Worst-Case Privacy*

Tianhao Liu†

November 13, 2025

(*Click Here for the latest version.*)

Abstract

How should privacy loss be assessed? Motivated by current privacy discourse, I formalize the principle of uniform protection: if an information structure violates a privacy standard, it should also be deemed unacceptable even when used infrequently. Together with a standard (Blackwell) monotonicity requirement, uniform protection defines the class of *Worst-Case Privacy* measures. I show that any such measure can be decomposed into the losses of individual signals and aggregated through their maximum. When the privacy threats are predictions of protected attributes, the loss of a signal depends on how much it distinguishes between each pair of attributes, measured by log-likelihood ratios. I apply these measures to canonical economic settings. In matching markets, a sharp tradeoff arises: as stable mechanisms shift from firm-optimal to worker-optimal, workers' welfare improves but their privacy deteriorates. In a voting application, the optimal privacy-constrained rule exhibits choice reversal when a candidate's vote count is extreme.

Keywords: Prediction Privacy, Uniform Protection, Matching, Voting

*I am indebted to my advisor, Navin Kartik, as well as Yeon-Koo Che, Elliot Lipnowski, and Jacopo Perego for their consistent support on this project and beyond. I am grateful for the valuable feedback from Mark Dean, Laura Doval, Kevin He, Qingmin Liu, Kai Hao Yang, and Yangfan Zhou. I thank César Barilla, Benjamin Brooks, Rachel Cummings, Rahul Deb, Jinshuo Dong, Francesco Fabbri, Yingni Guo, Jan Knoepfle, Zihao Li, Erik Madsen, Xiaosheng Mu, Juan Ortner, Aniko Öry, Alessandro Pavan, Daniel Rappoport, Ludvig Sinander, Alexander Teytelboym, Yu Fu Wong, Yuzhao Yang, and Mengchu Zheng for helpful comments. I also received useful feedback from audiences at Columbia, Oxford, Jason Hartline's lab at Northwestern, and the 2025 IO Theory Conference.

†Department of Economics, Columbia University. E-mail: tl3014@columbia.edu.

1. Introduction

Technological advances have greatly expanded the ability to collect and process personal data. Online retailers, for example, can track consumers' browsing and purchase histories to infer their preferences. On the one hand, such information allows retailers to personalize recommendations and improve matching efficiency; on the other, it raises privacy concerns such as the potential for price discrimination.[1] Even when retailers use consumer data responsibly, there remains a risk that the data may be leaked and misused by third parties—for instance, through unauthorized resale or fraud. These forces create a fundamental tradeoff: more information can enhance consumer utility and firm profit, but simultaneously increases privacy loss by expanding the scope for misuse. To analyze these tradeoffs, the first question we must ask is: how should privacy loss be assessed? In this paper, I introduce economically grounded measures of privacy loss and develop a framework to study the tradeoffs between the value of information and privacy loss, with applications to canonical economic settings.

The backbone of my approach is the principle of *uniform protection*: if an information structure violates a privacy standard, it should also be deemed unacceptable even when used infrequently or on a small fraction of individuals. For example, if it is unacceptable for an online retailer to collect sensitive consumer characteristics, then collecting such information for even one percent of consumers is unacceptable as well. This principle of uniform protection is motivated by privacy regulation that views privacy as a fundamental individual right. For instance, U.S. Census confidentiality laws (Title 13, U.S.C. §9) prohibit "any publication whereby the data furnished by any particular individual can be identified."[2]

Uniform protection has important implications for comparing the privacy losses of different data-collection methods. To illustrate, consider a retailer who wishes to learn the individual demand (Low or High) of a unit mass of consumers. One method is to collect information to reveal every consumer's demand, but such full revelation may entail excessive privacy loss. An alternative is to collect such information from only a random sample: for each consumer, the retailer collects his demand information with probability 0.1, so that in aggregate the retailer learns the demand of 10% of the consumers. These data-collection

[1] Section 3.2 of Acquisti, Taylor, and Wagman (2016) surveys empirical evidence on how consumer data enables personalized pricing and surplus extraction. Chapter 3 of Goldfarb and Tucker (2024) discusses privacy concerns on digital platforms.

[2] Relatedly, the Universal Declaration of Human Rights states that "No one shall be subjected to arbitrary interference with his privacy" (United Nations, Article 12, 1948); the General Data Protection Regulation (GDPR) requires that "personal data shall be processed lawfully, fairly, and in a transparent manner in relation to the data subject" (that is, with respect to each natural person).

Table 1: Two alternative information structures for data collection. Structure A fully reveals each consumer's demand; structure B reveals each consumer's demand with probability 0.1 and otherwise nothing. Entries are conditional probabilities of each signal given the state.

(a) A: Fully revealing

State \ Signal	low	high
Low	1	0
High	0	1

(b) B: Fully revealing w.p. 0.1

State \ Signal	low	high	neutral
Low	0.1	0	0.9
High	0	0.1	0.9

methods can be modeled as applying to each consumer an *information structure*, i.e., a mapping from a consumer's demand to a signal distribution, as shown in Table 1. Information structure A perfectly reveals a consumer's demand by sending a signal "low" or "high" that matches the true demand; structure B does so with probability 0.1, and otherwise sends an uninformative "neutral" signal. Should B be considered more private than A? Uniform protection would say no. If B were assigned a strictly lower privacy loss than A, then there would exist some number δ, interpreted as a legal threshold on acceptable privacy loss, such that A's loss exceeds δ while B's falls below it. This implies that applying an unacceptable data-collection method to a small fraction of consumers makes it acceptable overall, thereby violating uniform protection.

A *privacy (loss) measure* is a function that assigns a numerical value to each information structure. I formalize uniform protection as an axiom of privacy measures, which I call *Worst-Case Protection*. The axiom requires that if an information structure can be implemented by "mixing" over two other information structures, then the privacy loss of the overall structure is determined by the higher loss of the two.[3] In Table 1, for example, B can be implemented by applying A with probability 0.1 and otherwise applying an uninformative information structure that always outputs signal "neutral." Worst-Case Protection requires the privacy loss of B to equal the higher of the losses of A and the uninformative information structure, ensuring uniform protection.

Together with a standard axiom of Blackwell Monotonicity, which requires the privacy measure to be monotonic with respect to informativeness, Worst-Case Protection defines a class of privacy measures, which I call *Worst-Case Privacy*. Theorem 1 shows that any worst-case privacy measure must have a "max-separable" representation: the privacy loss of an

[3] To be precise, the idea of uniform protection is captured by one direction of Worst-Case Protection: an information structure's privacy loss is not lower than any of its components'. Worst-Case Protection also imposes that the privacy loss is not higher; this can be viewed as without loss of generality, as explained in Subsection 2.1.

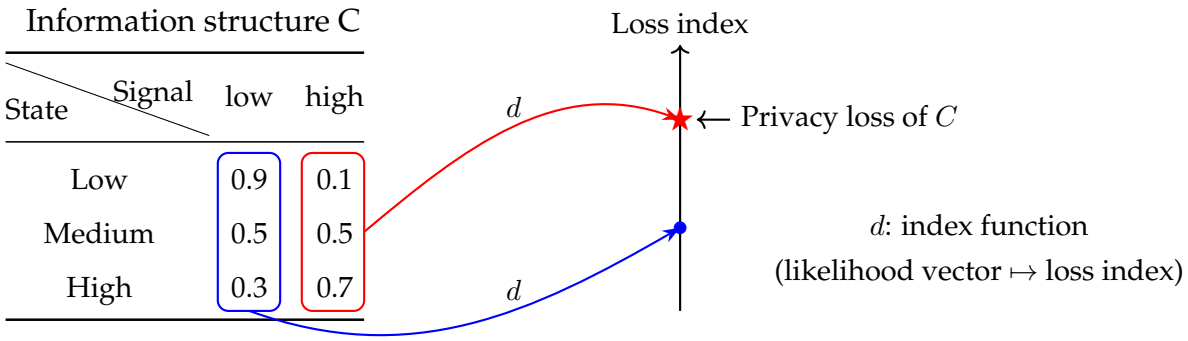

Figure 1: Illustration of a worst-case privacy measure. Each column of information structure C (highlighted in boxes) represents a likelihood vector. The index function d maps each vector to a loss index. The overall privacy loss equals the largest index (red star).

information structure can be evaluated signal by signal, and then aggregated by taking the maximum. Figure 1 illustrates this representation for a specific information structure, C. Each column of C represents a likelihood vector—the conditional probability of that signal under each state. Each likelihood vector is assigned a loss index using some *index function* d, which maps likelihood vectors to numerical indices.[4] The privacy loss of C is then given by the maximum of these indices, which is denoted by the red star. In short, the privacy measure is represented by d through (i) assigning a loss index to each likelihood vector using d, and (ii) taking the maximum across these indices.

The representation in Theorem 1 is not only conceptually useful, but also has practical implications. First, it renders the analysis of performance–privacy tradeoffs especially tractable. The max-separable structure implies that, given a privacy budget δ, it suffices to check whether the loss index of any signal falls below δ, without the need to allocate the budget optimally across signals. Leveraging this property, I derive tools for solving worst-case privacy-constrained problems in Subsection 4.1. Second, from a policy standpoint, the separability across signals also makes privacy regulation easier to monitor: the regulator only needs to check whether the likelihood vector associated with each signal is acceptable according to d. For example, if a retailer can infer a single consumer's private attribute, the privacy violation is conclusive.[5]

One important question remains: which particular worst-case privacy measure should be used in practice, or equivalently, how should the index function d be chosen? The answer depends on the adversaries—those who may misuse the information—that the mea-

[4] Moreover, the index function d has to be quasi-convex to ensure the privacy measure satisfies Blackwell Monotonicity.

[5] By contrast, under alternatives such as average-case privacy measures (e.g., mutual information), a single violation is insufficient; one must also verify that such violations occur frequently enough.

sure is designed to guard against. For example, whether an information structure that reveals more about gender has a higher loss than one that reveals more about race depends on whether adversaries discriminate by gender or by race. I model such adversaries as a class of *decision problems* $\mathcal{C}$, each representing a way in which an adversary could use information against the data owner's interest—for example, to discriminate, to infer identity, or to target a scam. When $\mathcal{C}$ denotes the set of adversarial decision problems that a regulator aims to protect against, a natural property of privacy measures is $\mathcal{C}$*-Monotonicity*: one information structure has a higher privacy loss than another whenever it enables all adversaries in $\mathcal{C}$ to achieve a higher expected gain. Different classes $\mathcal{C}$ will identify different worst-case privacy measures, or equivalently, different index functions d. When $\mathcal{C}$ contains all decision problems, we know from Blackwell (1953) that $\mathcal{C}$-Monotonicity amounts to Blackwell Monotonicity.

To refine the set of index functions, I focus on a class of commonly encountered adversarial problems: *prediction problems*. In a prediction problem, an adversary gains a positive payoff if he correctly predicts the true state, and zero otherwise. Instances of prediction problems include a hacker predicting the owner of anonymized data or a retailer predicting a consumer's demographic attributes. Theorem 2 shows that when $\mathcal{C}$ is restricted to prediction problems, $\mathcal{C}$-Monotonicity implies a pairwise structure of the index function d: the loss index of a likelihood vector is determined by the maximum of (possibly weighted) log-likelihood ratios between pairs of states. For example, for information structure C in Figure 1, modulo the weights, the red likelihood vector has a loss index of $\log 7$, corresponding to the largest among $\log(7/5)$, $\log 5$, and $\log 7$. This pairwise structure stems from the fact that comparison between predicting any two states depends only on the belief about that pair, since both predictions yield zero payoff when the true state is a third one.

I refer to this subclass of Worst-Case Privacy that protects against prediction problems as *Prediction Privacy*. For prediction privacy measures, the only parameters left unspecified are the weights on the log-likelihood ratios. These weights are chosen by the designer and capture normative judgments about distinguishing which pairs of states are more harmful. For example, distinguishing "HIV" from "healthy" may be viewed as more harmful than distinguishing "fever" from "healthy." By contrast, in contexts where the states represent demographic groups such as gender or race, one may impose identical weights across pairs to reflect equal treatment.

Prediction Privacy, and more generally Worst-Case Privacy, can be deployed to both evaluate tradeoffs between performance and privacy, and to study design problems under privacy constraints.

My first application studies how privacy and welfare trade off in matching markets. This analysis is motivated by the concern that matching outcomes—such as student placements or medical residencies—may inadvertently reveal participants' private preferences, which can themselves be sensitive or correlated with other sensitive information (such as location or special needs). I focus on worker privacy in a two-sided matching market between workers and firms. The analysis reveals a sharp tradeoff between privacy and welfare: the firm-optimal stable mechanism, while minimizing worker welfare, maximizes worker privacy. Moreover, within a natural class of stable mechanisms, any increase in worker welfare necessarily entails a (weakly) higher privacy loss for workers. The underlying mechanism is intuitive: improvements in worker welfare make the realized matching outcomes more correlated with workers' preferences. This stronger correlation raises the privacy loss.

Next, I turn to economic design problems and ask how to maximize welfare or profit under privacy constraints. I first formulate a general framework with worst-case privacy constraints and develop tractable tools using a duality-based approach to characterize the optimal solution, and then apply these tools to two economic settings.

The first setting concerns a voting environment in which a social planner aggregates voters' private signals about an underlying state while protecting voter privacy. Such protections are important because failure to do so can facilitate vote buying or retaliation, thereby undermining collective decision-making. I show that the optimal privacy-constrained voting rule exhibits extreme reversal: when the vote count for one outcome becomes sufficiently extreme, the opposite outcome is chosen instead.[6] Under this rule, an outcome may be selected either because it has a moderate majority or because the opposite outcome has an extreme majority. Consequently, the chosen outcome conveys little information about any individual vote.

The second setting examines credit screening, where a credit agency recommends to a lender whether to grant loans to borrowers while protecting borrowers' demographic attributes. Such protection is motivated by the Equal Credit Opportunity Act, which prohibits lenders from discriminating based on demographic attributes. The optimal privacy-constrained recommendation rule exhibits a form of "affirmative action": the credit agency sets a targeted range of rejection rates for the demographic groups. Whenever a group's initial rejection rate falls outside this range, the credit agency recommends more or fewer within that group until the rejection rate lies in the range. This rule ensures that a rec-

[6] This rule aims to illustrate the forces that prediction privacy brings to voting rather than suggesting it is practical. Subsection 4.2 discusses how to interpret this result, and Subsection 5.1 discusses alternative privacy measures that do not produce extreme reversals.

ommendation—whether to accept or reject—reveals limited information about borrowers' demographic attributes.

Related Literature

Theoretical treatment of privacy has been studied through several distinct approaches. One strand of the literature microfounds agents' preferences for privacy in how personal information may be used against them. For example, Taylor (2004) and Calzolari and Pavan (2006) capture concerns about potential exploitation in future interactions, while Ichihashi (2020), Hidir and Vellodi (2021), and Galperti and Perego (2023) model consumers' decisions to share data based on how data might be used. This approach provides concrete interpretations of privacy but often relies heavily on the details of the environment, limiting its portability.

An alternative strand takes a reduced-form approach, modeling privacy as an exogenous cost. Such costs are sometimes specified as a binary penalty for revealing one's data record (Choi, Jeon, and Kim, 2019; Galperti, Liu, and Perego, 2024). Eilat, Eliaz, and Mu (2021) propose more fine-grained measures of Bayesian privacy—an ex ante version based on mutual information and an ex post version, which is an instance of Worst-Case Privacy. They apply Bayesian privacy to study monopolistic screening and auctions (Eilat, Eliaz, and Mu, 2023). Although this approach is tractable and portable, it offers only limited interpretability of the underlying costs.

My approach combines the advantages of both: it provides reduced-form privacy measures that are grounded in transparent principles. These measures are derived by asking whose privacy to protect (the worst case) and from whom to protect it (the relevant class of adversaries). Other approaches to privacy include Strack and Yang (2024), He, Sandomirskiy, and Tamuz (2021), and Strack and Yang (2025), where privacy of certain information needs to be protected strictly, and Haupt and Hitzig (2021), who introduce a notion of contextual privacy, capturing knowledge that is superfluous given the context. By quantifying privacy loss, my approach complements theirs by making it possible to trace out the frontier between performance and privacy.

Privacy is also a central topic in computer science. A leading notion is differential privacy (Dwork, McSherry, Nissim, and Smith, 2006; Dwork and Roth, 2014), which captures the idea that the inclusion of any single individual's data should have only a limited effect on the released output. For economic connections, Pai and Roth (2013) relate differential privacy to approximate strategy-proofness in mechanism design, while Schmutte and Yoder (2022) study optimal publication mechanisms subject to a differential privacy constraint. Differential privacy is an instance of Worst-Case Privacy, and I contribute to this

literature by providing it a decision-theoretic interpretation through prediction problems comparable to Prediction Privacy in Subsection 5.1.

Two other privacy notions in computer science are closely related to my framework. The first is the maximum adversarial threat model in the quantitative information flow (QIF) literature (Alvim, Chatzikokolakis, McIver, Morgan, Palamidessi, and Smith, 2016; Zarrabian and Sadeghi, 2025). This model shares the same max-separable form as Worst-Case Privacy but differs conceptually. In QIF, both the max-separable structure and the function d are *assumed*, representing the maximum gain of a specified adversary. By contrast, in my framework, they are *derived* from the economic principles of uniform protection and consistency with comparisons across a class of adversarial decision problems. The second related notion is the Pufferfish framework of Kifer and Machanavajjhala (2014). It shares a similar functional form with Prediction Privacy but differs in two key respects. First, Prediction Privacy allows asymmetric treatment of pairs of protected attributes, whereas Pufferfish corresponds to the special case of equal weighting across all pairs. Second, Prediction Privacy is grounded in decision-theoretic principles, while Pufferfish does not provide such an interpretation.

Some work in computer science also adopts an axiomatic approach to privacy. I note two related frameworks. Gilboa-Freedman and Smorodinsky (2020) axiomatize differential privacy in a binary-state setting. An insight from my Theorem 1 is that the natural extensions of their axioms beyond the binary case have much broader implications, and are insufficient to pin down differential privacy.[7] Subsection 5.1 complements this discussion by providing a decision-theoretic interpretation for differential privacy. Su (2024) shows that any measure defined through pairwise state comparisons can be characterized via hypothesis testing. Prediction problems provide an economic context in which such pairwise comparisons are sufficient for evaluating privacy loss.

Finally, this paper broadly connects to the literature on the cost of information acquisition, as both are related to the informativeness of information structures. A canonical assumption in that literature is convexity: if an information structure randomizes between two other information structures, the cost of the overall structure is no greater than their convex combination (Caplin and Dean, 2015; De Oliveira, Denti, Mihm, and Ozbek, 2017). Similar average-case reasoning appears in axiomatic models of information costs, such as Pomatto, Strack, and Tamuz (2023) and Bordoli and Iijima (2025). This property is fundamentally at odds with uniform protection, since it implies that any information structure becomes acceptable when applied with sufficiently small probability. Accordingly, I depart

[7] In fact, any worst-case privacy measure satisfies their main axioms.

from this literature by replacing convexity with the axiom of Worst-Case Protection.

Organization of the Paper. The remainder of the paper is structured as follows. Section 2 characterizes the class of privacy measures consistent with uniform protection, and studies its specialization to prediction problems. Section 3 analyzes the privacy and welfare tradeoff in a canonical matching market. Section 4 develops optimization tools for worst-case privacy constraints, and applies them to two privacy-constrained design problems, voting and credit screening. Section 5 discusses the technical extensions of the framework. Section 6 concludes.

2. Worst-Case Privacy

Let $\Theta = \{\theta_1, \ldots, \theta_K\}$ denote a finite set of states whose information we seek to protect. Let $(S, \mathcal{S})$ be an uncountable measurable signal space representing what an observer can see.[8] An information structure is a mapping $x : \Theta \to \Delta(S)$, where $\Delta(S)$ denotes the set of Borel probability measures over S. The information structure describes what information is conveyed by the signals. For most of the paper, I focus for simplicity on information structures with finite support: for each $\theta \in \Theta$, the signal distribution $x(\cdot|\theta)$ is supported on finitely many signals.[9] Let

$$S_x := \{s \in S : s \in \operatorname{Supp} x(\cdot|\theta) \text{ for some } \theta\}$$

denote the finite set of signals in the support of x, and let $\mathcal{X}_f \subset \mathcal{X}$ denote the set of finitely supported structures.

Privacy loss arises when the observer may misuse the information conveyed by the signals. For example, when the states encode an individual's characteristics, the observer may exploit this information to discriminate or to target a scam. I take a reduced-form approach and capture such privacy losses by a *privacy (loss) measure*, defined as a function L assigning a nonnegative numerical loss to each information structure:[10]

$$L : \mathcal{X}_f \to \overline{\mathbb{R}}_+.$$

[8] More precisely, $(S, \mathcal{S})$ is assumed to be a standard Borel space. It is also assumed to be uncountable to ensure that any other standard Borel space can be embedded into it by Borel isomorphism.

[9] Subsection 5.2 extends the main results to privacy measures defined on the full domain of information structures $\mathcal{X}$.

[10] Since all my axioms are ordinal, they put no restriction on the magnitude or units of L. In fact, in the general framework where the domain of L is $\mathcal{X}$, the privacy measure can be equivalently defined by a total order on $\mathcal{X}$. See Remark 4 for details.

Modeling privacy loss in reduced form has two advantages. First, it is portable, as it applies broadly to environments where information can be represented by an information structure. Second, it enables the formulation of precise privacy standards by requiring any information structure x to satisfy $L(x) \leq \delta$. From this perspective, δ represents a legal or regulatory threshold of acceptable privacy loss, and $\{x \in \mathcal{X}_f : L(x) \leq \delta\}$ characterizes the set of legally permissible information structures. Next, I introduce axioms that guide the choice of L.

2.1. The Axioms of Worst-Case Privacy

The first axiom captures the principle of uniform protection: if an information structure violates a privacy standard, it should also be deemed unacceptable even when used infrequently or on a small fraction of individuals. This principle is motivated by the privacy regulation that views privacy as a fundamental individual right (See fn. 2). I formalize this principle via the following axiom of *Worst-Case Protection*.

