x_{k_{\min'}} \\ &= \sum_{k\in\Phi} p_k x_{k_{\min}} + \sum_{k\in\Omega} \max\{p_\pi x_{\pi_{\min}} + b_k,\ p_k x_{k_{\min}}\}.\end{aligned} \tag{116}$$

In addition, similarly to $\mathbf{Y}$, we define $\mathbf{Z}$ as follows.

$$\mathbf{Z} \triangleq \sum_{k\in\Phi} p_k x_{k_{\min}} + \sum_{k\in\Omega}(p_\pi x_{\pi_{\min}} + b_k). \tag{117}$$

Clearly, the RHS of $\mathbf{Z}$ is not greater than the RHS of $\mathbf{Y}$, and thus we have $\mathbf{X} \geq \mathbf{Y} \geq \mathbf{Z}$. Define $\mathbf{W} \triangleq \mathbf{X} - \mathbf{Z}$, based on (115) and (117), we have

$$\begin{aligned}\mathbf{W} &\triangleq \mathbf{X} - \mathbf{Z} \\ &= \sum_{k\in\Omega}[(p_k - p_\theta y_{\theta_{\min}}) - (p_\pi x_{\pi_{\min}} + b_k)] \\ &= \sum_{k\in\Omega}\left[(p_k - p_\theta y_{\theta_{\min}}) - \left(p_\pi x_{\pi_{\min}} + p_k - \beta\sum_{i=1}^{m} p_i\right)\right] \\ &= \sum_{k\in\Omega}\left[\beta\sum_{i=1}^{m} p_i - p_\theta y_{\theta_{\min}} - p_\pi x_{\pi_{\min}}\right] \\ &= \omega\left[\beta\sum_{i=1}^{m} p_i - p_\theta y_{\theta_{\min}} - p_\pi x_{\pi_{\min}}\right].\end{aligned} \tag{118}$$

Define $\mathbf{C} \triangleq \beta\sum_{i=1}^{m} p_i - p_\theta y_{\theta_{\min}} - p_\pi x_{\pi_{\min}}$, the constant term in (118). Since $\mathbf{X} \geq \mathbf{Z}$, we have $\mathbf{W} = \omega\mathbf{C} \geq 0$, and because $\omega$ is the cardinality of $\Omega$, it is non-negative, and thus $\mathbf{C} \geq 0$. Since $\mathbf{C} \geq 0$, and from Lemma 11, $\omega \leq \frac{1-\beta}{\beta}$, we thus have

$$\mathbf{Y} \geq \mathbf{Z} = \mathbf{X} - (\mathbf{X} - \mathbf{Z}) = \mathbf{X} - \mathbf{W} = \mathbf{X} - \omega\mathbf{C} \geq \mathbf{X} - \frac{1-\beta}{\beta}\mathbf{C}. \tag{119}$$

Note that since $y_\theta = y_{\theta_{\min}}$, the RHS of (119) becomes

$$\begin{aligned}
&\mathbf{X} - \frac{1-\beta}{\beta}\mathbf{C} \\
&= \left[\frac{1-\beta}{\beta}p_\theta x_\theta - \frac{1}{\beta}b_\theta\right] - \frac{1-\beta}{\beta}\mathbf{C} \\
&= \frac{1-\beta}{\beta}p_\theta x_\theta - \frac{1}{\beta}b_\theta - (1-\beta)\sum_{i=1}^{m}p_i + \frac{1-\beta}{\beta}p_\theta y_{\theta_{\min}} + \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} \\
&= \frac{1-\beta}{\beta}p_\theta(x_\theta + y_\theta) - \frac{1}{\beta}b_\theta - (1-\beta)\sum_{i=1}^{m}p_i + \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} \qquad (120)\\
&= \frac{1-\beta}{\beta}p_\theta - \frac{1-\beta}{\beta}b_\theta - (1-\beta)\sum_{i=1}^{m}p_i + \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} - b_\theta \\
&= \frac{1-\beta}{\beta}b_\theta - \frac{1-\beta}{\beta}b_\theta + \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} - b_\theta \\
&= \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} - b_\theta.
\end{aligned}$$

Therefore, we obtain

$$\sum_{\substack{k=1\\k\neq\theta}}^{m}p_k x_{k_{\min'}} = \mathbf{Y} \geq \mathbf{X} - \frac{1-\beta}{\beta}\mathbf{C} = \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} - b_\theta. \qquad (121)$$

Since $j = \theta$, by replacing $\theta$ in (121) by $j$, we obtain (99) and finish the proof of the converse.