Axiom 1 (Worst-Case Protection)**.** *For any $x', x'' \in \mathcal{X}_f$ with $S_{x'} \cap S_{x''} = \varnothing$ and any $\alpha \in (0,1)$, the privacy loss of their convex combination, $x = \alpha x' + (1-\alpha)x''$, satisfies*

$$L(x) \;=\; \max\{L(x'),\, L(x'')\}.$$

In words, Worst-Case Protection requires that if x can be implemented by randomly applying two other information structures x' and x'', then its privacy loss equals the higher loss of the two. The condition $S_{x'} \cap S_{x''} = \varnothing$ means that any realized signal s reveals whether it was generated by x' or x'', and thus the signals of x' and x'' do not interfere. The relationship $x = \alpha x' + (1-\alpha)x''$ means that the signal distribution of x conditional on θ equals $\alpha x'(\cdot|\theta) + (1-\alpha)x''(\cdot|\theta)$. Table 2 illustrates using the information structures in Table 1. Let x' denote information structure A and x'' the uninformative structure that always sends the "neutral" signal. Then information structure B equals $x = 0.1x' + 0.9x''$. Since the supports of x' ("low", "high") and x'' ("neutral") are disjoint, Worst-Case Protection requires the privacy loss of x to equal the higher of x' and x''.

The substantive requirement of Worst-Case Protection is $L(x) \geq \max\{L(x'), L(x'')\}$, which reflects the principle of uniform protection. To see how, suppose Θ is the type space of an individual, and x is applied independently to a unit mass of individuals, each with a type $\theta \in \Theta$. In this case, α fraction of the individuals will be applied x', and the other $1-\alpha$ fraction will be applied x''. Suppose the legally acceptable privacy loss is δ. Uniform protection requires that if either x' or x'' is unacceptable, meaning that $\max\{L(x'), L(x'')\} > \delta$, then x should also be unacceptable, meaning $L(x) > \delta$. We would

Table 2: Illustration of Worst-Case Protection with Information Structures in Table 1

(a) x': Information structure A

	low	high	neutral
Low	1	0	0
High	0	1	0

(b) x'': Uninformative

	low	high	neutral
Low	0	0	1
High	0	0	1

(c) x: Information structure B

	low	high	neutral
Low	0.1	0	0.9
High	0	0.1	0.9

like this condition to hold for all privacy standards, that is, for all δ. This is equivalent to requiring $L(x) \geq \max\{L(x'), L(x'')\}$.

The converse requirement $L(x) \leq \max\{L(x'), L(x'')\}$ is without loss in the following sense. Note that an equivalent way to implement x is to first toss a coin and, depending on its outcome (with probabilities α and $1-\alpha$), use x' or x''. Hence, whenever both x' and x'' satisfy a privacy standard (i.e., $\max\{L(x'), L(x'')\} \leq \delta$), the overall structure x can be implemented in this way, so it is without loss to assume $L(x) \leq \delta$. This condition holds for all thresholds δ if and only if $L(x) \leq \max\{L(x'), L(x'')\}$.

It is instructive to contrast Worst-Case Protection with convexity, a standard assumption for costs of information acquisition (Caplin and Dean, 2015; De Oliveira et al., 2017). Specifically, a cost function C is convex if for any $x', x'' \in \mathcal{X}$ with $S_{x'} \cap S_{x''} = \varnothing$ and $\alpha \in (0,1)$, it holds that $C(x) \leq \alpha C(x') + (1-\alpha)C(x'')$. This assumption is natural when C describes the cost of information acquisition. However, for a privacy measure, convexity implies that the low privacy loss to a large group can be "averaged out" with the high loss to a small group, which violates uniform protection. Worst-Case Protection instead ensures that mixing an information structure with a more private one does not dilute its privacy loss.

Finally, I note that the interpretation of Worst-Case Protection depends on what Θ represents. When Θ represents the type space of one individual, it captures uniform protection across individuals, as discussed earlier. However, when Θ represents datasets containing types of multiple individuals, the worst case is taken over datasets rather than over individuals within a dataset. For example, if $\Theta = T_1 \times T_2$ denotes the type profiles of two individuals, the privacy measure L can assign a privacy loss of 1 to revealing the first individual's type and 2 to revealing the second. Therefore, individuals *within* a dataset are not protected uniformly. Nonetheless, Worst-Case Protection imposes restrictions *across* datasets: if revealing the second individual's type is unacceptable, then it should not be revealed in any dataset. In short, Worst-Case Protection ensures that an unacceptable information structure is never used, but it does not specify which information structures are unacceptable.

The second axiom provides a standard principle to specify which information structures have higher privacy losses. As noted earlier, privacy loss arises because information may be misused. To capture this idea, I model potential misuse through *adversaries* whose gain comes at society's loss. For example, when an adversary targets a scam, the adversary's gain corresponds to the victim's loss. Each adversary faces a *decision problem* represented by a triple (μ_0, A, u), where $\mu_0 \in \Delta(\Theta)$ is a prior, A is a set of actions, and $u : \Theta \times A \to \mathbb{R}$ is a measurable utility function. Information enables adversaries to make better decisions and thereby imposes higher social loss.

A benchmark is to take an ignorant view of adversaries by assuming that any decision problem is plausible. In this case, a natural criterion is to deem an information structure x to have a higher privacy loss than x' when all adversaries' expected utility under x is at least as high as under x', or equivalently, when x is more (Blackwell) informative than x'.[11]

Axiom 2 (Blackwell Monotonicity)**.** *If x is more informative than x', then $L(x) \geq L(x')$.*

By Blackwell's (1953) theorem, the axiom is equivalent to requiring that post-processing, such as adding noise to signals, can never increase privacy loss. As an illustration, consider Table 2. Both information structures x'' and x can be derived by adding noise to x', so Blackwell Monotonicity requires $L(x''), L(x) \leq L(x')$. Therefore, Blackwell Monotonicity and Worst-Case Protection jointly imply that $L(x) = L(x')$.

An immediate implication of Blackwell Monotonicity is that privacy loss depends only on the informational content of an information structure, but not on its formats. For example, relabeling or duplicating signals does not affect L. In particular, if an information structure is uninformative, such as x'' in Table 2, then its privacy loss is no greater than any other information structure. As a *normalization*, I set the privacy loss of uninformative information structures to zero.

Combining the two axioms (and normalization), I define:

Definition 1 (Worst-Case Privacy)**.** Privacy measure L is a *worst-case privacy* measure if it satisfies Worst-Case Protection, Blackwell Monotonicity, and normalization.

2.2. Characterization of Worst-Case Privacy

Checking whether a privacy measure satisfies Worst-Case Protection requires decomposing an information structure into all possible randomizations over other information

[11] That is, $\int_{\Delta(\Theta)} \sup_{a \in A} \left(\sum_\theta u(\theta, a)\mu(\theta) \right) d\tau_x^{\mu_0}(\mu) \geq \int_{\Delta(\Theta)} \sup_{a \in A} \left(\sum_\theta u(\theta, a)\mu(\theta) \right) d\tau_{x'}^{\mu_0}(\mu)$, where $\tau_x^{\mu_0}$ denotes the posterior distribution induced by information structure x under prior μ_0.

structures and comparing the privacy losses of them. This process becomes combinatorially burdensome as the number of signals gets large. Theorem 1 simplifies by characterizing the entire class of Worst-Case Privacy. For any $x \in \mathcal{X}_f$ and $s \in S_x$, let $x^s := (x(s|\theta_1), ..., x(s|\theta_K))$ denote the likelihood vector associated with signal s.

Theorem 1. *Privacy measure L is a worst-case privacy measure if and only if*

$$L(x) = \max_{s \in S_x} d(x^s), \ \forall x \in \mathcal{X}_f \tag{WCP}$$

for some function $d : \mathbb{R}_+^K \to \overline{\mathbb{R}}_+$ that is:

1. *homogeneous of degree 0;*
2. *$d(c) = 0$ for all constant vectors $c \in \mathbb{R}_+^K$;*
3. *quasi-convex.*

Theorem 1 shows that any worst-case privacy measure can be parameterized by an *index function* d defined on the set of likelihood vectors, and the privacy loss of x is aggregated in a max-separable form across signals in its support. See Figure 1 for a graphical illustration. The index function d has three properties. First, d is homogeneous of degree 0, which means that scaling a likelihood vector by a positive constant does not change its loss index. Therefore, only the update direction of a signal matters for its loss index, not the probability of the signal. Second, d vanishes on constant vectors, which guarantees the normalization condition. Third, d is quasi-convex, which implies that if a likelihood vector is a convex combination of other likelihood vectors, then its loss index is no higher than the maximum of them. This property ensures Blackwell Monotonicity.

The representation in Theorem 1 is not only conceptually useful, but also has practical implications. First, it renders the analysis of performance–privacy tradeoffs tractable. The max-separable structure implies that, given a privacy budget δ, it suffices to check whether the loss index of each signal falls below δ, without the need to allocate the budget optimally across signals. I demonstrate how this property leads to tractable tools for solving worst-case privacy-constrained problems in Subsection 4.1. Second, from a policy standpoint, the separability across signals also makes privacy regulation easier to monitor: the regulator only needs to check whether the likelihood vector associated with each signal is acceptable according to d.

Next, I discuss two technical aspects of Theorem 1; readers can skip to Subsection 2.3 without loss of continuity.

Table 3: Illustration of the "Only If" Direction of Theorem 1

(a) Original structure x

	x	
	s_1	s_2
θ_1	0.4	0.6
θ_2	0.2	0.8

(b) $0.5x' + 0.5x''$

	x' (0.5)		x'' (0.5)	
	s_1'	s_2'	s_1''	s_2''
θ_1	0.2	0.8	0.7	0.3
θ_2	0.1	0.9	0.6	0.4

(c) $0.25x + 0.75\tilde{x}$

	x (0.25)		$\tilde{x}$ (0.75)	
	s_1'	s_2''	s_2'	s_1''
θ_1	0.4	0.6	$\frac{8}{15}$	$\frac{7}{15}$
θ_2	0.2	0.8	0.6	0.4

Proof Idea of Theorem 1. The substantive content of Theorem 1 is that any worst-case privacy measure can be decomposed into likelihood vectors, *not just information structures*. The key idea is to construct the loss index of a likelihood vector ν as the infimum across the privacy losses of all information structures that contain $k\nu$ as a likelihood vector for some $k > 0$:

$$d(\nu) := \inf_{x \in \mathcal{X}_f:\ x^s = k\nu \text{ for some } k>0, s \in S_x} L(x). \tag{1}$$

Intuitively, if representation (WCP) holds, then $d(\nu) \leq L(x)$ for all such x. Equation 1 thus defines d as its largest possible value. With this definition, it is immediate that $\max_{s \in S_x} d(x^s) \leq L(x)$ for all $x \in \mathcal{X}_f$. Next, I illustrate why the other direction $L(x) \leq \max_{s \in S_x} d(x^s)$ holds.

The main idea is to show that for any x, we can construct an information structure that contains x as a component and has a privacy loss of $\max_{s \in S_x} d(x^s)$. For illustration, consider the information structure x in Table 3a. The likelihood vectors are highlighted in red for s_1 and in blue for s_2. For each of them, take auxiliary information structures x' and x'' that approximately attain their respective loss indices, so that $L(x') \approx d(x^{s_1})$, $L(x'') \approx d(x^{s_2})$. An example of x' and x'', and their mixture, $0.5x' + 0.5x''$, are shown in Table 3b. This mixture is our desired construction. By Worst-Case Protection, its loss is $L(0.5x' + 0.5x'') = \max\{L(x'), L(x'')\} \approx \max\{d(x^{s_1}), d(x^{s_2})\}$. On the other hand, $0.5x' + 0.5x''$ can also be written as a mixture of x and another structure $\tilde{x}$, specifically $0.25x + 0.75\tilde{x}$, as illustrated in Table 3c. By Worst-Case Protection and Blackwell Monotonicity, this equivalence implies

$$L(x) \leq L(0.25x + 0.75\tilde{x}) = L(0.5x' + 0.5x'') \approx \max\{d(x^{s_1}), d(x^{s_2})\}.$$

The formal proof extends this reasoning to arbitrary information structures and removes the approximation.

A Belief-Based Formulation of Theorem 1. There is an alternative formulation of Theorem 1 in terms of *posterior distributions*. Fix an arbitrary full-support reference measure

$\lambda \in \Delta(\Theta)$, which can be interpreted as a prior. Given λ, any information structure x induces a distribution of posteriors over Θ with mean λ, denoted by τ_x^λ. Under Blackwell Monotonicity, all information structures that induce the same distribution of posteriors entail the same privacy loss. Hence, the privacy measure L can equivalently be defined on the space of posterior distributions with mean λ. From this perspective, Worst-Case Protection requires that for any posterior distributions τ', τ'' with mean λ and any $\alpha \in (0,1)$, $L(\alpha\tau' + (1-\alpha)\tau'') = \max\{L(\tau'), L(\tau'')\}$. Moreover, Theorem 1 implies that L admits a max-separable representation across posteriors through a *belief-based index function* $\tilde{d} : \Delta(\Theta) \to \overline{\mathbb{R}}_+$, defined by[12]

$$\tilde{d}(\mu) := d\left(\left(\frac{\mu(\theta_k)}{\lambda(\theta_k)}\right)_{k=1}^{K}\right), \quad \forall\, \mu \in \Delta(\Theta); \qquad L(x) = \max_{\mu \in \operatorname{Supp} \tau_x^\lambda} \tilde{d}(\mu), \quad \forall\, x \in \mathcal{X}_f.$$

The function $\tilde{d}$ assigns a loss index to each posterior according to the loss index of the likelihood vector that induces it. Because d is homogeneous of degree 0, $\tilde{d}(\mu)$ depends only on the posterior itself and not on the specific likelihood vector generating it. Moreover, $\tilde{d}(\lambda) = 0$ since d vanishes on constant vectors, and $\tilde{d}$ is quasi-convex whenever d is, implying $L(\tau') \leq L(\tau'')$ whenever τ'' is a mean-preserving spread of τ'. Finally, the max-separable representation implies that L is decomposable not only across convex combinations of posterior distributions with mean λ, but also across posteriors, which need not have mean λ when treated as degenerate posterior distributions.

2.3. Specialization of Worst-Case Privacy

Theorem 1 establishes that any worst-case privacy measure admits a max-separable representation parameterized by an index function d. The functional form of d depends on judgments about which kinds of information are more harmful to reveal, which in turn reflects the objectives of potential adversaries. Blackwell Monotonicity captures the most permissive version of this idea: it requires $L(x) \geq L(x')$ only when x performs at least as well as x' in all decision problems. This axiom, however, is silent when some adversaries prefer x while others prefer x', leaving the index function d underdetermined.

In practice, when the relevant adversarial decision problems are known to belong to a class $\mathcal{C}$, it is natural to strengthen the axiom by requiring L to be monotonic with respect to $\mathcal{C}$. Formally, say x *$\mathcal{C}$-dominates* x' if, for every decision problem in $\mathcal{C}$, the expected utility under x is at least as high as under x'.

[12] A formal statement of this equivalence is provided in Lemma A.1.

Definition 2 ($\mathcal{C}$-Monotonicity)**.** Privacy measure L is *$\mathcal{C}$-monotonic* if for any $x, x' \in \mathcal{X}_f$ with x $\mathcal{C}$-dominating x', it holds that $L(x) \geq L(x')$.

Blackwell Monotonicity corresponds to the case where $\mathcal{C}$ is the class of all decision problems. Restricting $\mathcal{C}$ forces more pairs of information structures to be comparable and thus narrows the class of admissible privacy measures.

2.3.1. Prediction Problems

A natural class of adversarial decision problems is the class of prediction problems, where the adversary aims to predict some aspect of the state and receives a positive payoff only when the prediction is correct. To model such problems, I assume the state space has a Cartesian product structure $\Theta = \times_{n=1}^{N} T_n$, with a full-support prior $\mu_0 \in \Delta(\Theta)$.[13] Here, each T_n represents an *aspect* of the state that some adversaries may seek to predict. For example, these aspects may correspond to an individual's demographic characteristics (e.g., T_1 and T_2 denote race and gender), or to the types of different individuals (e.g., T_n is the type space of individual n). I denote by θ_n the n-th coordinate of the state θ, which takes value $t_n \in T_n$.

The class of prediction problems captures all adversaries interested in predicting one of the aspects.

Definition 3 (Prediction Problems)**.** Given (Θ, μ_0), decision problem (μ_0, A, u) is a *prediction problem* if for some $n \in \{1, ..., N\}$ and $c : T_n \to \mathbb{R}_+$, it holds that $A = T_n$ and

$$u(\theta, a) = \begin{cases} c(a), & \text{if } \theta_n = a \\ 0, & \text{otherwise} \end{cases}.$$

The class of prediction problems is denoted by $\mathcal{C}^P_{\mu_0}$.

In a prediction problem, the adversary receives a payoff only when correctly predicting the n-th aspect of θ for some n. Prominent examples include attribute inference attacks and re-identification attacks, which capture privacy threats frequently encountered in practice (Yeom, Giacomelli, Fredrikson, and Jha, 2018; Sweeney, 2002).

Example 1 (Attribute Inference Attack)**.** In an attribute inference attack, the adversary's goal is to infer an individual's characteristic that is missing. He is interested in an aspect

[13] This setting admits three generalizations. First, the state space may include auxiliary information O that does not require protection, so that $\Theta = \times_{n=1}^{N} T_n \times O$. Second, each protected aspect need not form a partition of Θ; they may be any disjoint nonempty subsets. Third, the prior need only have full support on each marginal T_n, instead of on the entire Θ. See Theorem A.1 for the formal statement.

T_n, which can represent gender or race. Θ contains the complete profile of the individual, S is the space of coarsened profiles, and x is a garbling algorithm. The adversary, upon observing the garbled profile, attempts to predict aspect T_n. □

Example 2 (Re-identification Attack)**.** In a re-identification attack, the adversary's goal is to infer the data owner's identity from anonymized data. He is interested in an aspect T_n, which is the identity of the n-th individual. Θ contains all possible datasets with personal identities, S is the space of anonymized data, and x is the anonymization procedure. The adversary, upon observing the anonymized data, attempts to predict the individual n's identity. □

2.3.2. Prediction Privacy

Since prediction problems represent many commonly encountered privacy threats, one might be interested in privacy measures that are monotonic in this class of decision problems, which are defined as *prediction privacy* (PP) measures.

Definition 4 (Prediction Privacy)**.** Privacy measure L is a prediction privacy measure if it satisfies Worst-Case Protection, $C^P_{\mu_0}$-Monotonicity, and normalization.

Next, I characterize the class of prediction privacy measures. For two elements $t_n, t'_n \in T_n$ of aspect n, define the log-likelihood ratio (LLR) update after observing signal s as[14]

$$\ell_{t_n,t'_n}(x^s) := \log \frac{\sum_\theta x(s|\theta)\,\mu_0(\theta|t_n)}{\sum_\theta x(s|\theta)\,\mu_0(\theta|t'_n)}. \tag{2}$$

A larger value of ℓ_{t_n,t'_n} means that, under x, signal s provides stronger evidence toward t_n relative to t'_n. Finally, note that if there is only one aspect such that $\Theta = T$, prior μ_0 is irrelevant.

Theorem 2. *Privacy measure L is a prediction privacy measure if and only if*

$$L(x) = \max_{s\in S_x,\ 1\le n\le N,\ t_n,t'_n\in T_n} d_{t_n,t'_n}\left(\ell_{t_n,t'_n}(x^s)\right), \quad \forall x \in \mathcal{X}_f, \tag{PP}$$

for some functions $d_{t_n,t'_n} : \overline{\mathbb{R}} \to \overline{\mathbb{R}}_+$ where each d_{t_n,t'_n} is increasing with $d_{t_n,t'_n}(0) = 0$.

Theorem 2 shows that when prediction problems are the relevant threats, the index function necessarily takes a pairwise form:

$$d(x^s) = \max_{1\le n\le N,\ t_n,t'_n\in T_n} d_{t_n,t'_n}\left(\ell_{t_n,t'_n}(x^s)\right).$$

[14] I use the convention that $0/0 = 1$.

For each signal s, the loss index depends on how strongly the signal shifts beliefs toward t_n relative to t'_n, captured by the log-likelihood ratios $\ell_{t_n,t'_n}(x^s)$ and transformed by the corresponding *weighting functions* d_{t_n,t'_n}. The overall loss is then given by the maximum across all signals s, aspects n, and pairs in T_n. This pairwise decomposition substantially restricts admissible index functions.[15]

This pairwise structure is due to a key feature of prediction problems: the comparison between predicting t_n and t'_n depends only on the belief about this pair, since both predictions yield zero payoff when θ_n is a third element of T_n under the true state. For instance, when deciding whether the owner of the anonymized data is Alice or Bob, it is irrelevant whether the probability of the owner being Charlie is high or low. This feature gives rise to the observation that, if the maximal LLR between t_n and t'_n under x is at least as large as under x', i.e.

$$\max_{s\in S_x} \ell_{t_n,t'_n}(x^s) \;\geq\; \max_{s\in S_{x'}} \ell_{t_n,t'_n}(x'^s),$$

for all n and $t_n, t'_n \in T_n$, then it follows that $L(x) \geq L(x')$. Building on this observation, one can construct functions d_{t_n,t'_n} in a manner analogous to Equation 1.

Another feature of representation (PP) is the flexibility of the weighting functions d_{t_n,t'_n}. Because these functions may differ across pairs, the measure accommodates asymmetric treatment of aspects or of types within an aspect. For example, updates about health status may be judged more harmful than those about residential address; within health status, updates distinguishing between "healthy" and "genetic disease" may be considered more sensitive than those between "healthy" and "regular disease." In practice, the choice of weighting functions depends on normative judgments about which kinds of information are more sensitive and therefore warrant stricter protection.

A benchmark case is when all weighting functions are identical and, without loss of generality, taken to be the identity function.[16] Such a choice is particularly appropriate when the T_n's represent demographic groups of individuals, such as race or gender, where identical weights across pairs reflect equal treatment between types and across individuals.

[15] As a heuristic illustration, consider the case of one aspect, $\Theta = T$, and suppose each dimension of $\overline{\mathbb{R}}_+^K$ is discretized into m values. Specifying a general index function $d : \mathbb{R}_+^K \to \overline{\mathbb{R}}_+$ would then require assigning a value to each of the m^K grid points, with each assignment chosen from m possibilities, for a total of m^{m^K} specifications. By contrast, specifying the pairwise functions $d_{t,t'} : \overline{\mathbb{R}} \to \overline{\mathbb{R}}_+$ requires approximately $K^2 m^m$ specifications. As the number of states K grows, this latter quantity increases at a much slower rate.

[16] This is without loss in the sense that any prediction privacy measure with identical weights coincides with prediction privacy under identity weights up to an increasing transformation.

I refer to this measure as the *benchmark prediction privacy* (BPP):

$$L_{BPP}(x) = \max_{s \in S_x,\, 1 \le n \le N,\, t_n, t'_n \in T_n} \ell_{t_n, t'_n}(x^s), \quad \forall\, x \in \mathcal{X}_f. \tag{BPP}$$

As an illustration, consider information structure C in Figure 1. The loss index of the red likelihood vector equals the maximum of $\log(7/5)$, $\log 5$, and $\log 7$ (and their negatives), which is $\log 7$. Similarly, the blue likelihood vector has loss index $\log 3$. Hence, under BPP, the privacy loss of information structure C, represented by the red star, is $\log 7$.

I conclude this section by showing that local differential privacy (LDP; Kasiviswanathan, Lee, Nissim, Raskhodnikova, and Smith 2011), a widely used notion in the computer science literature, is a special case of BPP when there is only one aspect. Hence, Theorem 2 provides a decision-theoretic interpretation of LDP.

Example 3 (Local Differential Privacy)**.** Suppose there is a single aspect so that $\Theta = T$. Then BPP takes the form

$$L(x) = \max_{s \in S_x,\, \theta, \theta' \in \Theta} \log \frac{x(s|\theta)}{x(s|\theta')},$$

which coincides exactly with the definition of local differential privacy. The prior μ_0 is irrelevant since each $t \in T$ corresponds to a single state. Intuitively, LDP limits how much information can be inferred between any two of an individual's types. These protections are typically applied before data are transmitted to a centralized system, which explains the term "local." □

Remark 1. The widely used notion of differential privacy (Dwork et al., 2006) admits a similar interpretation. A detailed discussion and a comparison with BPP are provided in Subsection 5.1.

Figure 2 summarizes the privacy measures introduced in this section and their relationships. The regulatory principle of uniform protection characterizes the workhorse class of Worst-Case Privacy. Focusing on prediction problems—a class of salient privacy threats in practice—yields the pairwise structure of the index function d, which defines Prediction Privacy. Finally, the benchmark prediction privacy measure corresponds to the case in which all weighting functions are identity functions, reflecting equal treatments.

3. Privacy in Two-Sided Matching

Privacy concerns have been highlighted in market design (Roth, 2018). One salient context is matching markets, as matching outcomes can reveal sensitive information about participants. For example, in school choice, a student's placement may reflect her home

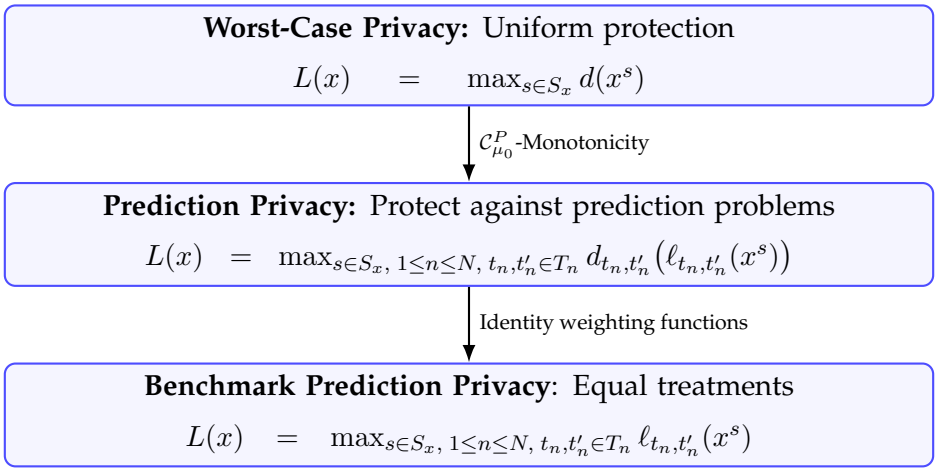

Figure 2: Summary of Privacy Measures

address or special academic needs. In this section, I use the benchmark prediction privacy to evaluate privacy properties of stable matching mechanisms in a canonical two-sided, many-to-one matching market.[17] I demonstrate a sharp tradeoff between welfare and privacy.[18]

3.1. Model Setup

Consider a finite set of workers W and a finite set of firms F. Each firm f has a publicly known capacity $q_f \geq 1$. I assume that all matches are mutually acceptable, $|F| \geq 2$, and $|W| \geq \sum_{f \in F} q_f$. These assumptions imply that all capacities of firms will be filled in any stable matching. Each worker has a strict preference ranking over firms. Firms' preferences are assumed to be responsive to a strict individual ranking over workers.[19]

Let Θ_w denote the set of strict rankings over firms for worker w, and let Θ_f denote the set of strict rankings over workers for firm f. The set of complete preference profiles is Θ, with a typical element $\boldsymbol{\theta} = (\boldsymbol{\theta}^W, \boldsymbol{\theta}^F)$. I assume workers' preferences are drawn i.i.d. according to a full-support distribution over Θ_w, while firms' preferences are drawn independently and uniformly at random.[20] Let μ_0 denote such a prior on Θ. In Appendix B, I extend this assumption to allow correlation across firms' preferences.

[17] For a classic reference, see Chapter 5 of Roth and Sotomayor (1990).

[18] In this application, "welfare" refers specifically to the quality of matches, not including privacy concerns.

[19] Formally, a firm's preference over sets of workers is responsive to a strict individual ordering $\succ$ if for any set S with $|S| < q_f$ and any worker $w' \notin S$, it holds that $S \cup \{w'\} \succ_f S$, and for any $w, w' \notin S$, $S \cup \{w'\} \succ_f S \cup \{w\}$ if and only if $w' \succ w$.

[20] These assumptions are standard in large random matching markets; see, e.g., Pittel (1989) and Ashlagi, Kanoria, and Leshno (2017). Firms' preferences over sets of workers can be arbitrary as long as it is responsive to the realized ranking over individual workers.

A matching is a mapping $m : W \to F \cup \{\emptyset\}$. A matching m is stable under a preference profile $\boldsymbol{\theta}$ if (i) no firm's capacity is exceeded, and (ii) there is no worker–firm pair (w, f) such that w prefers f to $m(w)$ and f either has remaining capacity or prefers w to some worker it is currently matched with. Let $M(\boldsymbol{\theta})$ denote the set of stable matchings under profile $\boldsymbol{\theta}$, and let M be the set of all possible stable matching outcomes. A *stable matching mechanism* is a mapping $x : \boldsymbol{\theta} \mapsto \Delta(M(\boldsymbol{\theta}))$.

Privacy Concern. I focus on protecting workers' privacy, assuming that the matching outcome m is publicly observable and thus treated as a signal. This notion is natural in matching markets where final assignments are announced and can reveal sensitive information about workers' preferences. In this case, each worker's preference is treated as a protected aspect, and firms' preferences are auxiliary information (see fn. 13). I apply the benchmark prediction privacy measure, which takes the following form:

$$L_{BPP}(x) = \max_{m \in M,\, w \in W,\, \theta_w, \theta'_w \in \Theta_w} \log \frac{\Pr_x(m|\theta_w)}{\Pr_x(m|\theta'_w)},$$

where $\Pr_x$ denotes the joint distribution over $(\boldsymbol{\theta}, m)$ under mechanism x and prior μ_0.

3.2. The Privacy-Welfare Tradeoff

Given a preference profile $\boldsymbol{\theta}$, define the partial order $>_W$ on the set of stable matchings $M(\boldsymbol{\theta})$ by $m >_W m'$ if all workers weakly prefer m to m' and at least one strictly prefers it. The analogous order $>_F$ is defined for firms. It is a classic result (see, e.g., Corollary 5.32 of Roth and Sotomayor (1990)) that $M(\boldsymbol{\theta})$ forms a lattice under either order, with $m >_W m'$ if and only if $m <_F m'$.[21] In particular, there exist two canonical stable matchings at the extremes: the worker-optimal and the firm-optimal. All workers (resp. firms) prefer the worker-optimal (resp. firm-optimal) matching to any other stable matchings. I refer to the deterministic mechanisms that always select these most-preferred stable matchings as the worker-optimal and firm-optimal stable matching mechanisms.

Proposition 1. *The firm-optimal stable matching mechanism x^F minimizes worker privacy loss among all stable matching mechanisms: $L_{BPP}(x^F) \leq L_{BPP}(x)$ for all stable matching mechanisms x. In particular, $L_{BPP}(x^F) > 0$, so no stable matching mechanism can have zero privacy loss.*

The underlying mechanism behind Proposition 1 is intuitive: improvements in worker welfare make the realized matching outcomes more correlated with workers' preferences.

[21] A partially ordered set is a lattice if every pair of elements has a least upper bound and a greatest lower bound.

This stronger correlation increases the LLR between worker types who rank the matched firm higher and lower, thereby raising the benchmark prediction privacy loss. As a result, the firm-optimal stable mechanism, which is worst for worker welfare, minimizes these LLRs. Note that this argument depends on the LLR representation of the benchmark prediction privacy measure.[22]

To provide more details, I introduce two natural classes of stable matching mechanisms. A mechanism is *symmetric* if it treats all workers identically ex-ante. Formally, for any permutation π on W, x is symmetric if $x(\pi(m)|\pi(\boldsymbol{\theta})) = x(m|\boldsymbol{\theta})$, where $\pi(m)$ is the matching obtained by relabeling workers in m and $\pi(\boldsymbol{\theta})$ is the profile where workers' preferences and their ranks in firms' preferences are relabeled according to π. A mechanism is *monotonic* if a worker is weakly more likely to be matched with a firm when she ranks it higher: for any two preference lists θ_w^i, θ_w^j for worker w where firm f is ranked i-th and j-th respectively with $i < j$, it holds that $\Pr_x(m(w) = f|\theta_w^i) \geq \Pr_x(m(w) = f|\theta_w^j)$.[23] Both the firm- and worker-optimal stable matching mechanisms are symmetric and monotonic (see Lemma B.3 in Appendix B).

The argument for Proposition 1 is as follows. I first show that, in minimizing privacy loss, it is without loss to focus on symmetric mechanisms. I then show that, for symmetric mechanisms, the benchmark prediction privacy L_{BPP} can be characterized by the probabilities of matching with a firm at a given rank in the workers' preference lists, with no need to consider the entire matching m. In particular, for monotonic mechanisms such as the firm-optimal stable mechanism, L_{BPP} is driven by the likelihood ratio of a worker being matched with her first and last choices. Since the firm-optimal mechanism minimizes this ratio among all symmetric mechanisms, we conclude that it is the most private.[24]

This logic extends to a more general characterization of the welfare–privacy tradeoff. For any mechanism x, let $F_w^x(\boldsymbol{\theta})$ denote the distribution of the rank of worker w's assigned firm under $x(\cdot|\boldsymbol{\theta})$ (with unmatched status treated as rank $|F|+1$). Say that x is *better for worker welfare* than x' if, for every w and $\boldsymbol{\theta}$, the distribution of ranks $F_w^{x'}(\boldsymbol{\theta})$ first-order stochastically dominates $F_w^x(\boldsymbol{\theta})$ (i.e., workers are weakly more likely to be matched to higher-ranked firms under x than under x').

[22] Indeed, I show in Appendix B.1 that the worker-optimal stable mechanism does not Blackwell dominate the firm-optimal one regarding their informational content about a worker's preference.

[23] A sufficient condition for monotonicity is $x(m(w) = f|\theta_w^i, \boldsymbol{\theta}_{-w}) \geq x(m(w) = f|\theta_w^j, \boldsymbol{\theta}_{-w})$ for all w and $\boldsymbol{\theta}_{-w}$. This condition is not vacuous because, if $m(w) = f$ and m is stable under $(\theta_w^j, \boldsymbol{\theta}_{-w})$, then m is also stable under $(\theta_w^i, \boldsymbol{\theta}_{-w})$.

[24] Note that Proposition 1 does not claim the firm-optimal mechanism is the uniquely most private mechanism. In fact, there can be other mechanisms that are as private as the firm-optimal one but strictly better for worker welfare. A full characterization of the welfare–privacy frontier is left for future research.

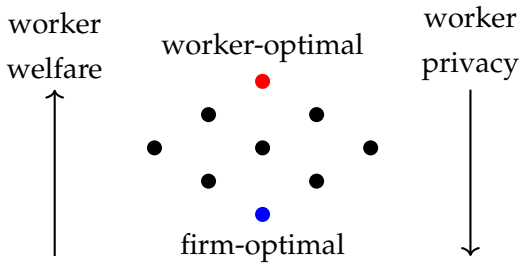

Figure 3: Tradeoff between worker welfare and privacy. The dots represent symmetric and monotonic stable matching mechanisms, whose vertical positions reflect the relation "better for worker welfare."

Proposition 2. *Suppose x' is a symmetric and monotonic stable mechanism. If a stable mechanism x is better for worker welfare than x', then $L_{BPP}(x) \geq L_{BPP}(x')$.*

Proposition 2 formalizes the welfare–privacy tradeoff, which is illustrated schematically in Figure 3. Within the class of symmetric and monotonic stable mechanisms, any improvement in worker welfare necessarily entails (weakly) higher privacy loss. The worker-optimal and firm-optimal mechanisms thus lie at the opposite ends of this frontier. The first part of Proposition 1 follows directly as a corollary of this more general result.

A sharp implication arises in balanced one-to-one matching markets with preferences drawn independently and uniformly at random. Restricting attention to stable mechanisms that are symmetric and monotonic on both sides, moving along the lattice of stable matchings from the worker-optimal to the firm-optimal mechanism produces a four-way simultaneous shift: workers' welfare falls while their privacy improves, and firms' welfare rises while their privacy worsens.

4. Optimal Design under Worst-Case Privacy

Worst-Case Privacy is also a useful benchmark for privacy-aware designs. In this section, I formulate optimization problems under worst-case privacy constraints and establish a certification result that serves as a tool for identifying optimal privacy-constrained solutions. I then apply these tools to two settings: voting and credit screening.

4.1. Worst-Case Privacy Constrained Optimization

I begin with a general framework for privacy-constrained optimization, which encompasses various settings, including information design and mechanism design. Let A be a finite set, and let the designer choose a mapping $x : \Theta \to \Delta(A)$, generally referred to as a

mechanism.[25] The designer's reduced-form utility is $u(x) \in [-\infty, \infty)$. In information design problems, the set A can be interpreted as action recommendations. If $\hat{u} : \Theta \times A \to \mathbb{R}$ is the state-contingent payoff function and μ_0 is the prior, then one standard specification of u is

$$u(x) = \sum_{\theta,a} \hat{u}(\theta, a)\, x(a|\theta)\mu_0(\theta),$$

whenever x satisfies the relevant feasibility constraints (e.g., obedience or IC constraint), and $u(x) = -\infty$ otherwise. In mechanism design problems, set A can instead represent allocations, and u admits an analogous microfoundation. Thus, the standard problem can be compactly written as $\sup_{x:\Theta\to\Delta(A)} u(x)$.

The use of a mechanism x may, however, raise privacy concerns, since the realized action $a \in A$ conveys information about the state θ through x. To address this, privacy-aware design imposes a worst-case privacy constraint: $L(x) \leq \delta$ for some privacy budget $\delta \geq 0$. The designer's privacy-constrained optimization problem is therefore

$$\begin{aligned} U(\delta) := \sup_{x:\Theta\to\Delta(A)} \quad & u(x) \\ \text{s.t.} \quad & L(x) \leq \delta, \end{aligned} \tag{$\mathcal{P}_\delta$}$$

where $L(x) = \max_{a\in A} d(x^a)$ is a worst-case privacy measure as characterized in Theorem 1.

Next, I provide a certification result that can be used to find the optimal solutions of Program ($\mathcal{P}_\delta$). A key feature of this setup is that the worst-case constraint can be separated across actions: mechanism x satisfies the constraint if and only if $x(a|\cdot) \in C_\delta$ for all a, where

$$C_\delta := \{\nu \in \mathbb{R}^K_+ : d(\nu) \leq \delta\}$$

is the set of likelihood vectors whose loss index is no higher than δ. As a lower-contour set of d, the set C_δ is a nonempty convex cone. Given a full-support reference measure $\mu_0 \in \Delta(\Theta)$, which is usually taken to be the prior in applications, define the *μ_0-weighted polar cone* of C_δ as

$$P_\delta^{\mu_0} := \left\{ p \in \mathbb{R}^K : \sum_\theta p(\theta)\nu(\theta)\mu_0(\theta) \leq 0,\ \forall \nu \in C_\delta \right\}.$$

[25] In standard information and mechanism design settings, it is without loss to treat A as the signal space. This rests on two principles: (i) the revelation principle, which justifies restricting attention to direct (recommendation) mechanisms without loss of optimality, and (ii) Blackwell Monotonicity, which implies that pooling signals that lead to the same action can only weakly reduce privacy loss. In mechanism design or social choice settings, it is implicitly assumed that the realized allocation a is publicly observable. The framework can be adapted to alternative cases—for instance, when only part of the allocation is observable.

Let $p(\theta, a) \in \mathbb{R}$ be the shadow price of action a in state θ. Define the following program as the "priced" version of Program ($\mathcal{P}_\delta$):

$$W(p) := \sup_{x:\Theta\to\Delta(A)} \quad u(x) - \sum_{\theta,a} p(\theta, a)x(a|\theta)\mu_0(\theta) \tag{$\mathcal{W}_p$}$$

The next result shows that if x solves Program ($\mathcal{W}_p$) for an appropriate set of prices, then x is an optimal solution to ($\mathcal{P}_\delta$).

Theorem 3 (Sufficient Condition for Optimality)**.** *Suppose mechanism x satisfies $L(x) \leq \delta$ and there exist shadow prices $p(\theta, a)$ such that:*

1. ***(Price Validity)*** *$p(\cdot, a) \in P_\delta^{\mu_0}$ for all $a \in A$.*
2. ***(Priced Optimality)*** *x solves the priced problem ($\mathcal{W}_p$).*
3. ***(Complementary Slackness)*** *$\sum_\theta p(\theta, a)x(a|\theta)\mu_0(\theta) = 0$ for all $a \in A$.*

Then x solves ($\mathcal{P}_\delta$).

Remark 2. In Appendix C.1, I establish the converse direction under additional assumptions. In that case, ($\mathcal{P}_\delta$) can be solved by minimizing the priced problem ($\mathcal{W}_p$) over all p such that $p(\cdot, a) \in P_\delta^{\mu_0}$ for all $a \in A$.

The conditions of Theorem 3 can be viewed as a competitive equilibrium in a "privacy market." Each action a plays the role of a seller, offering likelihood vectors $x(a|\cdot)$ from its feasible set C_δ, while the designer acts as the buyer, facing prices $p(\theta, a)\mu_0(\theta)$. Price Validity ensures competitiveness so that no seller can have a positive profit, Priced Optimality ensures the buyer's choice is utility-maximizing given these prices, and Complementary Slackness guarantees that sellers break even. This market interpretation clarifies why worst-case privacy constraints tend to be analytically tractable: the "supply" of x decouples across actions, allowing each to be priced separately. Such separability is distinctive to worst-case privacy measures. By contrast, under average-case measures such as mutual information, the shadow price of information revealed through one action depends on the informational contents of all others, entangling the problem across actions.

One potential hurdle in applying Theorem 3 is characterizing the polar cone $P_\delta^{\mu_0}$. Fortunately, for the benchmark prediction privacy measure, the polar cone admits a clean description. Let $\mathcal{H}$ be a set of disjoint nonempty subsets of Θ. Let $d_{\mu_0,\mathcal{H}} := \max_{H_i,H_j\in\mathcal{H}} \ell_{H_i,H_j}$, where ℓ_{H_i,H_j} is the log-likelihood ratio update between H_i and H_j defined in Equation 2. For a price vector $p : \Theta \to \mathbb{R}$, define $p_i := \max_{\theta\in H_i} p(\theta)$ and its positive/negative parts $p_i^+ := \max\{p_i, 0\}$ and $p_i^- := \max\{-p_i, 0\}$.

Lemma 1. *1. For $d_{\mu_0,\mathcal{H}}$, a price vector p lies in $P_\delta^{\mu_0}$ if and only if*

$$\sum_i p_i^- \, \mu_0(H_i) \;\geq\; e^\delta \sum_i p_i^+ \, \mu_0(H_i), \qquad \textit{and} \qquad p(\theta) \leq 0 \;\; \forall \theta \notin \cup_i H_i.$$

2. If $d = \max_n d_n$ for finitely many index functions d_n with μ_0-weighted polar cones $P_{\delta,n}^{\mu_0}$, then their Minkowski sum satisfies $\sum_n P_{\delta,n}^{\mu_0} \subset P_\delta^{\mu_0}$.

Lemma 1 provides explicit tools for applications. In particular, it shows that for the benchmark prediction privacy measure L_{BPP}, one can construct shadow prices piecewise: if for every aspect n, price p_n is valid for d_{μ_0,T_n}, then $\sum_n p_n$ is valid for the index function of L_{BPP}. This makes Theorem 3 straightforward to apply in practice.

Remark 3 (Data Minimization). A complementary perspective is to minimize privacy loss, subject to achieving a target level of utility. This corresponds directly to the legal principle of data minimization.[26] Proposition C.1 shows that both utility maximization and data minimization trace out the same privacy–utility frontier and that the same machinery developed in this section can also be applied to address data minimization problems.

4.2. Application to Voting

Privacy concerns in voting, such as the possibility of vote revelation, have attracted growing attention in law and policy debates (Adler and Hall, 2013; Congressional Research Service, 2024; Kuriwaki, Lewis, and Morse, 2025). Failure to protect voter privacy can facilitate vote buying or retaliation, thereby undermining collective decision-making (Keyssar, 2009; Mares, 2015). Yet despite these concerns, formal economic analysis of how privacy protection shapes voting remains limited. To fill this gap, I apply my framework to study optimal voting rules subject to the benchmark prediction privacy constraint.

I consider a stylized model in which the state of the world is binary, $\Omega = \{\omega_0, \omega_1\}$, with a uniform prior. There are an odd number of $n \geq 3$ voters, each receiving a private binary signal $t_i \in T := \{t^0, t^1\}$. The signals are conditionally independent and symmetric, with quality $q = \Pr(t^1|\omega_1) = \Pr(t^0|\omega_0) \in (1/2, 1)$. Let $\Theta = T^n$ be the set of signal profiles. To isolate the role of privacy, I assume the voters report truthfully.