**We summarize Case 2.2 in the following:**

$$\begin{aligned}
&R_j^* \geq 1 \text{ and } \frac{1-\beta}{\beta}p_\pi x_{\pi_{\min}} - b_j \leq \sum_{\substack{k=1\\k\neq j}}^{m}p_k x_{k_{\min'}} \\
\iff &y_\theta = y_{\theta_{\min}} \text{ and } y_k = y_{k_{\max'}}, \forall k = 1,\ldots,m, k \neq \theta \\
\iff &x_j^\dagger = \frac{1}{p_j}\left\{p_j - \frac{\beta}{1-\beta}\sum_{\substack{k=1\\k\neq j}}^{m}p_k y_{k_{\max'}}\right\} \triangleq x_j^{\dagger 0} \\
\iff &\beta \geq \frac{p_\theta y_{\theta_{\min}}}{p_\theta y_{\theta_{\min}} + \sum_{\substack{k=1\\k\neq\theta}}^{m}p_k y_{k_{\max'}}} \triangleq \beta_0.
\end{aligned}$$

Therefore, if (S1) (S4) are true, (OPT-1) has feasible solutions for arbitrary $x_{k_{\min}}$ and $x_{k_{\max}}$, $\forall k = 1,\ldots,m$, $k \neq j$, which implies (OPT-1) has feasible solutions for all the cases (Case 1, Case 2.1, and Case 2.2), which requires $\beta$ to be greater than the minimum $\beta$ in each of the above cases, i.e.,

$\beta \geq \max\{\beta_0, \beta_1, \beta_{p_j}\}$, where

$$\beta_0 = \frac{p_\theta y_{\theta\min}}{p_\theta y_{\theta\min} + \sum_{\substack{k=1 \\ k\neq\theta}}^{m} p_k y_{k\max'}},$$
$$\beta_1 = \frac{p_\pi x_{\pi\min}}{p_\pi x_{\pi\min} + \sum_{\substack{k=1 \\ k\neq\pi}}^{m} p_k x_{k\max'}},$$
$$\beta_{p_j} = \frac{p_\pi x_{\pi\min} + p_j y_{j\min}}{\sum_{k=1}^{m} p_k},$$

and the corresponding optimal objective value $x_j^\dagger$ and its corresponding optimal solutions are

$$\begin{aligned}
\beta = \beta_0 \iff x_j^\dagger &= \frac{1}{p_j}\left\{p_j - \frac{\beta}{1-\beta}\sum_{\substack{k=1 \\ k\neq j}}^{m} p_k y_{k\max'}\right\} \triangleq x_j^{\dagger 0} \\
&\iff y_j = y_\theta = y_{\theta\min} \\
&\quad\ y_k = y_{k\max'}, \forall k = 1, \ldots, m,\ k \neq \theta, \\
\beta = \beta_1 \iff x_j^\dagger &= \frac{1}{p_j}\left\{\frac{1-\beta}{\beta} p_\pi x_{\pi\min} - \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_{k\max'}\right\} \triangleq x_j^{\dagger 1} \\
&\iff x_\pi = x_{\pi\min} \\
&\quad\ x_k = x_{k\max'}, \forall k = 1, \ldots, m,\ k \neq j, \pi, \\
\beta = \beta_{p_j} \iff x_j^\dagger &= \frac{1}{p_j}\left\{p_\pi x_{\pi\min} + p_j - \beta\sum_{k=1}^{m} p_k\right\} \triangleq x_j^{\dagger p} \\
&\iff x_\pi = x_{\pi\min} \\
&\quad\ \sum_{\substack{k=1 \\ k\neq j,\pi}}^{m} p_k x_k = \frac{1-2\beta}{\beta} p_\pi x_{\pi\min} - p_j + \beta\sum_{k=1}^{m} p_k.
\end{aligned}$$

We thus finish the proof of Lemma 9.

## ACKNOWLEDGEMENTS

We would like to thank Muhammad Naveed, Bhaskar Krishnamachari, Konstantinos Psounis, Shang-Hua Teng, and the anonymous referees for their insightful and constructive comments.