A social planner chooses a collective action $a \in \{a_0, a_1\}$ to match the state, with $v(\omega_1, a_1) = v(\omega_0, a_0) = 1$ and zero otherwise. The planner's interim expected payoff from action a_1,

[26] For example, GDPR (Article 5(1)(c)) requires personal data to be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed."

given a signal profile θ with m signals of t^1, is the posterior probability of state ω_1:

$$\Pr(\omega_1|m) = \frac{1}{1 + \left(\frac{1-q}{q}\right)^{2m-n}}.$$

The planner's net benefit from choosing a_1 over a_0 when there are m signals of t^1 is

$$\Delta v(m) := \Pr(\omega_1|m) - \Pr(\omega_0|m) = 2\Pr(\omega_1|m) - 1. \tag{3}$$

A voting rule is a mapping $x : \Theta \to \Delta(A)$. The planner chooses x to maximize expected utility subject to the benchmark prediction privacy constraint. Formally, the planner's problem is:

$$\begin{aligned} \max_{x:\Theta\to\Delta(A)} \quad & \sum_{\theta,a} \Pr(\omega_a|\theta)\, x(a|\theta)\, \mu_0(\theta) \\ \text{s.t.} \quad & L_{BPP}(x) = \max_{i,a,t_i,t_i'} \log \frac{\Pr_x(a|t_i)}{\Pr_x(a|t_i')} \leq \delta, \end{aligned} \tag{$\mathcal{V}_{BPP}$}$$

where $\Pr_x$ denotes the joint distribution under x and the prior.

Without a privacy constraint (i.e., $\delta = \infty$), the simple majority rule is optimal. However, under simple majority, observing the collective action reveals that each voter is more likely to have voted for that action, which can violate the privacy constraint when δ is small. In this case, the following proposition characterizes the optimal privacy-constrained voting rule. Let $m(\theta)$ denote the number of t^1 signals in profile θ.

Proposition 3. *The optimal solution to Program ($\mathcal{V}_{BPP}$) has a cutoff structure and is unique up to randomization at its thresholds. Specifically, there exists a threshold $\hat{m} \in (\frac{n}{2}, n]$ such that*

$$x(a_1|\theta) = \begin{cases} 1, & \text{if } n/2 < m(\theta) < \hat{m} \text{ or } m(\theta) < n - \hat{m}, \\ 0, & \text{if } n - \hat{m} < m(\theta) < n/2 \text{ or } m(\theta) > \hat{m}, \end{cases}$$

and the rule may randomize at the thresholds $m(\theta) = \hat{m}$ and $m(\theta) = n - \hat{m}$.

In words, the optimal voting rule exhibits extreme reversal: when the majority for an action becomes too large, the opposite action is instead selected. This reversal protects privacy by making the realized action ambiguous about the votes. When the collective decision is a_1, for instance, there are two possibilities: either there was a moderate majority for a_1, or there was an extreme majority for a_0. Because these two possibilities offset each other, the observed action reveals little about any individual's vote.

The intuition for why extreme reversal is optimal can be understood through the shadow

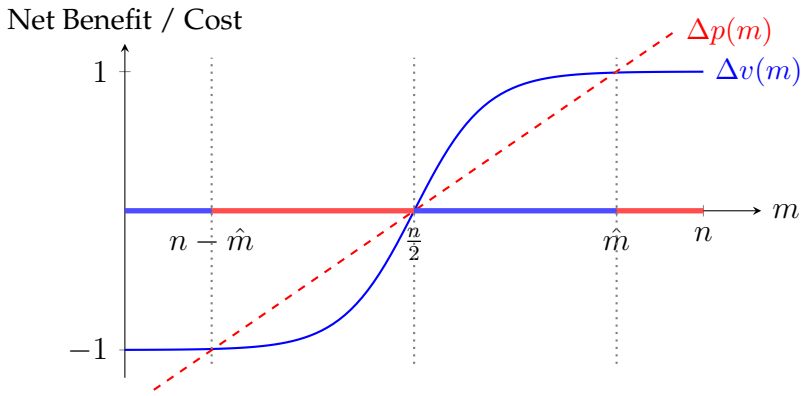

Figure 4: The optimal voting rule. The x-axis represents the number of t^1 signals, m. The blue curve is the net benefit of a_1, $\Delta v(m)$, and the red dashed line is the linear net privacy cost, $\Delta p(m)$. The colored interval on the axis indicates the chosen action: blue for a_1 and red for a_0.

prices of the privacy constraint. The planner's problem is to maximize expected utility, subject to a privacy cost. The net benefit of choosing a_1, $\Delta v(m)$, given by Equation 3, is S-shaped in m, as shown by the blue curve in Figure 4.

On the other hand, based on the symmetric structure of the problem, one can verify that the shadow prices of the privacy constraint are:

$$p(\theta, a_1) = m(\theta)\hat{p} - (n - m(\theta))e^{\delta}\hat{p}, \quad p(\theta, a_0) = (n - m(\theta))\hat{p} - m(\theta)e^{\delta}\hat{p},$$

for some price level $\hat{p} \geq 0$. Intuitively, $\hat{p}$ represents the shadow cost of assigning one additional t^1 signal to a_1 (or equivalently, one additional t^0 signal to a_0). Assigning a mismatched pair instead yields a relaxed cost of $e^{\delta}\hat{p}$, and this construction is guided by Lemma 1. Thus, when the profile contains $m(\theta)$ signals of t^1, the total shadow price is obtained by summing up these contributions, which explains the construction of $p(\theta, a_1)$ and $p(\theta, a_0)$.

The planner's priced problem is then to choose a_1 when $\Delta v(m)$ exceeds the net privacy cost of a_1, $\Delta p(m) := p(a_1, m) - p(a_0, m)$, which is linear in m. This tradeoff is illustrated in Figure 4. When the privacy constraint is not binding, $\hat{p} = 0$, and we recover the first-best majority rule. When the constraint binds, $\hat{p} > 0$. The S-shaped benefit and linear cost cross three times, leading to a cutoff structure where the planner follows the majority for moderate signal profiles but reverses the decision for extreme ones. The price level $\hat{p}$ (and thus the threshold $\hat{m}$) is pinned down by the binding privacy constraint.

This optimal voting rule has two welfare implications.

Corollary 1. *Let $v^*_{BPP}(\delta; n)$ be the optimal value of Program ($\mathcal{V}_{BPP}$). Then:*

1. *As* $n \to \infty$, $v^*_{BPP}(\delta; n) \to 1$ *if and only if* $\log \frac{q}{1-q} \leq \delta$.

2. *There is nontrivial improvement at perfect privacy:* $v^*_{BPP}(0; n) > 1/2$.

The first result shows that as the electorate grows, full information aggregation is possible if and only if signals are sufficiently *imprecise* relative to the privacy budget. The reason is that the benchmark prediction privacy restricts indirect inference about voters' signals through the state ω. Under the first-best voting rule (simple majority), as $n \to \infty$, the collective action a becomes almost perfectly informative about ω. When signals are very precise (large q), this makes a strongly revealing of each individual's signal, which violates the benchmark prediction privacy constraint.

The second result states that even under perfect privacy ($\delta = 0$), the planner can still improve upon a constant voting rule. This may seem surprising, as one might think that if a is informative about ω, it must be informative about voters' signals. However, it is possible for ω to be correlated with both a and t_i while a and t_i are themselves independent.[27] I illustrate how a marginal, welfare-improving deviation from a constant voting rule can be constructed while preserving perfect privacy in the following example.

Example 4. Suppose $n = 3$ and $\delta = 0$. Consider a marginal deviation from a rule that always picks a_0. Let the new rule x be such that $x(a_1|m = 2) = \varepsilon$ and $x(a_1|m = 0) = \gamma\varepsilon$ for some small $\varepsilon > 0$. The choice to select a_1 when $m = 2$ is welfare-improving, as a_1 is more likely to be correct. To maintain perfect privacy ($\delta = 0$), this must be balanced by selecting a_1 in a state where s_0 is more likely, such as $m = 0$. The perfect privacy condition, $\Pr_x(a_1|t_i = t^1) = \Pr_x(a_1|t_i = t^0)$, pins down the required compensation: $\gamma = \frac{q(1-q)}{q^3+(1-q)^3}$.

The net change in expected welfare from this deviation is

$$\Delta v(2)\mu_0(m = 2)\varepsilon \; + \; \Delta v(0)\mu_0(m = 0)\gamma\varepsilon.$$

The first term reflects the welfare gain from improving the decision when $m = 2$, while the second captures the welfare loss from worsening the decision when $m = 0$. The deviation is profitable because the gain outweighs the loss. The simplest case to see this is when q is close to 1. In this limit, $\Delta v(2) \approx -\Delta v(0)$: the informational value of two t^1 signals in favor of ω_1 is nearly the same as that of three t^0 signals in favor of ω_0. At the same time, simple algebra shows that $\mu_0(m = 2)\varepsilon = 3\,\mu_0(m = 0)\gamma\varepsilon$. The factor of 3 arises because the signal-count difference is 1 at $m = 2$ but 3 at $m = 0$, so preserving perfect privacy requires the

[27] Relatedly, He et al. (2021) studies the boundary of informativeness about an underlying state under the requirement that signals are independent.

former to be three times more likely. Combining these welfare and probability calculations, we see that the net change in expected welfare is strictly positive. □

Finally, I remark that the pattern of extreme reversal also appears in the optimal voting mechanisms studied by Chwe (2010), Kattwinkel and Winter (2024), and Best, Quigley, Saeedi, and Shourideh (2025). In their settings, however, the reversal arises from incentive-compatibility constraints of biased voters, whereas here it emerges as the most effective way to aggregate information while protecting privacy. Proposition 3 is intended to illustrate the forces that the benchmark prediction privacy introduces into voting, rather than to suggest that the mechanism is practical. In fact, different privacy considerations can yield distinct implications. I show in Subsection 5.1 that other natural privacy measures need not produce extreme reversals.

4.3. Application to Credit Screening

The Equal Credit Opportunity Act prohibits lenders from discriminating based on protected characteristics such as race or gender. One way to enforce this principle is to limit the information disclosed to lenders so that it does not reveal too much about these characteristics, thereby preventing discrimination at its source. This section examines how such privacy constraints can be implemented and their welfare implications for borrowers.

A credit agency (e.g., Experian) advises a lender on whether to approve a loan to an individual of type $(t, \theta) \in [0,1] \times \Theta$, where t is a credit score and θ is a demographic group. The prior on θ is μ_0, and the score t follows distribution F_θ conditional on θ, assumed to be continuous and strictly increasing. The lender's profit is $v(t, \theta)$ if the individual is approved and 0 otherwise, where v is continuous, strictly increasing in t, and crosses zero at some $t^*_\theta \in (0,1)$.

The credit agency chooses a recommendation mechanism $\tilde{x}(a|t, \theta)$ for $a \in \{a_0, a_1\}$ (reject, approve) to maximize the lender's expected profit subject to the BPP constraint on θ. Three simplifications apply. First, by the revelation principle and Blackwell Monotonicity, it is without loss to focus on recommendation mechanisms (see fn. 25). Second, since the credit agency and the lender share aligned incentives, non-obedient mechanisms are never optimal, so obedience constraints can be ignored. Third, because the payoff is single-crossing in (t, a), any optimal mechanism is a threshold rule within each group θ, represented by its rejection rate $q_\theta := x(a_0|\theta)$. Thus, we can restrict attention to mechanisms $x : \Theta \to \Delta(A)$, with $\tilde{x}$ implicitly rejecting individuals with the lowest credit scores in each group θ given x. This places the problem squarely within the framework of Subsection 4.1.

The credit agency's problem reduces to choosing $(q_\theta)_{\theta\in\Theta} \in [0,1]^K$ to maximize

$$u(x) := \sum_{\theta\in\Theta} \mu_0(\theta) \int_{q_\theta}^1 v(F_\theta^{-1}(s), \theta)\, ds,$$

subject to the BPP constraint

$$\max_{a\in A,\, \theta,\theta'\in\Theta} \log \frac{x(a|\theta)}{x(a|\theta')} \le \delta.$$

This is an instance of Program ($\mathcal{P}_\delta$).

Let the first-best (unconstrained) rejection rates be $q_\theta^* := F_\theta(t_\theta^*)$. To reduce case discussions, assume $q_\theta^* \le 1/2$, so that within each group, most individuals are approved under the first-best. Under this assumption,

$$\delta^* := \max_{\theta,\theta'} \log \frac{q_\theta^*}{q_{\theta'}^*}$$

is the privacy loss of the first-best mechanism. For $\delta \ge \delta^*$, the first-best is feasible; for $\delta < \delta^*$, the privacy constraint binds.

Proposition 4. *In the credit screening application, when $\delta \in [0, \delta^*]$, there exists a threshold $q(\delta)$ such that the uniquely optimal rejection rates are*

$$q_\theta = \max\{\, q(\delta),\ \min\{q_\theta^*,\ e^\delta q(\delta)\}\}.$$

Moreover, $q(\delta)$ is strictly decreasing and $e^\delta q(\delta)$ is strictly increasing in δ, with $q(\delta^*) = \min_\theta q_\theta^*$ and $e^{\delta^*} q(\delta^*) = \max_\theta q_\theta^*$.

Proposition 4 shows that the BPP constraint can be equivalently implemented by imposing a range for rejection rates $[q, e^\delta q]$ and requiring that each group's rejection rate q_θ lie within it. Groups with $q_\theta^* \in [q, e^\delta q]$ are unaffected, while those with $q_\theta^* < q$ must be rejected more often to prevent the "reject" decision from revealing membership in a low-credit-score group. Conversely, groups with $q_\theta^* > e^\delta q$ must be rejected less often.

When the privacy budget $\delta = 0$, rejection rates are equalized across groups, recovering the optimal quantile information structure in Strack and Yang (2024). As δ increases from 0 to δ^*, the admissible range $[q(\delta), e^\delta q(\delta)]$ expands in a nested way, interpolating between the privacy-preserving rule and the first-best. Thus, stricter privacy protection harms groups with initially low rejection rates and benefits those with high rejection rates. This mechanism is illustrated in Figure 5. Note that although this mechanism resembles affirmative

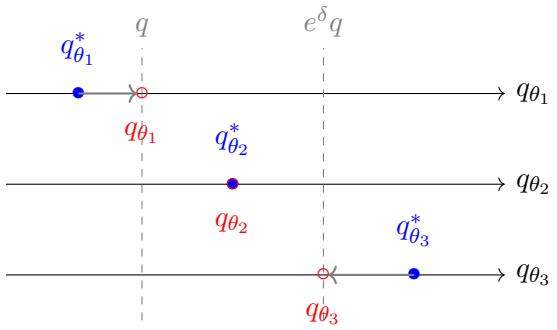

Figure 5: First-best rejection rates (q^*_θ, solid dots) are mapped to privacy-constrained rates (q_θ, open circles). Rejection rates outside $[q, e^\delta q]$ are clamped to the boundaries.

action in its effect, the two arise from distinct motivations: affirmative action aims to ensure fairness across groups, whereas here, the equalization of rejection rates is driven by privacy considerations.

5. Further Discussions of Worst-Case Privacy

In this section, I discuss several additional aspects of the worst-case framework. First, I provide a decision-theoretic interpretation of differential privacy. Second, I extend the representation results to information structures with arbitrary, possibly infinite, signal spaces. Third, I analyze the role of priors and consider how to address cases where the prior is unknown or contested.

5.1. Connection to Differential Privacy

Consider the case $\Theta = \times_{n=1}^{N} T_n$, where each T_n is interpreted as the type space of individual n. Therefore, each $\theta \in \Theta$ is a dataset. Recall that prediction privacy measures restrict an adversary's ability to predict each T_n, thereby aiming to keep each individual's type *secret*. A different notion—often described as protecting privacy as *control* (Dwork and Naor, 2010; Dwork, 2011)—takes a conditional perspective: it limits what can be inferred about an individual's type *given that the adversary already knows the types of all others*. Here, each individual retains control over her own data but not over others'. Therefore, the adversary may learn t_n from t_{-n}, but not directly from t_n itself.

Formally, let $\mu_0 \in \Delta(\Theta)$ be a full-support prior. Denote by $\mathcal{C}^{DP}_{\mu_0}$ the class of decision problems such that for some $n \in \{1, ..., N\}$, $c : T_n \to \mathbb{R}_+$, and $t_{-n} \in T_{-n}$, it holds that $A = T_n$ and

$$u(\theta, a) = \begin{cases} c(a), & \text{if } \theta_n = a, \theta_{-n} = t_{-n} \\ 0, & \text{otherwise} \end{cases}.$$

In words, an adversary in $\mathcal{C}^{DP}_{\mu_0}$ seeks to predict an individual's type t_n only when the others' types are fixed at t_{-n}. This captures the situation in which the adversary already knows the others' types t_{-n} and only the focal individual's data remains private. Denote $\theta \mathcal{N} \theta'$ if θ and θ' differ in exactly one coordinate. Following the same logic as in Theorem 2, we have:

Proposition 5. *A privacy measure L satisfies Worst-Case Protection, $\mathcal{C}^{DP}_{\mu_0}$-Monotonicity, and normalization if and only if*

$$L(x) = \max_{s \in S_x,\, \theta \mathcal{N} \theta'} d_{\theta,\theta'}(\ell_{\theta,\theta'}(x^s)), \quad \forall\, x \in \mathcal{X}_f, \tag{4}$$

for some increasing functions $d_{\theta,\theta'} : \overline{\mathbb{R}} \to \overline{\mathbb{R}}_+$ with $d_{\theta,\theta'}(0) = 0$.

Intuitively, for two datasets θ and θ' that differ by one individual's type (say individual n), $\ell_{\theta,\theta'}$ measures the informativeness about t_n in θ relative to t'_n in θ', conditional on knowing t_{-n}. When this log-likelihood ratio is small, the adversary learns little about t_n versus t'_n given t_{-n}.

In particular, when all weighting functions are chosen as identity functions—reflecting equal treatment—Equation 4 reduces to

$$L_{DP}(x) = \max_{s \in S_x,\, \theta \mathcal{N} \theta'} \ell_{\theta,\theta'}(x^s) = \max_{s \in S_x,\, \theta \mathcal{N} \theta'} \log \frac{x(s|\theta)}{x(s|\theta')},$$

which is precisely the definition of differential privacy (DP). Hence, Proposition 5 provides a decision-theoretic interpretation of DP.

It is instructive to compare DP and BPP. When $N = 1$, both coincide with LDP. For $N > 1$, however, they embody distinct notions: DP permits strong inferences about an individual's type t_n when these arise from correlations with others' data t_{-n}, but restricts inferences conditional on t_{-n}. By contrast, BPP restricts strong inferences after marginalizing over t_{-n}, yet allows a signal to reveal substantial information about t_n conditional on a particular t_{-n}. I highlight this contrast by revisiting the voting application below.

Voting under Differential Privacy. Consider the voting model from Subsection 4.2, but suppose the planner now faces a DP constraint:

$$\begin{aligned} \max_{x:\Theta \to \Delta(A)} \quad & \sum_{\theta,a} \Pr(\omega_a|\theta)\, x(a|\theta)\, \mu_0(\theta) \\ \text{s.t.} \quad & L_{DP}(x) = \max_{a \in A,\, \theta \mathcal{N} \theta'} \log \frac{x(a|\theta)}{x(a|\theta')} \;\leq\; \delta. \end{aligned} \tag{$\mathcal{V}_{DP}$}$$

The next result characterizes the optimal voting rule under the DP constraint.[28] Let v^*_{DP} denote the optimal value of Program ($\mathcal{V}_{DP}$).

Proposition 6. *The optimal solution (unique when $\delta > 0$) to Program ($\mathcal{V}_{DP}$) is*

$$x(a_1|\theta) = \frac{e^{-(n-m(\theta)-\frac{n-1}{2})\delta}}{1+e^{-\delta}} \quad \text{if } m(\theta) < \tfrac{n}{2}, \qquad x(a_0|\theta) = \frac{e^{-(m(\theta)-\frac{n-1}{2})\delta}}{1+e^{-\delta}} \quad \text{if } m(\theta) > \tfrac{n}{2}.$$

In particular, for any $\delta > 0$, $v^*_{DP}(\delta; n) \to 1$ as $n \to \infty$. If $\delta = 0$, then $v^*_{DP}(0; n) = 1/2$ for all n.

Unlike under BPP, when $\delta > 0$, DP always permits full aggregation as the electorate grows large. This is because DP allows indirect inference about a voter's signal via correlations with others. As the influence of each voter vanishes with n, the DP constraint becomes asymptotically non-binding, allowing the planner to approximate the simple majority rule and achieve first-best welfare.

By contrast, when $\delta = 0$, DP requires that no single voter can ever be pivotal, which forces the rule to be constant and therefore uninformative. Under BPP, however, variations are still possible: a voter may be pivotal (as in the cutoff rule characterized in Proposition 3), provided that the collective action does not reveal information about any individual's signal once others' signals are marginalized over.

5.2. Extension to Infinitely Supported Information Structures

I extend the main representation results to $\mathcal{X}$, the set of information structures with arbitrary support. In particular, a privacy measure is a function $L : \mathcal{X} \to \overline{\mathbb{R}}_+$. The following example shows why, without additional structure, L may fail to admit a max-separable representation as in Theorem 1.

Example 5. Let Θ be binary, so a posterior μ can be identified with a number in $[0, 1]$. Let the reference measure be uniform, $\lambda = 1/2$. Define a measure L on the space of posterior distributions τ (which must have mean $1/2$): set $L(\tau) = 1$ if τ has a point mass at $\mu = 1$, $L(\tau) = \frac{1}{2}$ if $\mu = 1$ is in the support of τ but not a point mass, and $L(\tau) = 0$ otherwise.

This measure satisfies Worst-Case Protection and Blackwell Monotonicity. Worst-Case Protection: For a mixture $\tau = \alpha\tau' + (1-\alpha)\tau''$, τ has a point mass at 1 if and only if at least one of τ' or τ'' does, and 1 is in the support of τ if and only if at least one of τ' or τ'' does. Thus, $L(\tau) = \max\{L(\tau'), L(\tau'')\}$. Blackwell Monotonicity: Suppose τ' is a mean-preserving

[28] The optimal rule under DP can also be obtained as a corollary of Theorem 3 in Schmutte and Yoder (2022). In Appendix C.4, I provide an alternative proof using the shadow prices of the privacy constraint, comparable to the argument in Proposition 3.

spread of τ. If τ' does not have a point mass at 1, neither does τ; if 1 is not in the support of τ', neither is it in the support of τ.

However, this measure cannot be represented as in Theorem 1. If a representation existed, from the representation for $\mathcal{X}_f$, the (belief-based) index function would have to be $\tilde{d}(\mu) = 1$ if $\mu = 1$ and $\tilde{d}(\mu) = 0$ otherwise. But this index function cannot represent posterior distributions in which 1 is in the support without having a point mass, such as the uniform distribution on $[0, 1]$. □

The failure in Example 5 arises from discontinuities that can occur when moving from finitely to infinitely supported information structures. For example, when belief $\mu = 1$ shifts from being merely in the support to having positive mass, privacy loss jumps discontinuously. To rule out such pathologies, I impose a continuity requirement on L.

Say a sequence of information structures x_n converges to x if $\tau^\lambda_{x_n}$ converges weakly to τ^λ_x.[29] The next axiom requires L to be lower-semicontinuous with respect to this convergence notion. It can be interpreted as a robustness requirement: if all information structures in a sequence satisfy some privacy standard (i.e., $L(x_n) \leq \delta$ for all n), then so does their limit.[30]

Axiom 3 (Lower-Semicontinuity)**.** *If $x_n \to x$, then $L(x) \leq \liminf_n L(x_n)$.*

For L defined on $\mathcal{X}$, I say L is a worst-case privacy measure if it satisfies Worst-Case Protection, Blackwell Monotonicity, and Lower-Semicontinuity. In Appendix A, I show that $L : \mathcal{X} \to \overline{\mathbb{R}}_+$ is a worst-case privacy measure if and only if it takes the form of representation (WCP), with the additional requirement that d is lower-semicontinuous. Moreover, Theorem 2 also extends with the additional requirements that the weighting functions d_{t_n,t'_n} are lower-semicontinuous. The index functions of both BPP and DP satisfy this requirement.

Remark 4. With the axiom of Lower-Semicontinuity, the privacy measure L can be equivalently defined through a total order on $\mathcal{X}$. Indeed, by Rader's representation theorem (Rader, 1963), any such order admits a lower-semicontinuous cardinal representation, and L can be defined as such a cardinal representation. Through this representation, the axioms of Worst-Case Protection and Blackwell Monotonicity can be equivalently phrased as axioms on the total order.

[29] The topology induced on $\mathcal{X}$ is independent of the (full-support) reference measure $\lambda \in \Delta(\Theta)$.

[30] Upper-semicontinuity, by contrast, conflicts with Worst-Case Protection. For example, if $x_n := \frac{1}{n}\hat{x} + (1 - \frac{1}{n})x_0$ where $\hat{x}$ is perfectly revealing and x_0 is uninformative, Worst-Case Protection requires $L(x_n) = L(\hat{x})$ while upper-semicontinuity requires $\limsup_n L(x_n) \leq L(x_0) = 0$. The only measure satisfying both is the trivial one.

5.3. Known vs. Unknown Prior

Another issue is whether a privacy measure should depend on the prior belief over states. The worst-case axiomatic framework is agnostic on this point, since its analysis can be carried out for each fixed prior μ_0. For this reason, the characterization in Theorem 1 accommodates both prior-independent privacy measures, such as DP, and prior-dependent ones, such as Prediction Privacy. The following gives another example of a prior-dependent worst-case privacy measure.

Example 6 (Ex Post Bayesian Privacy)**.** Ex post Bayesian privacy (Evfimievski, Gehrke, and Srikant, 2003; Eilat et al., 2021) is defined by $L(x) = \sup_{\mu \in \tau_x^{\mu_0}} \tilde{d}(\mu)$, where the belief-based index function is

$$\tilde{d}(\mu) = D_{KL}(\mu \| \mu_0) := \sum_{\theta} \mu(\theta) \log \frac{\mu(\theta)}{\mu_0(\theta)},$$

and D_{KL} is the Kullback–Leibler divergence, which satisfies quasi-convexity (and lower-semicontinuity). Different choices of μ_0 yield different measures L, so ex post Bayesian privacy is inherently prior-dependent: it evaluates the worst-case leakage relative to the specified prior. □

From a normative perspective, incorporating available prior information may improve the evaluation of an information structure. The following example illustrates.

Example 7. Consider the following information structure where the state is (ω, θ):

x	(ω_1, θ_1)	(ω_1, θ_2)	(ω_2, θ_1)	(ω_2, θ_2)
s_1	$\frac{1}{2}$	$\frac{1}{2}$	1	0
s_2	$\frac{1}{2}$	$\frac{1}{2}$	0	1

Suppose the goal is to protect θ. If $\mu_0(\omega_1) = 1$, the information structure is uninformative about θ and thus privacy-preserving. If $\mu_0(\omega_2) = 1$, the information structure is perfectly revealing and maximally harmful. If the conditional distributions $\mu_0(\cdot|\theta_i)$ are known, one can evaluate privacy loss by deriving the implied information structure for θ. Ignoring this prior information would lead to a poor assessment of the privacy loss. □

The challenge, therefore, is not whether to use prior information when it is available, but how to evaluate privacy when the prior is unknown or contested. The next result provides a simple solution.

Proposition 7. *Let Γ be a finite index set, and let $\mathcal{C}_\gamma$, $\gamma \in \Gamma$, denote classes of decision problems. Then L is a worst-case privacy measure that is $\cup_{\gamma\in\Gamma}\mathcal{C}_\gamma$-monotonic if and only if*

$$L = \sup_{\gamma\in\Gamma} L_\gamma,$$

where each L_γ is a worst-case privacy measure that is $\mathcal{C}_\gamma$-monotonic. Moreover, the "if" direction continues to hold when Γ is infinite.

Proposition 7 implies that taking the pointwise supremum of privacy measures defined under different priors yields a measure that protects uniformly across them. Formally, let $\{\mathcal{C}_{\mu_0}\}_{\mu_0\in\Delta(\Theta)}$ denote classes of decision problems that differ only in their priors. If each L_{μ_0} is a worst-case privacy measure that is $\mathcal{C}_{\mu_0}$-monotonic, then $L(x) := \sup_{\mu_0\in\Delta(\Theta)} L_{\mu_0}(x)$ is itself a worst-case privacy measure that is monotonic with respect to the same class of decision problems but with arbitrary priors. In practice, requiring $L \leq \delta$ is equivalent to requiring $L_{\mu_0} \leq \delta$ for all μ_0, thus providing uniform protection across priors.

6. Concluding Remarks

This paper develops an axiomatic framework for privacy measures grounded in economic principles. At its core is the axiom of Worst-Case Protection, motivated by modern privacy law, which requires that the privacy loss of an information structure equal the highest among its components. The resulting class of Worst-Case Privacy admits a max-separable representation over likelihood vectors. Within this workhorse model, I show that restricting attention to particular classes of adversarial decision problems yields concrete privacy measures. In particular, specializing to prediction problems leads to a pairwise structure that provides an economic foundation for standards such as local differential privacy and differential privacy. I also derive Prediction Privacy as a class that ensures the secrecy of all individuals' types in a dataset. These measures are then applied to study economic settings such as matching and voting.

In two-sided matching, I formalize a sharp welfare–privacy tradeoff: within a natural class of stable mechanisms, higher worker welfare necessarily entails greater privacy loss. In voting, I show that the optimal voting rule that protects voters' signals may require reversing the collective action when there is an extreme majority. By contrast, if privacy is understood "as control," the optimal voting rule is monotonic. These results highlight that different notions of privacy have distinct implications for the efficiency of information aggregation, underscoring that privacy concerns involve not only how much to protect, but also what kind of information to protect.

Next, I discuss several limitations of the current analysis and directions for future research.

Alternative Classes of Adversarial Problems. This paper focuses on prediction problems, where adversarial payoffs are binary (correct vs. incorrect) for a given action. Other adversarial objectives are also natural. For instance, an adversary may seek to reconstruct or estimate sensitive attributes, with payoffs determined by the distance between the estimate and the truth (Chatzikokolakis, Andrés, Bordenabe, and Palamidessi, 2013). In this case, $\mathcal{C}$ can be taken as the class of estimation problems, and it would be valuable to characterize the corresponding worst-case privacy measures that are $\mathcal{C}$-monotonic. More generally, the same machinery illustrated in Figure 2 can be applied to derive privacy measures suited to different applications, which I leave for future work.

Economic Applications with Incentive Constraints. The applications in this paper abstract from incentive constraints, but more generally, privacy often interacts with incentives. For example, Eilat et al. (2021) study Bayesian privacy in monopolistic screening, Strack and Yang (2025) analyze privacy-preserving personalized pricing, and Eilat et al. (2023) examine auctions with privacy concerns. A common feature of these settings is that the designer cannot freely choose any mechanism; she must either induce truth-telling from the data owner or ensure obedience from the receiver. In such cases, Theorem 3 still provides a certification tool for identifying optimal solutions. The central challenge is to understand how the shadow prices of privacy interact with incentive constraints. This remains an important direction for future research.

Cardinal Representations. Although I study a cardinal measure of privacy loss, its foundation is essentially ordinal: all properties imposed on L are ordinal in nature. Consequently, the cardinal representation is identified only up to an increasing transformation. Sharpening the representation requires additional structure. Natural candidates include composability (sub-additivity), which requires that the privacy loss of the composition of information structures not exceed the sum of their individual losses (Dwork and Roth, 2014; Bordoli and Iijima, 2025), and additivity, which requires equality in this relation (Pomatto et al., 2023). In practice, composability is particularly important: when a mechanism is designed in parts, it guarantees that the overall privacy loss can be aggregated consistently and does not blow up as the design scales. Exploring such refinements may help further operationalize Worst-Case Privacy.

Weakening Worst-Case Protection. Worst-Case Protection is a demanding axiom: it implies that privacy loss is determined by the signal with the highest loss index, no matter how unlikely it is. This ensures robustness—protection even in extreme scenarios—but comes at the cost of disregarding likelihoods that may matter in practice. By contrast, alternative approaches such as average-case weigh signals by their probabilities, yielding measures that are typically more responsive but less resilient to rare harmful events. A promising direction for future work is to explore intermediate notions that blend these perspectives. For example, one could relax the axiom so that signals contribute to overall loss according to a probability weighting function that interpolates between two extremes: treating all signals as equally important (worst-case) and weighting them exactly by their probabilities (average-case). Such relaxations would preserve the spirit of uniform protection while allowing for more nuanced tradeoffs between robustness and efficiency.

References

ACQUISTI, A., C. R. TAYLOR, AND L. WAGMAN (2016): "The Economics of Privacy," *Journal of Economic Literature*, 54, 442–492.

ADLER, E. S. AND T. E. HALL (2013): "Ballots, transparency, and democracy," *Election Law Journal*, 12, 146–161.

ALIPRANTIS, C. D. AND K. C. BORDER (2006): *Infinite dimensional analysis: a hitchhiker's guide*, Springer.

ALVIM, M. S., K. CHATZIKOKOLAKIS, A. MCIVER, C. MORGAN, C. PALAMIDESSI, AND G. SMITH (2016): "Axioms for information leakage," in *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, IEEE, 77–92.

ASHLAGI, I., Y. KANORIA, AND J. D. LESHNO (2017): "Unbalanced random matching markets: The stark effect of competition," *Journal of Political Economy*, 125, 69–98.

BEST, J., D. QUIGLEY, M. SAEEDI, AND A. SHOURIDEH (2025): "Divide or Confer: Aggregating Information without Verification," *Working Paper*.

BLACKWELL, D. (1953): "Equivalent comparisons of experiments," *The annals of mathematical statistics*, 265–272.

BORDOLI, D. AND R. IIJIMA (2025): "Convex Cost of Information via Statistical Divergence," *Working Paper*.

CALZOLARI, G. AND A. PAVAN (2006): "On the optimality of privacy in sequential contracting," *Journal of Economic theory*, 130, 168–204.

CAPLIN, A. AND M. DEAN (2015): "Revealed preference, rational inattention, and costly information acquisition," *American Economic Review*, 105, 2183–2203.

CHATZIKOKOLAKIS, K., M. E. ANDRÉS, N. E. BORDENABE, AND C. PALAMIDESSI (2013): "Broadening the scope of differential privacy using metrics," in *international symposium on privacy enhancing technologies symposium*, Springer, 82–102.

CHOI, J. P., D.-S. JEON, AND B.-C. KIM (2019): "Privacy and personal data collection with information externalities," *Journal of Public Economics*, 173, 113–124.

CHWE, M. S.-Y. (2010): "Anonymous Procedures for Condorcet's Model: Robustness, Nonmonotonicity, and Optimality," *Quarterly Journal of Political Science*, 5, 45–70.

CONGRESSIONAL RESEARCH SERVICE (2024): "AI Deepfakes and the 2024 Elections," Tech. Rep. IN12389, Congressional Research Service, accessed: 2025-09-03.

DE OLIVEIRA, H., T. DENTI, M. MIHM, AND K. OZBEK (2017): "Rationally inattentive preferences and hidden information costs," *Theoretical Economics*, 12, 621–654.

DWORK, C. (2011): "A firm foundation for private data analysis," *Communications of the ACM*, 54, 86–95.

DWORK, C., F. MCSHERRY, K. NISSIM, AND A. SMITH (2006): "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*, Springer, 265–284.

DWORK, C. AND M. NAOR (2010): "On the difficulties of disclosure prevention in statistical databases or the case for differential privacy," *Journal of Privacy and Confidentiality*, 2.

DWORK, C. AND A. ROTH (2014): "The algorithmic foundations of differential privacy," *Foundations and trends® in theoretical computer science*, 9, 211–407.

EILAT, R., K. ELIAZ, AND X. MU (2021): "Bayesian privacy," *Theoretical Economics*, 16, 1557–1603.

——— (2023): "Privacy Preserving Auctions," Tech. rep., CEPR Discussion Papers.

EVFIMIEVSKI, A., J. GEHRKE, AND R. SRIKANT (2003): "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 211–222.

GALPERTI, S., T. LIU, AND J. PEREGO (2024): "Competitive markets for personal data," in *ACM EC*, ACM Economics and Computation.

GALPERTI, S. AND J. PEREGO (2023): "Privacy and the Value of Data," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 113, 197–203.

GILBOA-FREEDMAN, G. AND R. SMORODINSKY (2020): "On the behavioral implications of differential privacy," *Theoretical Computer Science*, 841, 84–93.

GOLDFARB, A. AND C. TUCKER (2024): *The Economics of Privacy*, University of Chicago Press.

HAUPT, A. AND Z. HITZIG (2021): "Contextually private mechanisms," *arXiv preprint arXiv:2112.10812*.

HE, K., F. SANDOMIRSKIY, AND O. TAMUZ (2021): "Private private information," *arXiv preprint arXiv:2112.14356*.

HIDIR, S. AND N. VELLODI (2021): "Privacy, personalization, and price discrimination," *Journal of the European Economic Association*, 19, 1342–1363.

ICHIHASHI, S. (2020): "Online privacy and information disclosure by consumers," *American Economic Review*, 110, 569–595.

KASIVISWANATHAN, S. P., H. K. LEE, K. NISSIM, S. RASKHODNIKOVA, AND A. SMITH (2011): "What can we learn privately?" *SIAM Journal on Computing*, 40, 793–826.

KATTWINKEL, D. AND A. WINTER (2024): "Optimal Decision Mechanisms for Committees: Acquitting the Guilty," *arXiv preprint arXiv:2407.07293*.

KEYSSAR, A. (2009): *The right to vote: The contested history of democracy in the United States*, Basic Books (AZ).

KIFER, D. AND A. MACHANAVAJJHALA (2014): "Pufferfish: A framework for mathematical privacy definitions," *ACM Transactions on Database Systems (TODS)*, 39, 1–36.

KURIWAKI, S., J. B. LEWIS, AND M. MORSE (2025): "Privacy violations in election results," *Science Advances*, 11, eadt1512.

MARES, I. (2015): *From open secrets to secret voting: Democratic electoral reforms and voter autonomy*, Cambridge University Press.

PAI, M. M. AND A. ROTH (2013): "Privacy and mechanism design," *ACM SIGecom Exchanges*, 12, 8–29.

PITTEL, B. (1989): "The average number of stable matchings," *SIAM Journal on Discrete Mathematics*, 2, 530–549.

POMATTO, L., P. STRACK, AND O. TAMUZ (2023): "The cost of information: The case of constant marginal costs," *American Economic Review*, 113, 1360–1393.

RADER, T. (1963): "The existence of a utility function to represent preferences," *The Review of Economic Studies*, 30, 229–232.

ROCKAFELLAR, R. T. (1997): *Convex analysis*, vol. 28, Princeton university press.

ROTH, A. E. (1986): "On the allocation of residents to rural hospitals: a general property of two-sided matching markets," *Econometrica: Journal of the Econometric Society*, 425–427.

——— (2018): "Marketplaces, markets, and market design," *American Economic Review*, 108, 1609–1658.

ROTH, A. E. AND M. A. O. SOTOMAYOR (1990): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, vol. 18 of *Econometric Society Monographs*, Cambridge: Cambridge University Press.

SCHMUTTE, I. M. AND N. YODER (2022): "Information Design for Differential Privacy," in *Proceedings of the 23rd ACM Conference on Economics and Computation*, 1142–1143.

STRACK, P. AND K. H. YANG (2024): "Privacy-Preserving Signals," *Econometrica*, 92, 1907–1938.

——— (2025): "Non-Discriminatory Personalized Pricing," *arXiv preprint arXiv:2506.20925*.

SU, W. J. (2024): "A Statistical Viewpoint on Differential Privacy: Hypothesis Testing, Representation, and Blackwell's Theorem," *Annual Review of Statistics and Its Application*, 12.

SWEENEY, L. (2002): "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, 10, 557–570.

TAYLOR, C. R. (2004): "Consumer privacy and the market for customer information," *RAND Journal of Economics*, 631–650.

YEOM, S., I. GIACOMELLI, M. FREDRIKSON, AND S. JHA (2018): "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*, IEEE, 268–282.

ZARRABIAN, M. A. AND P. SADEGHI (2025): "An extension of the adversarial threat model in quantitative information flow," in *2025 IEEE 38th Computer Security Foundations Symposium (CSF)*, IEEE, 585–600.

Appendices

A. Proofs of Section 2

In this section, I consider the general case where $L : \mathcal{X} \to \overline{\mathbb{R}}_+$ is defined on the set of all information structures. For L defined on $\mathcal{X}_f$, all proofs are analogous, except that L and d may not be lower-semicontinuous (l.s.c.). I will comment in "Remark" environments on how the statement and proof change after each result.

I first formally establish that the likelihood-vector-based and belief-based representations are equivalent. Fix a full-support reference measure $\lambda \in \Delta(\Theta)$. Let D be the set of functions $d : \mathbb{R}_+^K \to \overline{\mathbb{R}}_+$ that are homogeneous of degree 0 and satisfy $d(\mathbf{0}) = 0$. Let $\tilde{D}$ be the set of functions $\tilde{d} : \Delta(\Theta) \to \overline{\mathbb{R}}_+$. Define the mapping $\Phi : D \to \tilde{D}$ such that for any $d \in D$, $\tilde{d} = \Phi(d)$ is given by:

$$\tilde{d}(\mu) := d\left(\left(\frac{\mu(\theta_k)}{\lambda(\theta_k)}\right)_{k=1}^K\right),$$

for all $\mu \in \Delta(\Theta)$. It is easy to see that this mapping is a bijection.

The following lemma establishes that this bijection preserves the key properties required by the representation theorems.

Lemma A.1 (Equivalence of Representations)**.** *Let $d \in D$ and $\tilde{d} \in \tilde{D}$ be such that $\tilde{d} = \Phi(d)$. Then d is quasi-convex and lower-semicontinuous with $d(c) = 0$ for all constant vectors $c \in \mathbb{R}_+^K$ if and only if $\tilde{d}$ is quasi-convex and lower-semicontinuous with $\tilde{d}(\lambda) = 0$. Moreover, if these conditions hold, then for any information structure $x \in \mathcal{X}$, we have:*

$$\sup_{\hat{S} \in \mathcal{S}} d(x(\hat{S}|\cdot)) = \sup_{\mu \in \operatorname{Supp} \tau_x^\lambda} \tilde{d}(\mu).$$

Proof. The proof proceeds in two parts. First, we show the equivalence of the properties of the functions, and second, we prove the equality of the representations.

(Part 1: Equivalence of Properties) We first show the "only if" direction. Assume d is quasi-convex, lower-semicontinuous, and $d(c) = 0$ for constant vectors c.

- ($\tilde{d}$ is quasi-convex): Let $\mu', \mu'' \in \Delta(\Theta)$ and $\alpha \in (0,1)$. Let $\mu = \alpha\mu' + (1-\alpha)\mu''$. Let ν', ν'', ν be the likelihood vectors corresponding to μ', μ'', μ (e.g., $\nu_k = \mu_k/\lambda_k$). Then ν is a convex combination of ν' and ν''. Since d is quasi-convex, $d(\nu) \leq \max\{d(\nu'), d(\nu'')\}$. By definition, this implies $\tilde{d}(\mu) \leq \max\{\tilde{d}(\mu'), \tilde{d}(\mu'')\}$.
- ($\tilde{d}$ is lower-semicontinuous): The mapping $\mu \mapsto (\mu_k/\lambda_k)_k$ is continuous. Since d is lower-semicontinuous, their composition $\tilde{d}$ is also lower-semicontinuous.

- ($\tilde{d}(\lambda) = 0$): $\tilde{d}(\lambda) = d((\lambda_k/\lambda_k)_k) = d(\mathbf{1}) = 0$.

Next, we show the "if" direction. Assume $\tilde{d}$ is quasi-convex, lower-semicontinuous, and $\tilde{d}(\lambda) = 0$. For a non-zero likelihood vector $\nu \in \mathbb{R}^K_+$, let the corresponding posterior be $\mu(\nu) \in \Delta(\Theta)$, with components $\mu_k(\nu) = \frac{\nu_k \lambda_k}{\sum_{j=1}^K \nu_j \lambda_j}$. Let $d = \Phi^{-1}(\tilde{d})$, where $d(\nu) = \tilde{d}(\mu(\nu))$ if $\nu \neq \mathbf{0}$ and $d(\mathbf{0}) = 0$.