## REFERENCES

[1] [n.d.]. Admissions Transparency Data, New College of the Humanities, London, United Kingdom. https://t.ly/9bo0.
[2] [n.d.]. Admissions Transparency Implementation Working Group, Department of Education, Skills, and Employment, Australian Government. https://t.ly/qaaZ. Accessed: 2021-02-19.
[3] [n.d.]. Open Government. https://www.oecd.org/gov/open-government/. Accessed: 2021-02-19.
[4] [n.d.]. U.S. Census Bureau Historical Income Tables: People. https://goo.gl/UDoF64. Accessed: 2018-10-01.
[5] Alessandro Acquisti and Ralph Gross. 2009. Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences* (2009), PNAS–0904891106.
[6] Anita L. Allen. 2016. Protecting one's own privacy in a big data economy. *Harv. L. Rev. F.* 130 (2016), 71.
[7] Mário S. Alvim, Miguel E. Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. 2011. Differential privacy: On the trade-off between utility and information leakage. *Formal Aspects in Security and Trust* 7140 (2011), 39–54.

[8] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 973–989.

[9] Kurt M. Anstreicher. 1999. Linear programming in $O(\frac{n^3}{\ln n}L)$ operations. *SIAM Journal on Optimization* 9, 4 (1999), 803–812.

[10] Daniel W. Apley. 2016. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468* (2016).

[11] Michael Barbaro, Tom Zeller, and Saul Hansell. 2006. A face is exposed for AOL searcher no. 4417749. *New York Times* 9, 2008 (2006), 8.

[12] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[13] Gilles Barthe and Boris Kopf. 2011. Information-theoretic bounds for differentially private mechanisms. In *Computer Security Foundations Symposium (CSF), 2011 IEEE 24th*. IEEE, 191–204.

[14] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207* (2017).

[15] Dan Biddle. 2006. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing* (2 ed.). Gower Publishing, Ltd.

[16] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

[17] Stephen Boyd, Lin Xiao, Almir Mutapcic, and Jacob Mattingley. 2007. Notes on decomposition methods. *Notes for EE364B, Stanford University* (2007), 1–36.

[18] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[19] Jianneng Cao and Panagiotis Karras. 2012. Publishing microdata with a robust privacy guarantee. *Proceedings of the VLDB Endowment* 5, 11 (2012), 1388–1399.

[20] Fred H. Cate, D. Annette Fields, and James K. McBain. 1994. The right to privacy and the public's right to know: The central purpose of the Freedom of Information Act. *Admin. L. Rev.* 46 (1994), 41.

[21] Chien-Lun Chen, Ranjan Pal, and Leana Golubchik. 2016. Oblivious mechanisms in differential privacy: Experiments, conjectures, and open questions. In *2016 IEEE Security and Privacy Workshops (SPW)*. IEEE, 41–48.

[22] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230* (2017).

[23] A. Datta, S. Sen, and Y. Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*. 598–617.

[24] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[25] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

[26] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014).

[27] Kelly Dilworth. [n.d.]. We still don't know a lot about how credit card applications are evaluated. https://t.ly/ifSo.

[28] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics*. IEEE, 0210–0215.

[29] Flávio du Pin Calmon and Nadia Fawaz. 2012. Privacy against statistical inference. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 1401–1408.

[30] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS'12)*. ACM, New York, NY, USA, 214–226.

[31] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*. Springer, 265–284.

[32] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[33] Michael D. Ekstrand, Rezvan Joshaghani, and Hoda Mehrpouyan. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*. 35–47.

[34] Barbara Espinoza and Geoffrey Smith. 2013. Min-entropy as a resource. *Information and Computation* 226 (2013), 57–75.

[35] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

[36] Katherine Fink. 2018. Opening the government's black boxes: Freedom of information and algorithmic accountability. *Information, Communication & Society* 21, 10 (2018), 1453–1471.

[37] Kate Finman. [n.d.]. CA state auditor report alleges UC admissions are biased, unfair. https://t.ly/7VcF. Accessed: 2021-02-19.