- (d is quasi-convex): Let $\nu', \nu'' \in \mathbb{R}^K_+$ and $\alpha \in (0,1)$. Let $\nu = \alpha\nu' + (1-\alpha)\nu''$. We want to show $d(\nu) \leq \max\{d(\nu'), d(\nu'')\}$. If either $\nu' = \mathbf{0}$ or $\nu'' = \mathbf{0}$, the result follows from homogeneity of degree 0. Now assume $\nu', \nu'' \neq \mathbf{0}$. The posterior $\mu(\nu)$ has components

$$\mu_k(\nu) = \frac{(\alpha\nu'_k + (1-\alpha)\nu''_k)\lambda_k}{\alpha\sum_j \nu'_j\lambda_j + (1-\alpha)\sum_j \nu''_j\lambda_j} = \beta\frac{\nu'_k\lambda_k}{\sum_j \nu'_j\lambda_j} + (1-\beta)\frac{\nu''_k\lambda_k}{\sum_j \nu''_j\lambda_j},$$

  where $\beta = \frac{\alpha\sum_j \nu'_j\lambda_j}{\alpha\sum_j \nu'_j\lambda_j + (1-\alpha)\sum_j \nu''_j\lambda_j} \in [0,1]$. Thus, $\mu(\nu) = \beta\mu(\nu') + (1-\beta)\mu(\nu'')$. Since $\tilde{d}$ is quasi-convex, we have

$$d(\nu) = \tilde{d}(\mu(\nu)) \leq \max\{\tilde{d}(\mu(\nu')), \tilde{d}(\mu(\nu''))\} = \max\{d(\nu'), d(\nu'')\}.$$

- (d is lower-semicontinuous): The mapping $\nu \mapsto \mu(\nu)$ is continuous for $\nu \neq \mathbf{0}$. Since $\tilde{d}$ is lower-semicontinuous, the composition $d(\nu) = \tilde{d}(\mu(\nu))$ is lower-semicontinuous on $\mathbb{R}^K_+ \setminus \{\mathbf{0}\}$. At $\nu = \mathbf{0}$, we have $d(\mathbf{0}) = 0$. Since $d(\nu) \geq 0$ for all ν, d is lower-semicontinuous at $\mathbf{0}$ as well.

- ($d(c) = 0$ for constant vectors): For a constant vector $\nu_k = c \geq 0$ for all k. If $c > 0$, the corresponding posterior is $\mu_k(c) = \frac{c\lambda_k}{\sum_j c\lambda_j} = \lambda_k$. Thus, $d(c) = \tilde{d}(\lambda) = 0$. If $c = 0$, ν is the zero vector, and $d(\mathbf{0}) = 0$ by definition.

(Part 2: Equality of Representations) Let $\tilde{d} = \Phi(d)$. Note for $\hat{S} \in \mathcal{S}$ such that $x(\hat{S}|\cdot) \neq \mathbf{0}$, by definition we have $d(x(\hat{S}|\cdot)) = \tilde{d}(\mu_{\hat{S}})$, where $\mu_{\hat{S}}$ is the posterior belief conditional on $\hat{S}$, i.e., $\mu_{\hat{S},k} = \frac{x(\hat{S}|\theta_k)\lambda_k}{\sum_j x(\hat{S}|\theta_j)\lambda_j}$. Let P be the distribution on $\Theta \times S$ by x and λ.

(Step 1) Show $\sup_{\mu \in \operatorname{Supp}\tau_x^\lambda} \tilde{d}(\mu) \leq \sup_{\hat{S}\in\mathcal{S}} d(x(\hat{S}|\cdot))$. Take any $\mu \in \operatorname{Supp}\tau_x^\lambda$. By definition of the support, there exists a sequence of sets of signals S_n with $P(S_n) > 0$ such that the corresponding posteriors $\mu_{S_n} \to \mu$. Since $\tilde{d}$ is lower-semicontinuous,

$$\tilde{d}(\mu) \leq \liminf_n \tilde{d}(\mu_{S_n}) = \liminf_n d(x(S_n|\cdot)) \leq \sup_{\hat{S}\in\mathcal{S}} d(x(\hat{S}|\cdot)).$$

Since this holds for any $\mu \in \operatorname{Supp}\tau_x^\lambda$, the inequality follows.

(Step 2) Show $\sup_{\hat{S}\in\mathcal{S}} d(x(\hat{S}|\cdot)) \leq \sup_{\mu\in\operatorname{Supp}\tau_x^\lambda} \tilde{d}(\mu)$. Take any $\hat{S} \in \mathcal{S}$ with $P(\hat{S}) > 0$. Let

$(\mu_s)_{s\in S}$ denote the regular conditional probabilities, which is a random vector defined on S. Then we have $\mu_{\hat{S}} = \mathbb{E}[\mu_s | s \in \hat{S}]$. Since $\tilde{d}$ is quasi-convex, we have $\tilde{d}(\mathbb{E}[\mu_s|s \in \hat{S}]) \leq \sup_{\mu\in \operatorname{Supp} \tau_x^\lambda} \tilde{d}(\mu)$.[31] Combining these gives

$$d(x(\hat{S}|\cdot)) = \tilde{d}(\mu_{\hat{S}}) \leq \sup_{\mu\in \operatorname{Supp} \tau_x^\lambda} \tilde{d}(\mu).$$

As this holds for any $\hat{S}$, the inequality follows. *Q.E.D.*

Remark. *In Part 1, d is lower-semicontinuous if and only if $\tilde{d}$ is, so this assumption can be simultaneously removed. In Part 2, lower-semicontinuity is only used in Step 1. When only $\mathcal{X}_f$ is concerned, it is sufficient to consider each signal in S_x without the need to take a sequence of sets S_n, so lower-semicontinuity is not needed.*

A.1. Proof of Theorem 1

In light of Lemma A.1, I prove the belief-based representation only; I use d to denote a belief-based index function as well. Fix a reference measure $\lambda \in \Delta(\Theta)$ with full support. We denote by $\mathcal{T} \subset \Delta(\Delta(\Theta))$ the set of posterior distributions with mean λ. Since Axiom 2 implies that the posterior distribution is a sufficient statistic for the privacy loss, we abuse notation and for any $\tau \in \mathcal{T}$, we use $L(\tau)$ to denote the privacy loss of any information structure x with $\tau_x^\lambda = \tau$.

Proof of Theorem 1: "If" direction. We show that if a function L has the representation given in the theorem, it satisfies the three axioms and the normalization.

(Axiom 1: Worst-Case Protection) Let $x = \alpha x' + (1-\alpha)x''$ for some $x', x'' \in \mathcal{X}$ with $x' \perp x''$ and $\alpha \in (0,1)$. The induced posterior distribution is $\tau_x = \alpha\tau_{x'} + (1-\alpha)\tau_{x''}$, which implies $\operatorname{Supp} \tau_x = \operatorname{Supp} \tau_{x'} \cup \operatorname{Supp} \tau_{x''}$. Therefore,

$$L(x) = \sup_{\mu\in \operatorname{Supp} \tau_x} d(\mu) = \max\left\{ \sup_{\mu'\in \operatorname{Supp} \tau_{x'}} d(\mu'), \sup_{\mu''\in \operatorname{Supp} \tau_{x''}} d(\mu'') \right\} = \max\{L(x'), L(x'')\}.$$

(Axiom 2: Blackwell Monotonicity) Let x be Blackwell more informative than x'. By Blackwell's theorem (Blackwell (1953)), this implies that τ_x is a mean-preserving spread of $\tau_{x'}$, and therefore $\operatorname{Supp} \tau_{x'} \subseteq \operatorname{conv}(\operatorname{Supp} \tau_x)$. Let $\delta := L(x)$. By definition, $d(\mu) \leq \delta$ for

[31] This is based on the following observation: if $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex and X is an integrable random vector, then $f(\mathbb{E}[X]) \leq \sup_{x\in \operatorname{Supp} X} f(x)$. To see this, let $M := \sup_{x\in \operatorname{Supp} X} f(x)$. Since f is quasi-convex, $A := \{x \in \mathbb{R}^n : f(x) \leq M\}$ is convex. Since $\mathbb{E}[X]$ is in the convex hull of the support of X, we conclude $\mathbb{E}[X] \in A$.

all $\mu \in \operatorname{Supp} \tau_x$. Since d is quasi-convex, the set $A := \{\mu \in \Delta\Theta : d(\mu) \leq \delta\}$ is convex. As $\operatorname{Supp} \tau_x \subseteq A$, it follows that $\operatorname{conv}(\operatorname{Supp} \tau_x) \subseteq A$. Because $\operatorname{Supp} \tau_{x'} \subseteq \operatorname{conv}(\operatorname{Supp} \tau_x)$, we have $\operatorname{Supp} \tau_{x'} \subseteq A$. This means $d(\mu') \leq \delta$ for all $\mu' \in \operatorname{Supp} \tau_{x'}$. Taking the supremum, we get $L(x') \leq \delta = L(x)$.

(Axiom 3: Lower-Semicontinuity) By Theorem 17.14 of Aliprantis and Border (2006), the support correspondence $\tau \mapsto \operatorname{Supp} \tau$ is lower hemicontinuous. Since d is l.s.c., by Berge's theorem (Lemma 17.29 of Aliprantis and Border (2006)), L satisfies Lower-Semicontinuity.

(Normalization) The range of L is $\overline{\mathbb{R}}_+$ since the range of d is $\overline{\mathbb{R}}_+$. For an uninformative experiment x, the posterior is always the prior λ, so $L(x) = d(\lambda) = 0$. This completes the proof. *Q.E.D.*

Remark. *Proof of Axiom 1 does not rely on the l.s.c. of d. Proof of Axiom 2 relies on l.s.c. d because A needs to be closed. However, when $x, x' \in \mathcal{X}_f$, closeness of A is not required, so l.s.c. of d is not needed in that case.*

Let $\mathcal{X}_c \subset \mathcal{X}$ be the set of information structures whose induced posterior distributions are countably supported. Similarly, let $\mathcal{T}_c$ and $\mathcal{T}_f$ be the sets of countably and finitely supported posterior distributions with mean λ.

The proof proceeds in two steps. I first establish a key technical lemma, which allows us to extend results proven for finitely supported experiments to all experiments. I then use this lemma to prove the main result.

Lemma A.2. *Suppose L satisfies Axiom 2 and Axiom 3. If the representation $L(x) = \sup_{\mu \in \operatorname{Supp} \tau_x} d(\mu)$ holds for all $x \in \mathcal{X}_f$ for some quasi-convex and lower-semicontinuous function d, then the equality also holds for all $x \in \mathcal{X}$.*

Proof. Let $\tilde{\mu}_x$ be the random posterior generated by x, with distribution τ_x. Let $\{\Sigma_n\}$ be an increasing sequence of finite partitions of $\Delta\Theta$ that generates the Borel σ-algebra.[32] Define the random variable $\tilde{\mu}_n := \mathbb{E}[\tilde{\mu}_x | \Sigma_n]$ and let τ_n be its distribution. We observe: (i) Since each Σ_n is finite, each τ_n is finitely supported. (ii) By Lévy's Upward Convergence Theorem, $\tilde{\mu}_n \to \tilde{\mu}_x$ almost surely, which implies weak convergence $\tau_n \to \tau_x$. By Axiom 3, this means $L(x) \leq \liminf_n L(\tau_n)$. (iii) Since each τ_n is a mean-preserving contraction of τ_x, Axiom 2 implies $L(\tau_n) \leq L(x)$ for all n. Combining (ii) and (iii) yields $L(x) = \lim_n L(\tau_n)$.

We now show that $\lim_n L(\tau_n) = \sup_{\mu \in \operatorname{Supp} \tau_x} d(\mu)$. First, since τ_n is a mean-preserving contraction of τ_x, we have $\operatorname{Supp} \tau_n \subset \operatorname{conv}(\operatorname{Supp} \tau_x)$ for all n. Since d is quasi-convex,

[32] There exists such a sequence because the Borel sigma-algebra $\mathcal{B}(\Delta\Theta)$ is countably generated. Specifically, suppose $\mathcal{B}(\Delta\Theta) = \sigma(A_1, A_2, ...)$. Then $\Sigma_n := \sigma(A_1, ..., A_n)$ forms such a sequence.

for any $\mu_n \in \operatorname{Supp}\tau_n$, $d(\mu_n) \leq \sup_{\mu\in\operatorname{Supp}\tau_x} d(\mu)$. As τ_n is finitely supported, $L(\tau_n) = \max_{\mu_n\in\operatorname{Supp}\tau_n} d(\mu_n) \leq \sup_{\mu\in\operatorname{Supp}\tau_x} d(\mu)$. Taking the limit gives $\lim_n L(\tau_n) \leq \sup_{\mu\in\operatorname{Supp}\tau_x} d(\mu)$.

For the other direction, since $\tilde{\mu}_n \to \tilde{\mu}_x$ almost surely, for every $\mu \in \operatorname{Supp}\tau_x$, there exists a sequence $(\mu_n)_{n=1}^{\infty}$ such that $\mu_n \in \operatorname{Supp}\tau_n$ and $\mu_n \to \mu$.[33] Since d is lower-semicontinuous, $d(\mu) \leq \liminf_n d(\mu_n)$. Moreover, since each τ_n is finitely supported, $d(\mu_n) \leq L(\tau_n)$. Thus, $d(\mu) \leq \liminf_n L(\tau_n)$. As this holds for any $\mu \in \operatorname{Supp}\tau_x$, we have $\sup_{\mu\in\operatorname{Supp}\tau_x} d(\mu) \leq \lim_n L(\tau_n)$.

Combining the two directions, we conclude that $L(x) = \lim_n L(\tau_n) = \sup_{\mu\in\operatorname{Supp}\tau_x} d(\mu)$. *Q.E.D.*

Remark. *This lemma is not needed if only $\mathcal{X}_f$ is concerned.*

With Lemma A.2, it is sufficient to prove the "only if" direction for experiments in $\mathcal{X}_f$.

Proof of Theorem 1: "Only if" direction. The proof proceeds in three main steps. First, we define a candidate index function d and show that it provides the desired representation for all experiments that induce finitely many posteriors. Second, we show that this function d has the required properties of quasi-convexity and lower-semicontinuity. Finally, we invoke Lemma A.2 to extend the representation to all experiments.

For any $\mu \in \Delta\Theta$, define the candidate index function d as:

$$d(\mu) := \inf_{\tau\in\mathcal{T}_c:\mu\in\operatorname{Supp}\tau} L(\tau). \tag{5}$$

By definition, $d(\mu) \geq 0$ for all μ, and $d(\lambda) = 0$ since an uninformative experiment has zero loss.

(Step 1: Representation for Finitely Supported Experiments) Let $x \in \mathcal{X}_f$ be an experiment with a finitely supported posterior distribution τ_x. By the definition of d, for any $\mu \in \operatorname{Supp}\tau_x$, we have $d(\mu) \leq L(\tau_x)$. Taking the maximum over the finite support gives

$$\max_{\mu\in\operatorname{Supp}\tau_x} d(\mu) \leq L(x).$$

For the converse, let $\operatorname{Supp}\tau_x = \{\mu_1, ..., \mu_J\}$. For any $\varepsilon > 0$ and each $j \in \{1, ..., J\}$, by the definition of d, there exists an experiment $x_j \in \mathcal{X}_c$ with posterior distribution τ_j such that

[33] To see this, we note that for every $\varepsilon, r > 0$, there exists N such that $\Pr(\{\|\tilde{\mu}_n - \tilde{\mu}_x\| < \varepsilon\} \cap \{\tilde{\mu}_x \in B_r(\mu)\}) > 0$ for all $n \geq N$. By the triangle inequality, we have $\Pr(\tilde{\mu}_n \in B_{r+\varepsilon}(\mu)) > 0$ for all $n \geq N$. The sequence is then constructed by taking $\varepsilon, r \to 0$.

$\mu_j \in \operatorname{Supp} \tau_j$ and $L(\tau_j) \leq d(\mu_j) + \varepsilon$. Define a new posterior distribution $\tilde{\tau}$ as the mixture

$$\tilde{\tau} := \sum_{j=1}^{J} \frac{\tau_x(\mu_j)/\tau_j(\mu_j)}{\sum_{j'=1}^{J} \tau_x(\mu_{j'})/\tau_{j'}(\mu_{j'})} \tau_j.$$

By Axiom 1, $L(\tilde{\tau}) = \max_j L(\tau_j) \leq \max_j d(\mu_j) + \varepsilon$. By construction, we can represent $\tilde{\tau}$ as

$$\tilde{\tau} = \frac{1}{\sum_{j'=1}^{J} \tau_x(\mu_{j'})/\tau_{j'}(\mu_{j'})} \tau_x + \left(1 - \frac{1}{\sum_{j'=1}^{J} \tau_x(\mu_{j'})/\tau_{j'}(\mu_{j'})}\right) \hat{\tau}$$

for some $\hat{\tau} \in \mathcal{T}_c$. Again by Axiom 1, this implies $L(\tilde{\tau}) \geq L(\tau_x)$. Combining these gives $L(x) \leq \max_j d(\mu_j) + \varepsilon$. Since ε was arbitrary, we have $L(x) \leq \max_j d(\mu_j)$.

(Step 2: Properties of the Index Function)

Quasi-Convexity. Let $\mu', \mu'' \in \Delta\Theta$ with $\mu' \neq \mu''$, and let $\mu = \alpha\mu' + (1-\alpha)\mu''$ for some $\alpha \in (0,1)$. For any $\varepsilon > 0$, choose $\tau', \tau'' \in \mathcal{T}_c$ such that $\mu' \in \operatorname{Supp} \tau'$, $\mu'' \in \operatorname{Supp} \tau''$, and $L(\tau') \leq d(\mu') + \varepsilon$, $L(\tau'') \leq d(\mu'') + \varepsilon$. Let τ be the mixture

$$\tau := \frac{\alpha/\tau'(\mu')}{\alpha/\tau'(\mu') + (1-\alpha)/\tau''(\mu'')} \tau' + \frac{(1-\alpha)/\tau''(\mu'')}{\alpha/\tau'(\mu') + (1-\alpha)/\tau''(\mu'')} \tau''.$$

By Axiom 1, $L(\tau) = \max\{L(\tau'), L(\tau'')\}$. Now, define a new distribution $\tilde{\tau} \in \mathcal{T}_c$ by taking the mass from μ' and μ'' in τ and placing it on their average, μ. Formally, $\tilde{\tau}(\mu) = \tau(\mu') + \tau(\mu'')$, $\tilde{\tau}(\mu') = \tilde{\tau}(\mu'') = 0$, and $\tilde{\tau}(\tilde{\mu}) = \tau(\tilde{\mu})$ for all other $\tilde{\mu}$. Since μ is a convex combination of μ' and μ'', $\tilde{\tau}$ is a mean-preserving contraction of τ. By Axiom 2, $L(\tilde{\tau}) \leq L(\tau)$. By construction, $\mu \in \operatorname{Supp} \tilde{\tau}$, so by the definition of d, $d(\mu) \leq L(\tilde{\tau})$. Combining these inequalities yields:

$$d(\mu) \leq L(\tilde{\tau}) \leq L(\tau) = \max\{L(\tau'), L(\tau'')\} \leq \max\{d(\mu'), d(\mu'')\} + \varepsilon.$$

Since this holds for any $\varepsilon > 0$, we have $d(\mu) \leq \max\{d(\mu'), d(\mu'')\}$.

Lower-Semicontinuity. Let $\mu_n \to \mu$. Let $\tilde{d} := \liminf_n d(\mu_n)$. We want to show $d(\mu) \leq \tilde{d}$. Without loss, assume $\tilde{d} < \infty$. For any $\varepsilon > 0$, there exists a subsequence, which we re-index by n, such that $d(\mu_n) \leq \tilde{d} + \varepsilon/2$ for all n. For each n, by definition of d, there exists $\tau_n \in \mathcal{T}_c$ such that $\mu_n \in \operatorname{Supp} \tau_n$ and $L(\tau_n) \leq d(\mu_n) + \varepsilon/2 \leq \tilde{d} + \varepsilon$.

Now, take any sequence $\{\alpha_n\}$ with $\alpha_n > 0$ and $\sum_{n=1}^{\infty} \alpha_n = 1$. Let

$$\tau := \sum_{n=1}^{\infty} \alpha_n \tau_n, \;\; \tilde{\tau}_m := \sum_{n=1}^{m} \frac{\alpha_n}{\sum_{n'=1}^{m} \alpha_{n'}} \tau_n.$$

Since $\tau_n \in \mathcal{T}_c$ for every n, we also have $\tau, \tilde{\tau}_m \in \mathcal{T}_c$. Moreover, by Axiom 1, we know that $L(\tilde{\tau}_m) = \max_{n=1,...,m} L(\tau_n) \leq \tilde{d} + \varepsilon$ for all m. By construction, $\tilde{\tau}_m$ converges pointwise to τ on $\Delta\Theta$ and $\tilde{\tau}_m \leq \frac{1}{\alpha_1}\tau$ for all m. As a result, $\tilde{\tau}_m$ weakly converges to τ.[34] By Axiom 3, $L(\tau) \leq \liminf_m L(\tilde{\tau}_m) \leq \tilde{d} + \varepsilon$. Moreover, note that by construction $\mu \in \operatorname{Supp}\tau$. Therefore, by definition of d we have $d(\mu) \leq L(\tau)$, which implies that $d(\mu) \leq \tilde{d} + \varepsilon$. Since ε was arbitrary, we conclude that $d(\mu) \leq \tilde{d}$.

(Step 3: Extension to All Experiments) Since we have shown that the function d defined in (5) is quasi-convex and lower-semicontinuous, and that the representation holds for all $x \in \mathcal{X}_f$, Lemma A.2 implies that the representation holds for all $x \in \mathcal{X}$. This completes the proof. *Q.E.D.*

Remark. *When only $\mathcal{X}_f$ is considered, d is defined by ranging over $\tau \in \mathcal{T}_f : \mu \in \operatorname{Supp}\tau$ in Equation 5. Only Step 1 and the first part of Step 2 are needed, with $\mathcal{T}_c$ replaced by $\mathcal{T}_f$.*

A.2. Proof of Theorem 2

In this section, I focus on a general setup for prediction problems that subsumes Theorem 2 and Proposition 5. The setup contains the following elements:

- Let $(\mathcal{H}_n)_{n=1}^N$ be a family such that each $\mathcal{H}_n$ contains disjoint nonempty subsets of Θ. These capture the "aspects" that were captured by T_n's in the main text.
- We fix a prior $\mu_0 \in \Delta(\Theta)$ such that μ_0 has full support on each $\mathcal{H}_n$: $\mu_0(H_n^i) > 0$ for all $H_n^i \in \mathcal{H}_n$ and all n. Each H_n^i is called a *category*.
- The class of prediction problems $\mathcal{C}_{\mu_0}^P$ contains all decision problems (μ_0, A, u) such that for some $n \in \{1, ..., N\}$ and $c : \mathcal{H}_n \to \mathbb{R}_+$, it holds that $A = \mathcal{H}_n$ and

$$u(\theta, a) = \begin{cases} c(a), & \text{if } \theta_n \in a \\ 0, & \text{otherwise} \end{cases}.$$

I show the following result:

Theorem A.1. *A privacy measure L satisfies Worst-Case Protection, $\mathcal{C}_{\mu_0}^P$-Monotonicity, (Lower-Semicontinuity), and normalization if and only if*

$$L(x) = \max_{s \in S_x,\, 1 \leq n \leq N,\, H_n, H_n' \in \mathcal{H}_n} d_{H_n, H_n'}\left(\ell_{H_n, H_n'}(x^s)\right), \quad \forall x \in \mathcal{X}_f,$$

[34] To see this, take any bounded function f on $\Delta\Theta$ and note that $\int f(\mu) d\tau(\mu) = \sum_{\mu \in \operatorname{Supp}\tau} f(\mu)\tau(\mu)$. Since $\tilde{\tau}_m$ converges to τ pointwise and each $\tilde{\tau}_m$ is dominated by $\frac{1}{\alpha_1}\tau$, by the dominated convergence theorem we conclude that the integrals also converge.

for some functions $d_{H_n,H_n'} : \overline{\mathbb{R}} \to \overline{\mathbb{R}}_+$ *where each* $d_{H_n,H_n'}$ *is increasing, (l.s.c.), with* $d_{H_n,H_n'}(0) = 0$.

In what follows, I index a particular element of $\mathcal{H}_n$ by H_n^i. For brevity, I slightly abuse notation and denote $d_{H_n^i,H_n^j}$ as d_{ij}^n. Let τ_x be the posterior distribution generated by information structure x and prior μ_0. Given a belief $\mu \in \Delta(\Theta)$, I denote the likelihood ratio of H_n^i relative to the prior as $\hat{\mu}_n^i := \mu(H_n^i)/\mu_0(H_n^i)$. The maximal log-likelihood ratio between two categories under x is denoted by

$$\overline{\ell}_{ij}^n(x) := \sup_{\mu \in \operatorname{Supp} \tau_x} \log\left(\frac{\hat{\mu}_n^i}{\hat{\mu}_n^j}\right) = \sup_{\hat{S} \in \mathcal{S}} \log \frac{\sum_{\theta \in H_n^i} x(\hat{S}|\theta)\mu_0(\theta|H_n^i)}{\sum_{\theta \in H_n^j} x(\hat{S}|\theta)\mu_0(\theta|H_n^j)}.$$

By Lemma A.1, the likelihood-vector-based representation in the theorem is equivalent to a belief-based representation of the form

$$L(x) = \sup_{\mu \in \operatorname{Supp} \tau_x} \max_{n,\, i \neq j} d_{ij}^n \left(\log \frac{\hat{\mu}_n^i}{\hat{\mu}_n^j}\right). \tag{6}$$

I prove the belief-based representation in what follows.

Proof of Theorem A.1: "If" direction. Suppose L is given by Equation 6. I first show that L is a worst-case privacy measure, and then that it satisfies $\mathcal{C}_{\mu_0,\mathcal{H}}$-Monotonicity.

(Step 1: Verifying that L is a Worst-Case Privacy Measure) Define the belief-based index function

$$d(\mu) := \max_{n,\, i \neq j} d_{ij}^n \left(\log \frac{\hat{\mu}_i}{\hat{\mu}_j}\right).$$

We verify that d is quasi-convex and lower-semicontinuous. For any n and any pair i, j, let $\tilde{d}_{ij}^n(\mu) := d_{ij}^n(\log(\hat{\mu}_n^i/\hat{\mu}_n^j))$. Since d_{ij}^n is increasing and lower-semicontinuous, the lower contour set $\{\mu : \tilde{d}_{ij}^n(\mu) \leq \delta\}$ is equivalent to the set $\{\mu : \log(\hat{\mu}_n^i/\hat{\mu}_n^i) \leq a\}$ for some $a \geq 0$. This is the set $\{\mu : \mu(H_n^i) \leq e^a \frac{\mu_0(H_n^i)}{\mu_0(H_n^i)}\mu(H_n^i)\}$, which is a closed convex set. Thus, each $\tilde{d}_{ij}^n$ is quasi-convex and lower-semicontinuous. As the maximum of such functions, d is also quasi-convex and lower-semicontinuous. Furthermore, at the prior μ_0, we have $\hat{\mu}_{0,n}^i = 1$ for all i and n, so $d(\mu_0) = 0$. By the "if" direction of Theorem 1, $L(x) = \sup_{\mu \in \operatorname{Supp} \tau_x} d(\mu)$ satisfies Worst-Case Protection, Lower-Semicontinuity, and normalization.

(Step 2: Verifying $\mathcal{C}_{\mu_0}^P$-Monotonicity) Let $x, x' \in \mathcal{X}$ be such that x $\mathcal{C}_{\mu_0}^P$-dominates x'. Suppose, for the sake of contradiction, that $L(x) < L(x')$. By the representation of L, this implies there exists some n and a pair i, j such that $d_{ij}^n(\overline{\ell}_{ij}^n(x)) < d_{ij}^n(\overline{\ell}_{ij}^n(x'))$. Since d_{ij}^n is increasing, we must have $\overline{\ell}_{ij}^n(x) < \overline{\ell}_{ij}^n(x')$.

Let $r := e^{\overline{\ell}_{ij}^n(x)}$. Note that $1 \leq r < \infty$. Consider the following prediction problem with

$A = \{a_n^1, ..., a_n^{|\mathcal{H}_n|}\}$ and payoffs:

$$u(\theta, a_n^i) = 1 \text{ for } \theta \in H_n^i, \quad u(\theta, a_n^j) = \frac{\mu_0(H_n^i)}{\mu_0(H_n^j)} r \text{ for } \theta \in H_n^j,$$

and $u(\theta, a_n^{i'}) = 0$ for all other $(\theta, a_n^{i'})$. An agent with posterior μ strictly prefers action a_n^i to a_n^j if and only if $\mu(H_n^i) > \frac{\mu_0(H_n^i)}{\mu_0(H_n^j)} r \mu(H_n^j)$, which is equivalent to $\log(\hat{\mu}_n^i / \hat{\mu}_n^j) > \log r = \overline{\ell}_{ij}^n(x)$.

Under x, for any realized posterior $\mu \in \operatorname{Supp} \tau_x$, we have $\log(\hat{\mu}_i / \hat{\mu}_j) \leq \overline{\ell}_{ij}(x)$. Thus, the adversary's optimal action is always a_n^j (or indifferent). The expected payoff under x is therefore $\mu_0(H_j) \frac{\mu_0(H_n^i)}{\mu_0(H_n^j)} r = \mu_0(H_n^i) r$.

Under x', since $\overline{\ell}_{ij}^n(x') > \overline{\ell}_{ij}^n(x)$, there exists a set of signals $\hat{S}$ with positive probability under which the realized posterior $\mu_{\hat{S}}$ satisfies $\log\left(\hat{\mu}_{\hat{S},n}^i / \hat{\mu}_{\hat{S},n}^j\right) > \overline{\ell}_{ij}^n(x)$. For these signals, the adversary strictly prefers action a_n^i. For all other signals, the agent prefers a_n^j. The expected payoff under x' is therefore strictly greater than $\mu_0(H_n^i) r$.

This implies that x' yields a strictly higher expected payoff than x in this specific prediction problem, which contradicts the premise that x $\mathcal{C}_{\mu_0}^P$-dominates x'. Therefore, we must have $L(x) \geq L(x')$. *Q.E.D.*

Remark. *If d_{ij}^n's are not l.s.c., the lower contour set in Step 1 can be represented by strict inequalities. These sets may not be closed but still convex, so d may not be l.s.c., but is still quasi-convex.*

The proof for the "only if" direction hinges on the following lemma, which establishes that for a privacy measure satisfying the axioms, dominance in terms of maximal log-likelihood ratios implies dominance in terms of privacy loss.

Lemma A.3. *Suppose L satisfies $\mathcal{C}_{\mu_0}^P$-Monotonicity and Axioms 1-3. If $x, x' \in \mathcal{X}_c$ satisfy that for all n and $i \neq j$, $\overline{\ell}_{ij}^n(x) \geq \overline{\ell}_{ij}^n(x')$, then $L(x) \geq L(x')$.*

Proof. Let $u_\sigma(\tau)$ denote the expected payoff for an adversary with decision problem $\sigma \in \mathcal{C}_{\mu_0}^P$ and posterior distribution τ. Denote the class of prediction problems corresponding to aspect n as $\mathcal{C}_{\mu_0,n}^P$. Note that $\mathcal{C}_{\mu_0}^P = \cup_{n=1}^N \mathcal{C}_{\mu_0,n}^P$. For $\sigma \in \mathcal{C}_{\mu_0,n}^P$, let $c_n^i := u(\theta, a_n^i) \geq 0$ for $\theta \in H_n^i$ and $\tilde{c}_n^i := c_n^i \mu_0(H_n^i)$. The interim payoff at posterior μ is $\max_i \tilde{c}_n^i \hat{\mu}_n^i$. The ex ante payoff is thus

$$u_\sigma(\tau) = \int_{\Delta\Theta} \max_i \tilde{c}_n^i \hat{\mu}_n^i d\tau(\mu). \tag{7}$$

Define A_{ij}^n as the following event:

$$A_{ij}^n := \left\{ \mu \in \Delta\Theta : \log \frac{\hat{\mu}_n^i}{\hat{\mu}_n^j} \geq \overline{\ell}_{ij}^n(\tau_{x'}) \right\}.$$

We first consider the case where $\tau_x(A^n_{ij}) > 0$ for all n and i, j such that $\overline{\ell}^n_{ij}(\tau_{x'}) > 0$. Let

$$a := \min_{n,\, i,j:\overline{\ell}^n_{ij}(\tau_{x'})>0} \tau_x(A^n_{ij}) > 0$$

be the smallest probability of such events. Define

$$b := \min_{n,\, i,j:\overline{\ell}^n_{ij}(\tau_{x'})>0} \mathbb{E}_{\tau_x}[\hat{\mu}^i_n | A^n_{ij}]\mu_0(H^i_n) > 0.$$

Let τ_0 be the degenerate distribution on μ_0. Next, we show that there exists $\beta > 0$ such that

$$\beta(u_\sigma(\tau_{x'}) - u_\sigma(\tau_0)) \leq u_\sigma(\tau_x) - u_\sigma(\tau_0) \tag{8}$$

for all $\sigma \in \mathcal{C}^P_{\mu_0}$. When $u_\sigma(\tau_{x'}) - u_\sigma(\tau_0) = 0$, the inequality always holds. Fix any $\sigma \in \mathcal{C}^P_{\mu_0}$ such that $u_\sigma(\tau_{x'}) - u_\sigma(\tau_0) > 0$. Let n be the aspect of concern in σ and i^* be a maximizer of $\tilde{c}^i_n$. Note that $u_\sigma(\tau_0) = \tilde{c}^{i^*}_n$. By Equation 7 and Bayes-plausibility, we have

$$u_\sigma(\tau_{x'}) - u_\sigma(\tau_0) = \int_{\Delta\Theta} \left(\max_i \{\tilde{c}^i_n \hat{\mu}^i_n\} - \tilde{c}^{i^*}_n \hat{\mu}^{i^*}_n \right) d\tau_{x'}(\mu).$$

Let $\xi \in \operatorname{Supp} \tau_{x'}$ be a belief that maximizes $\max_i\{\tilde{c}^i_n \hat{\mu}^i_n\} - \tilde{c}^{i^*}_n \mu^{i^*}_n$ within $\operatorname{Supp} \tau_{x'}$.[35] Then we have

$$0 < u_\sigma(\tau_{x'}) - u_\sigma(\tau_0) \leq \max_i \{\tilde{c}^i_n \hat{\xi}^i_n\} - \tilde{c}^{i^*}_n \hat{\xi}^{i^*}_n.$$

Let i' be a maximizer of $\tilde{c}^i_n \hat{\xi}^i_n$.

For $\mu \in A_{i',i^*}$, we have

$$\frac{\hat{\mu}^{i'}_n}{\hat{\mu}^{i^*}_n} \geq \frac{\hat{\xi}^{i'}_n}{\hat{\xi}^{i^*}_n} > \frac{\tilde{c}^{i^*}_n}{\tilde{c}^{i'}_n} \geq 1. \tag{9}$$

This implies that $\overline{\ell}^n_{i',i^*}(x') > 0$ and that

$$\begin{aligned}
u_\sigma(\tau_x) - u_\sigma(\tau_0) &\geq \tau_x(A^n_{i',i^*}) \mathbb{E}_{\tau_x}[\tilde{c}^{i'}_n \hat{\mu}^{i'}_n - \tilde{c}^{i^*}_n \hat{\mu}^{i^*}_n | A^n_{i',i^*}] \\
&\geq \tau_x(A^n_{i',i^*}) \mathbb{E}_{\tau_x}[\tilde{c}^{i'}_n \hat{\mu}^{i'}_n - \tilde{c}^{i^*}_n \hat{\xi}^{i^*}_n \frac{\hat{\mu}^{i'}_n}{\hat{\xi}^{i'}_n} | A^n_{i',i^*}] \\
&= \tau_x(A^n_{i',i^*}) \frac{\mathbb{E}_{\tau_x}[\hat{\mu}^{i'}_n | A^n_{i',i^*}]}{\hat{\xi}^{i'}_n} (\tilde{c}^{i'}_n \hat{\xi}^{i'}_n - \tilde{c}^{i^*}_n \hat{\xi}^{i^*}_n) \\
&\geq \tau_x(A^n_{i',i^*}) \mathbb{E}_{\tau_x}[\hat{\mu}^{i'}_n | A^n_{i',i^*}] \mu_0(H^{i'}_n) (\tilde{c}^{i'}_n \hat{\xi}^{i'}_n - \tilde{c}^{i^*}_n \hat{\xi}^{i^*}_n) \\
&\geq ab(\tilde{c}^{i'}_n \hat{\xi}^{i'}_n - \tilde{c}^{i^*}_n \hat{\xi}^{i^*}_n)
\end{aligned},$$

[35] Such a belief exists because the support set is compact and the value function is continuous.

where the first inequality follows Equation 7, the second inequality follows Equation 9, the third inequality follows $0 < \hat{\xi}_n^{i'} \leq 1/\mu_0(H_n^{i'})$, and the last inequality is by definition of a and b. Combined with $u_\sigma(\tau_{x'}) - u_\sigma(\tau_0) \leq \tilde{c}_n^{i'}\hat{\xi}_n^{i'} - \tilde{c}_n^{i^*}\hat{\xi}_n^{i^*}$, we conclude that Equation 8 holds for all $\sigma \in \mathcal{C}_{\mu_0}^P$ when $\beta := ab > 0$. This means that τ_x $\mathcal{C}_{\mu_0}^P$-dominates $\beta\tau_{x'} + (1-\beta)\tau_0$. By $\mathcal{C}_{\mu_0}^P$-Monotonicity, we know that $L(x) \geq L(\beta\tau_{x'} + (1-\beta)\tau_0) = L(\tau_{x'})$, where the equality follows from Axiom 1.

Finally, when $\tau_x(A_{ij}^n) = 0$ for some $\overline{\ell}_{ij}^n(x') > 0$, take an increasing sequence $\{\alpha_m\} \subset (0,1)$ such that $\alpha_m \to 1$. Define a sequence of posterior distributions τ_m' as follows. For each $\mu \in \operatorname{Supp} \tau_{x'}$, let $\tau_m'(\alpha_m\mu + (1-\alpha_m)\mu_0) = \tau_{x'}(\mu)$. One immediately checks that $\tau_m' \in \mathcal{T}_c$ and $\tau_m' \to \tau_{x'}$. By Axiom 3, we know that $L(x') \leq \liminf_m L(\tau_m')$. Moreover, note that by definition $\overline{\ell}_{ij}^n(\tau_m') < \overline{\ell}_{ij}^n(\tau_x)$ whenever $\overline{\ell}_{ij}^n(\tau_m') > 0$. Define

$$A_{ij,m}^n := \left\{ \mu \in \Delta\Theta : \log \frac{\hat{\mu}_n^i}{\hat{\mu}_n^j} \geq \overline{\ell}_{ij}^n(\tau_m') \right\}.$$

By construction $\tau_x(A_{ij,m}^n) > 0$ for all n and i, j such that $\overline{\ell}_{ij}^n(\tau_m') > 0$. The previous argument shows that $L(\tau_m') \leq L(x)$ and thus $L(x') \leq L(x)$. *Q.E.D.*

Remark. *When only $\mathcal{X}_f$ is concerned, it always holds that $\tau_x(A_{ij}^n) > 0$ for all n and i, j since τ_x has a finite support. Therefore, the first part of the argument suffices, and Axiom 3 is not needed.*

Now we prove the main result.

Proof of Theorem A.1: "only if" direction. For every $i \neq j$ and $l \in \overline{\mathbb{R}}$, define

$$d_{ij}^n(l) := \inf_{\tau \in \mathcal{T}_c : \overline{\ell}_{ij}^n(\tau) \geq l} L(\tau), \tag{10}$$

Note that by definition $d_{ij}^n(l) \geq 0$, $d_{ij}^n(0) = 0$, and d_{ij}^n is increasing in l. To see d_{ij}^n is lower-semicontinuous, take any converging sequence $l_m \to l$. Let $\tilde{d} := \liminf_m d_{ij}^n(l_m)$. We want to show $d_{ij}^n(l) \leq \tilde{d}$. Without loss, assume $\tilde{d} < \infty$. By definition of $\tilde{d}$, there exists a subsequence, re-labeled by m, such that $d_{ij}^n(l_m) \leq \tilde{d} + \varepsilon/2$ for all m. By definition of d_{ij}^n, there exists $\tau_m \in \mathcal{T}_c$ such that $\overline{\ell}_{ij}^n(\tau_m) \geq l_m$ and $L(\tau_m) \leq \tilde{d} + \varepsilon$ for all n. Now, take any sequence $\{\alpha_m\}$ with $\alpha_m > 0$ and $\sum_{m=1}^\infty \alpha_m = 1$. Let $\tau := \sum_{m=1}^\infty \alpha_m \tau_m$ and $\tilde{\tau}_h := \sum_{m=1}^h \frac{\alpha_m}{\sum_{m'=1}^h \alpha_{m'}} \tau_m$. Following the same argument as in the proof of Theorem 1 for lower-semicontinuity, we know that (i) $\tau \in \mathcal{T}_c$; (ii) $L(\tilde{\tau}_h) \leq \tilde{d} + \varepsilon$ for all h; (iii) $\tilde{\tau}_h \to \tau$. By Axiom 3, $L(\tau) \leq \tilde{d} + \varepsilon$. Moreover, note that by construction $\overline{\ell}_{ij}^n(\tau) \geq l$. Therefore, by definition of d_{ij}^n we have $d_{ij}^n(l) \leq L(\tau) \leq \tilde{d} + \varepsilon$. Since ε was arbitrary, we conclude that $d_{ij}^n(l) \leq \tilde{d}$.

Next, we show Equation 6 holds for all $x \in \mathcal{X}_f$. Fix any $x \in \mathcal{X}_f$. By definition, $d_{ij}^n(\overline{\ell}_{ij}^n(\tau_x)) \leq L(\tau_x)$ for all n and $i \neq j$, so $\max_{n,\, i\neq j} d_{ij}^n(\overline{\ell}_{ij}^n(\tau_x)) \leq L(x)$. It is left to argue that $L(x) \leq \max_{n,\, i\neq j} d_{ij}^n(\overline{\ell}_{ij}^n(\tau_x))$. Fix any $\varepsilon > 0$. For every n and $i \neq j$, take $\tau_{ij}^n \in \mathcal{T}_c$ such that $\overline{\ell}_{ij}^n(\tau_{ij}^n) \geq \overline{\ell}_{ij}^n(\tau_x)$ and $L(\tau_{ij}^n) \leq d_{ij}^n(\overline{\ell}_{ij}^n(\tau_x)) + \varepsilon$. Let $\tilde{\tau} \in \mathcal{T}_c$ be the uniform mixture across τ_{ij}^n for all n and $i \neq j$. By Axiom 1, we know that $L(\tilde{\tau}) = \max_{n,\, i\neq j} L(\tau_{ij}^n) \leq \max_{n,\, i\neq j} d_{ij}^n(\overline{\ell}_{ij}^n(\tau_x)) + \varepsilon$. Moreover, by construction we have $\overline{\ell}_{ij}^n(\tilde{\tau}) \geq \overline{\ell}_{ij}^n(\tau_x)$ for all n and $i \neq j$. By Lemma A.3, $L(x) \leq L(\tilde{\tau})$. Since ε was arbitrary, we conclude that $L(x) \leq \max_{n,\, i\neq j} d_{ij}^n(\overline{\ell}_{ij}^n(\tau_x))$. Finally, since Equation 6 holds for all $x \in \mathcal{X}_f$, by Lemma A.2 we know that it also holds for all $x \in \mathcal{X}$.[36] *Q.E.D.*

Remark. *When only $\mathcal{X}_f$ is concerned, d_{ij}^n is defined by ranging over all $\tau \in \mathcal{T}_f : \overline{\ell}_{ij}^n(\tau) \geq l$ in Equation 10. The proof for l.s.c. of d_{ij}^n is no longer needed, and thus Axiom 3 is not needed.*

Proof of Theorem 2. Theorem 2 follows directly from Theorem A.1 by setting each $\mathcal{H}_n$ to be the partition according to T_n. *Q.E.D.*

B. Proofs of Section 3

Let $C_f(S)$ denote firm f's most-preferred feasible subset of workers from a given set $S \subset W$, respecting its capacity constraint.

To set up the proof, we first introduce some notation. For a matching mechanism x, let P_x denote the joint distribution over $M \times \Theta$ induced by x and the prior μ_0. We use tildes (e.g., $\tilde{m}$, $\tilde{\theta}$) to denote random objects. For a permutation π on W, we define its action on matchings and preferences as in the main text. If we let $\boldsymbol{\theta}_{-w}$ denote the preference profile for participants other than w, which can be regarded as a set of preference profiles, then $\boldsymbol{\theta}'_{-\pi^{-1}(w)} = \pi(\boldsymbol{\theta}_{-w})$ denotes the set of preference profiles such that $\theta'_w = \theta_{\pi(w)}$ for all $w \neq \pi^{-1}(w)$ and $\theta'_f = \pi(\theta_f)$ for all f. Note that $\boldsymbol{\theta}'_{-\pi^{-1}(w)}$ is an element of $\Theta_{-\pi^{-1}(w)}$.

Our first result shows that it is without loss of generality to restrict attention to symmetric mechanisms. We say a welfare criterion for workers, $V(x)$, is *symmetric* if it is invariant to relabeling of workers, i.e., $V(x) = V(x^\pi)$ for any permutation π. A welfare criterion is *linear* if for any two mechanisms x_1, x_2 and any $\alpha \in [0,1]$, $V(\alpha x_1 + (1-\alpha)x_2) = \alpha V(x_1) + (1-\alpha)V(x_2)$. We say two mechanisms x and $\overline{x}$ are *welfare-equivalent* if they are judged to be equally good by any symmetric and linear welfare criterion.