[38] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. *arXiv preprint arXiv:1801.01489* (2018).
[39] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1322–1333.
[40] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*. 17–32.
[41] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* (2001), 1189–1232.
[42] Jerome H. Friedman, Bogdan E. Popescu, et al. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2, 3 (2008), 916–954.
[43] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65.
[44] Bryce Goodman and Seth Flaxman. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813* (2016).
[45] Brandon M. Greenwell, Bradley C. Boehmke, and Andrew J. McCarthy. 2018. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755* (2018).
[46] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 93.
[47] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.
[48] Harold V. Henderson and Shayle R. Searle. 1981. On deriving the inverse of a sum of matrices. *Siam Review* 23, 1 (1981), 53–60.
[49] Tamara E. Holmes. [n.d.]. How race affects your credit score. https://t.ly/6gfV. Accessed: 2021-02-19.
[50] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).
[51] Giles Hooker. 2004. Discovering additive structure in black box functions. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 575–580.
[52] Mikella Hurley and Julius Adebayo. 2016. Credit scoring in the era of big data. *Yale JL & Tech.* 18 (2016), 148.
[53] Farhad Kamali and Hilary Wynne. 2010. Pharmacogenetics of warfarin. *Annual Review of Medicine* 61 (2010), 63–75.
[54] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* 35, 3 (2013), 613–644.
[55] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases* (2012), 35–50.
[56] Igor Kononenko et al. 2010. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research* 11, Jan (2010), 1–18.
[57] Sanjay Krishnan and Eugene Wu. 2017. PALM: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM, 4.
[58] Martin Kučera, Petar Tsankov, Timon Gehr, Marco Guarnieri, and Martin Vechev. 2017. Synthesis of probabilistic privacy enforcement. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 391–408.
[59] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.
[60] Floridi Luciano and Taddeo Mariarosaria. 2016. What is data ethics? *Phil. Trans. R. Soc. A.37420160360* (2016).
[61] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. *l*-diversity: Privacy beyond *k*-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 24–24.
[62] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 3.
[63] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
[64] H. Mittelmann. [n.d.]. Benchmark of commercial LP solvers. http://plato.asu.edu/ftp/lpcom.html. Accessed: 2021-03-6.
[65] Tomas Monarrez and Kelia Washington. 2020. Racial and Ethnic Representation in Postsecondary Education. Research report. *Urban Institute* (2020).
[66] S. Alvim M'rio, Kostas Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. 2012. Measuring information leakage using generalized gain functions. In *2012 IEEE 25th Computer Security Foundations Symposium*. IEEE, 265–279.

[67] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE, 111–125.
[68] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.
[69] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814* (2016).
[70] Neel Patel, Reza Shokri, and Yair Zick. 2020. Model explanations with differential privacy. *arXiv:2006.09129* (2020).
[71] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
[72] Joel R. Reidenberg and Florian Schaub. 2018. Achieving big data privacy in education. *Theory and Research in Education* 16, 3 (2018), 263–279.
[73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.
[74] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.
[75] Pierangela Samarati and Latanya Sweeney. 1998. Generalizing data to provide anonymity when disclosing information. In *PODS*, Vol. 98. CiteSeer, 188.
[76] Pierangela Samarati and Latanya Sweeney. 1998. *Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression*. Technical Report. Technical report, SRI International.
[77] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Vol. 11700. Springer Nature.
[78] Nisha Shekhawat, Aakanksha Chauhan, and Sakthi Balan Muthiah. 2019. Algorithmic privacy and gender bias issues in Google ad settings. In *Proceedings of the 10th ACM Conference on Web Science*. 281–285.
[79] Reza Shokri, Martin Strobel, and Yair Zick. 2020. On the privacy risks of model explanations. *arXiv preprint arXiv:1907.00164v5* (2020).
[80] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *arXiv preprint arXiv:1610.05820* (2016).
[81] Robert H. Sloan and Richard Warner. 2018. When is an algorithm transparent? Predictive analytics, privacy, and public policy. *IEEE Security & Privacy* 16, 3 (2018), 18–25.
[82] Bernd Carsten Stahl and David Wright. 2018. Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy* 16, 3 (2018), 26–33.
[83] Latanya Sweeney. 1997. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* 25, 2–3 (1997), 98–110.
[84] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health* 671 (2000), 1–34.
[85] R. C. Thompson. 1978. Matrix type metric inequalities. *Linear and Multilinear Algebra* 5, 4 (1978), 303–319.
[86] Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1521–1528.
[87] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *USENIX Security Symposium*. 601–618.
[88] Ke Wang, Benjamin C. M. Fung, and Guozhu Dong. 2005. Integrating private databases for data analysis. In *International Conference on Intelligence and Security Informatics*. Springer, 171–182.
[89] Ke Wang, Benjamin C. M. Fung, and S. Yu Philip. 2007. Handicapping attacker's confidence: An alternative to $k$-anonymization. *Knowledge and Information Systems* 11, 3 (2007), 345–368.
[90] Teresa Watanabe. [n.d.]. UCLA professor wants to see data on whether UC illegally uses race in admissions decisions. https://t.ly/ny6t.
[91] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 191–200.
[92] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 325–333.
[93] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. 2009. The feature importance ranking measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 694–709.
[94] Stanley Zionts. 1968. Programming with linear fractional functionals. *Naval Research Logistics Quarterly* 15, 3 (1968), 449–451.

Received November 2020; revised March 2021; accepted April 2021