[36] In order to invoke Lemma A.2, we still need to check that $\max_{n,\, i\neq j} d_{ij}^n(\log(\mu_n^i/\mu_n^j))$ as a function of μ is quasi-convex and lower-semicontinuous. The argument for this is in Step 1 of the proof of Theorem 2's "if" direction.

Lemma B.1. *For any stable matching mechanism x, there exists a symmetric stable matching mechanism $\overline{x}$ that is welfare-equivalent to x and has a weakly lower privacy loss, $L(\overline{x}) \leq L(x)$.*

Proof. We show that for any stable matching mechanism x, we can construct a symmetric stable matching mechanism $\overline{x}$ that is welfare-equivalent and satisfies $L(\overline{x}) \leq L(x)$. Consider any permutation $\pi : W \to W$. Define x^π as the matching mechanism such that

$$x^\pi(m|\boldsymbol{\theta}) = x(\pi^{-1}(m)|\pi^{-1}(\boldsymbol{\theta})).$$

We first argue that x^π is a stable matching mechanism. Suppose not, then there exists a matching m and preference profile $\boldsymbol{\theta}$ such that m is not stable under $\boldsymbol{\theta}$ and $x^\pi(m|\boldsymbol{\theta}) > 0$. Let worker w and firm f be a blocking pair. By definition, we have $f \succ_w m(w)$ and f prefers w to one of its assigned workers.[37] Denote $m' := \pi^{-1}(m)$ and $\boldsymbol{\theta}' := \pi^{-1}(\boldsymbol{\theta})$. Then $x(m'|\boldsymbol{\theta}') > 0$. Under $\boldsymbol{\theta}'$, worker $\pi^{-1}(w)$ has the same preferences as w under $\boldsymbol{\theta}$, and firms' preferences are just relabeled. Therefore, $\pi^{-1}(w)$ and f form a blocking pair for m' under $\boldsymbol{\theta}'$, a contradiction.

Let N be the total number of permutations on W. Define the symmetrized matching mechanism $\overline{x}$ as

$$\overline{x} = \frac{1}{N} \sum_{\pi} x^\pi. \tag{11}$$

Since each x^π is stable, their mixture $\overline{x}$ is also stable. Moreover, $\overline{x}$ is symmetric by construction.

Next, we show welfare equivalence. Let $V(x)$ be any welfare criterion that is symmetric and linear in the mechanism. By linearity, we have $V(\overline{x}) = V(\frac{1}{N}\sum_\pi x^\pi) = \frac{1}{N}\sum_\pi V(x^\pi)$. By the symmetry of the criterion, $V(x^\pi) = V(x)$ for all π. Therefore, $V(\overline{x}) = \frac{1}{N}\sum_\pi V(x) = V(x)$. Thus, $\overline{x}$ is welfare-equivalent to x.

Finally, we show that $L(\overline{x}) \leq L(x)$. To see this, let $\hat{x}$ be the compound mechanism that first draws π uniformly and then implements x^π. In other words, $\hat{x}$ implements the same outcome as $\overline{x}$, but in addition also informs the observer about which permutation π has

[37] Recall that we maintain the assumption $|W| \geq \sum_f q_f$ so that all firms' capacities are filled.

been chosen. Note that for all π we have

$$\begin{aligned}
P_{x^\pi}(m|\theta_w) &= \sum_{\boldsymbol{\theta}_{-w}} x^\pi(m|\theta_w, \boldsymbol{\theta}_{-w})\mu_0(\boldsymbol{\theta}_{-w}) \\
&= \sum_{\boldsymbol{\theta}_{-w}} x(\pi^{-1}(m)|\pi^{-1}(\theta_w), \pi^{-1}(\boldsymbol{\theta}_{-w}))\mu_0(\pi^{-1}(\boldsymbol{\theta}_{-w})) \\
&= \sum_{\boldsymbol{\theta}_{-\pi^{-1}(w)}} x(\pi^{-1}(m)|\pi^{-1}(\theta_w), \boldsymbol{\theta}_{-\pi^{-1}(w)})\mu_0(\boldsymbol{\theta}_{-\pi^{-1}(w)}) \\
&= P_x(\pi^{-1}(m)|\pi^{-1}(\theta_w)).
\end{aligned}$$

The first equality holds because $\tilde{\theta}_w$ is independent of $\tilde{\boldsymbol{\theta}}_{-w}$. The second equality holds by definition of x^π and the symmetry of μ_0. The third equality holds because π^{-1} is a bijection between Θ_{-w} and $\Theta_{-\pi^{-1}(w)}$. As a result, we have

$$\begin{aligned}
L(x^\pi) &= \max_{m,w,\theta_w,\hat{\theta}_w \in \Theta_w} \log \frac{P_{x^\pi}(m|\theta_w)}{P_{x^\pi}(m|\hat{\theta}_w)} \\
&= \max_{m,w,\theta_w,\hat{\theta}_w \in \Theta_w} \log \frac{P_x(\pi^{-1}(m)|\pi^{-1}(\theta_w))}{P_x(\pi^{-1}(m)|\pi^{-1}(\hat{\theta}_w))} \\
&= L(x).
\end{aligned}$$

Therefore, by Axiom 1 we have $L(\hat{x}) = \max_\pi L(x^\pi) = L(x)$. Since $\overline{x}$ is a garbling of $\hat{x}$ (it withholds the information about π), by Axiom 2 we have $L(\overline{x}) \leq L(\hat{x}) = L(x)$. *Q.E.D.*

The next lemma implies that, under a symmetric matching mechanism, instead of checking each matching outcome, we only need to focus on the rank of the matched firm for workers.

Lemma B.2. *Fix any symmetric stable matching mechanism x. For any matching m such that all firms' capacities are filled, worker w, and preferences $\theta_w, \hat{\theta}_w \in \Theta_w$, we have*

$$\frac{P_x(m|\theta_w)}{P_x(m|\hat{\theta}_w)} = \frac{P_x(\tilde{m}(w) = m(w)|\theta_w)}{P_x(\tilde{m}(w) = m(w)|\hat{\theta}_w)}.$$

Proof. In this proof, we suppress the dependence on x and write P_x as P. We prove that for all matchings m' such that all firms' capacities are filled and $m'(w) = m(w)$, we have $P(m|\theta_w) = P(m'|\theta_w)$ for all $\theta_w \in \Theta_w$. This implies the desired equality since x is stable and thus $\tilde{m}$ is supported on matchings where all firms' capacities are filled.

Note that

$$P(m|\theta_w) = \sum_{\boldsymbol{\theta}_{-w}} x(m|\theta_w, \boldsymbol{\theta}_{-w})\mu_0(\boldsymbol{\theta}_{-w}) \tag{12}$$

because $\tilde{\theta}_w$ is independent of $\tilde{\boldsymbol{\theta}}_{-w}$. Since under m and m', each firm is matched to the same number of workers and $m(w) = m'(w)$, there is a permutation $\pi : W \to W$ such that $\pi(w) = w$ and $m' = \pi(m)$. We have:

$$\begin{aligned}
P(m'|\theta_w) &= \sum_{\boldsymbol{\theta}_{-w}} x(m'|\theta_w, \boldsymbol{\theta}_{-w})\mu_0(\boldsymbol{\theta}_{-w}) \\
&= \sum_{\boldsymbol{\theta}_{-w}} x(\pi(m)|\theta_w, \boldsymbol{\theta}_{-w})\mu_0(\boldsymbol{\theta}_{-w}) \\
&= \sum_{\boldsymbol{\theta}_{-w}} x(m|\theta_w, \pi^{-1}(\boldsymbol{\theta}_{-w}))\mu_0(\pi^{-1}(\boldsymbol{\theta}_{-w})) \\
&= \sum_{\boldsymbol{\theta}_{-w}} x(m|\theta_w, \boldsymbol{\theta}_{-w})\mu_0(\boldsymbol{\theta}_{-w}) \\
&= P(m|\theta_w).
\end{aligned}$$

The first and last equalities follow from Equation 12. The second equality follows from the definition of π. The third equality holds because both x and μ_0 are symmetric, and $\pi(w) = w$. The fourth equality holds because π^{-1} induces a bijection between Θ_{-w} and Θ_{-w}. As a result, the third and fourth lines are summing over the same terms. This completes the proof. *Q.E.D.*

The next lemma establishes a key monotonicity property for the canonical firm- and worker-optimal stable mechanisms.

Lemma B.3. *Let θ_w^i be a preference list for worker w where firm f is ranked in the i-th position, holding the relative ranking of other firms fixed.*

1. *Under the firm-optimal stable mechanism (x^F), the probability of worker w matching with firm f, $P_{x^F}(m(w) = f|\theta_w^i)$, is decreasing in i.*
2. *Under the worker-optimal stable mechanism (x^W), the probability of worker w matching with firm f, $P_{x^W}(m(w) = f|\theta_w^i)$, is decreasing in i.*

Proof. We prove each statement in turn. Let $i < j$. Let $\boldsymbol{\theta}^i = (\theta_w^i, \boldsymbol{\theta}_{-w})$ and $\boldsymbol{\theta}^j = (\theta_w^j, \boldsymbol{\theta}_{-w})$, where θ_w^i ranks firm f higher than θ_w^j. Let $m^i = x^F(\boldsymbol{\theta}^i)$ and $m^j = x^F(\boldsymbol{\theta}^j)$ be the firm-optimal stable matchings for the respective profiles. We will show that for any fixed $\boldsymbol{\theta}_{-w}$, if $m^j(w) = f$, then $m^i(w) = f$. Integrating over all $\boldsymbol{\theta}_{-w}$ then establishes the desired weak inequality.

(Part 1: Firm-Proposing DA) Suppose $m^j(w) = f$. We first show that m^j is also stable under the profile $\boldsymbol{\theta}^i$. Suppose not, then there is a blocking pair (w', f') for m^j under $\boldsymbol{\theta}^i$. Since all preferences other than worker w's are unchanged, and m^j is stable under $\boldsymbol{\theta}^j$, any new

blocking pair must involve w. So, $w' = w$. This means $f' \succ_w f$ under θ^i_w. Since f is ranked higher in θ^i_w than in θ^j_w, this implies $f' \succ_w f$ under θ^j_w as well. This means (w, f') would also have been a blocking pair for m^j under $\boldsymbol{\theta}^j$, a contradiction. Thus, m^j is stable under $\boldsymbol{\theta}^i$.

Since m^j is a stable matching under $\boldsymbol{\theta}^i$, by the lattice property of stable matchings, all firms must weakly prefer m^i to m^j. Next, we show that m^i is also stable under $\boldsymbol{\theta}^j$. Suppose not, then there is a blocking pair (w', f') under $\boldsymbol{\theta}^j$. Again, this pair must be (w, f'). This means $f' \succ_w m^i(w)$ under θ^j_w, but not under θ^i_w. This can only happen if $m^i(w) = f$. If $m^i(w) = f$, then $f' \succ_w f$ under θ^j_w. Since (w, f') blocks m^i under $\boldsymbol{\theta}^j$, firm f' must prefer w to its match in m^i. But since firms prefer m^i to m^j, this means f' also prefers w to its match in m^j. This implies (w, f') is a blocking pair for m^j under $\boldsymbol{\theta}^j$, a contradiction.

So, m^i is also stable under $\boldsymbol{\theta}^j$. Since m^j is the firm-optimal stable matching for that profile, firms must weakly prefer m^j to m^i. Combined with the fact that firms weakly prefer m^i to m^j, it must be that firms are indifferent between the two matchings, which implies $m^i = m^j$.[38] Therefore, if $m^j(w) = f$, then $m^i(w) = f$.

(Part 2: Worker-Proposing DA) Let $m^i = x^W(\boldsymbol{\theta}^i)$ and $m^j = x^W(\boldsymbol{\theta}^j)$. Suppose $m^j(w) = f$. We first show that m^j is also stable under $\boldsymbol{\theta}^i$. Suppose not, then there is a blocking pair (w', f') for m^j under $\boldsymbol{\theta}^i$. This pair must be (w, f'). This means $f' \succ_w f$ under θ^i_w. As f is ranked higher in θ^i_w than in θ^j_w, this implies $f' \succ_w f$ under θ^j_w as well. This contradicts the stability of m^j under $\boldsymbol{\theta}^j$. So, m^j is stable under $\boldsymbol{\theta}^i$.

Since m^j is a stable matching under profile $\boldsymbol{\theta}^i$, and m^i is the worker-optimal stable matching for that profile, worker w must weakly prefer her assignment in m^i to her assignment in m^j. That is, $m^i(w) \succeq_w f$ under preference θ^i_w.

Now, suppose for contradiction that $m^i(w) = f'$ with $f' \succ_w f$ under θ^i_w. We argue that m^i must also be stable under $\boldsymbol{\theta}^j$. Suppose not, then there is a blocking pair (w', f'') for m^i under $\boldsymbol{\theta}^j$. This pair must be (w, f''). This means $f'' \succ_w m^i(w)$ under θ^j_w, but not under θ^i_w. This is impossible, since the set of firms preferred to f' is the same under both preference lists. Thus, m^i is stable under $\boldsymbol{\theta}^j$. But this leads to a contradiction. We have a stable matching m^i under profile $\boldsymbol{\theta}^j$ where worker w is matched to $f' \succ_w f$ under θ^j_w. This contradicts the fact that m^j, in which w is matched to f, is the worker-optimal stable matching for profile $\boldsymbol{\theta}^j$. Therefore, it must be that $m^i(w) = f$. *Q.E.D.*

Lemma B.4. *For any symmetric, monotonic, stable matching mechanism x, there exist a worker $w \in W$, a firm $f \in F$, and preference lists $\theta_w, \hat{\theta}_w \in \Theta_w$ where f is ranked first in θ_w and last in $\hat{\theta}_w$,*

[38] This follows Theorem 5.26 of Roth and Sotomayor (1990), which states that firms have strict preferences over groups of workers that they may be assigned at stable matchings.

such that the privacy loss is given by

$$L(x) = \log \frac{P_x(\tilde{m}(w) = f|\theta_w)}{P_x(\tilde{m}(w) = f|\hat{\theta}_w)}.$$

Proof. Suppose that the maximum privacy loss $L(x)$ is attained at some matching m, worker w, and preferences $\theta_w, \hat{\theta}_w$. The privacy loss is therefore:

$$L(x) = \log \frac{P_x(m|\theta_w)}{P_x(m|\hat{\theta}_w)}. \tag{13}$$

We first argue that we can, without loss of generality, assume that w is matched under m. Suppose w is unmatched in m. By Lemma B.2, the likelihood ratio simplifies to:

$$\frac{P_x(m|\theta_w)}{P_x(m|\hat{\theta}_w)} = \frac{P_x(\tilde{m}(w) = \emptyset|\theta_w)}{P_x(\tilde{m}(w) = \emptyset|\hat{\theta}_w)}. \tag{14}$$

By the Rural Hospital Theorem (Roth (1986)), for any preference profile, the set of matched workers is the same in all stable matchings. This implies that for any fixed $\boldsymbol{\theta}$, $P_x(\tilde{m}(w) = \emptyset|\boldsymbol{\theta})$ is either 0 or 1. Furthermore, for any $\boldsymbol{\theta}_{-w}$, if w is unmatched in a stable matching under $(\theta_w, \boldsymbol{\theta}_{-w})$, that matching must also be stable under $(\hat{\theta}_w, \boldsymbol{\theta}_{-w})$ and thus she must also be unmatched in any stable matching under $(\hat{\theta}_w, \boldsymbol{\theta}_{-w})$. This is because any new blocking pair under $(\hat{\theta}_w, \boldsymbol{\theta}_{-w})$ must involve w, but since all matches are acceptable, any blocking pair w can form under $(\hat{\theta}_w, \boldsymbol{\theta}_{-w})$ if also a blocking pair under $(\theta_w, \boldsymbol{\theta}_{-w})$. Thus, $P_x(\tilde{m}(w) = \emptyset|\theta_w, \boldsymbol{\theta}_{-w}) = P_x(\tilde{m}(w) = \emptyset|\hat{\theta}_w, \boldsymbol{\theta}_{-w})$. Integrating over $\boldsymbol{\theta}_{-w}$ implies that the ratio in (14) is equal to 1. Therefore, it cannot be larger than any ratio where w is matched.

Let $f := m(w)$ be the firm worker w is matched to. Since x is symmetric, by Lemma B.2, the value of the likelihood ratio in (13) depends only on the event that w is matched to f. Since the mechanism x is monotonic by assumption, the probability $P_x(\tilde{m}(w) = f|\theta'_w)$ is decreasing in the rank of f in the preference list θ'_w. Therefore, the maximum is achieved by choosing θ_w such that f is ranked first and $\hat{\theta}_w$ such that f is ranked last. This completes the proof. *Q.E.D.*

Now we are ready to prove the main propositions.

Proof of Proposition 1 and Proposition 2. The proof proceeds in three parts. First, we show Proposition 2. As a corollary, we show that the firm-optimal stable matching mechanism x^F minimizes worker privacy loss among all stable mechanisms. Finally, we show that this minimal privacy loss is strictly greater than zero.

(Part 1: Higher Welfare Implies Lower Privacy) Let x be a stable matching mechanism and x' a symmetric, monotonic, stable matching mechanism, such that x is better for worker welfare than x'. Let $\overline{x}$ denote the symmetrized matching mechanism of x defined by Equation 11. I first argue that $\overline{x}$ is also better for worker welfare than x'. Let $F_w^x(\boldsymbol{\theta})$ denote the distribution of the rank of worker w's matched firm under $x(\cdot|\boldsymbol{\theta})$. Then we have that for all w and $\boldsymbol{\theta}$,

$$F_w^{\overline{x}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{\pi} F_w^{x^\pi}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{\pi} F^x_{\pi^{-1}(w)}(\pi^{-1}(\boldsymbol{\theta})),$$

where N is the total number of permutations. Moreover, by symmetry, we have for all $w, \boldsymbol{\theta}$ and permutations π on W,

$$F_w^{x'}(\boldsymbol{\theta}) = F^{x'}_{\pi^{-1}(w)}(\pi^{-1}(\boldsymbol{\theta}))$$

Since each $F^x_{\pi^{-1}(w)}(\pi^{-1}(\boldsymbol{\theta}))$ is first-order stochastically dominated by $F^{x'}_{\pi^{-1}(w)}(\pi^{-1}(\boldsymbol{\theta}))$, we conclude that $F_w^{\overline{x}}(\boldsymbol{\theta})$ is first-order stochastically dominated by $F_w^{x'}(\boldsymbol{\theta})$.

Since x' is symmetric and monotonic, by Lemma B.4, there exist $w, f, \theta_w, \hat{\theta}_w$ where f is ranked first in θ_w and last in $\hat{\theta}_w$ such that

$$L(x) = \log \frac{P_x(\tilde{m}(w) = f|\theta_w)}{P_x(\tilde{m}(w) = f|\hat{\theta}_w)}.$$

Now consider the symmetric mechanism $\overline{x}$. Take any matching m where all firms are filled and $m(w) = f$. By Lemma B.2, we have:

$$L(\overline{x}) \geq \log \frac{P_{\overline{x}}(m|\theta_w)}{P_{\overline{x}}(m|\hat{\theta}_w)} = \log \frac{P_{\overline{x}}(\tilde{m}(w) = f|\theta_w)}{P_{\overline{x}}(\tilde{m}(w) = f|\hat{\theta}_w)}.$$

For any fixed $\boldsymbol{\theta}$, w is either matched or unmatched in all stable matchings by the Rural Hospital Theorem. When w is unmatched, the probabilities of matching with f under both $\overline{x}$ and x are zero. When w is matched, since $F_w^{\overline{x}}(\boldsymbol{\theta})$ is first-order stochastically dominated by $F_w^{x'}(\boldsymbol{\theta})$, and f is ranked first in θ_w and last in $\hat{\theta}_w$, we have that for all $\boldsymbol{\theta}_{-w}$:

$$P_{\overline{x}}(\tilde{m}(w) = f|\theta_w, \boldsymbol{\theta}_{-w}) \geq P_{x'}(\tilde{m}(w) = f|\theta_w, \boldsymbol{\theta}_{-w}),$$

$$P_{\overline{x}}(\tilde{m}(w) = f|\hat{\theta}_w, \boldsymbol{\theta}_{-w}) \leq P_{x'}(\tilde{m}(w) = f|\hat{\theta}_w, \boldsymbol{\theta}_{-w}).$$

Therefore, integrating over the prior on $\boldsymbol{\theta}_{-w}$, we have:

$$P_{x'}(\tilde{m}(w) = f|\theta_w) \leq P_{\overline{x}}(\tilde{m}(w) = f|\theta_w), \quad P_{x'}(\tilde{m}(w) = f|\hat{\theta}_w) \geq P_{\overline{x}}(\tilde{m}(w) = f|\hat{\theta}_w).$$

Combining these inequalities directly implies that the likelihood ratio for x' is smaller than

that for $\overline{x}$, and thus $L(x') \leq L(\overline{x})$. By Lemma B.1, $L(\overline{x}) \leq L(x)$ and thus $L(x') \leq L(x)$.

(Part 2: Optimality of the Firm-Proposing DA) Note that x^F is symmetric, and by Lemma B.3, it is also monotonic. Moreover, by the lattice property, $F_w^{x^F}(\boldsymbol{\theta})$ FOSD $F_w^x(\boldsymbol{\theta})$ for all $w, \boldsymbol{\theta}$, and stable mechanism x. By Part 1, we conclude that $L(x^F) \leq L(x)$.

(Part 3: Impossibility of Perfect Privacy) Let P denote P_{x_F} for brevity. We aim to show that there exist a matching m, a worker w, and preferences $\theta_w, \hat{\theta}_w$ such that $P(m|\theta_w) > P(m|\hat{\theta}_w)$, implying that the privacy loss cannot be zero.

Consider any matching m in which all firms' capacities are filled. Let w be any matched worker under m, and let $f = m(w)$. Since $|F| > 1$, there exists another firm $f' \neq f$. Define θ_w so that f is ranked first and f' second. Let $\hat{\theta}_w$ be obtained from θ_w by swapping the ranks of f and f'. From the proof of Lemma B.3, we know that $P(m|\theta_w, \boldsymbol{\theta}_{-w}) \geq P(m|\hat{\theta}_w, \boldsymbol{\theta}_{-w})$ for all $\boldsymbol{\theta}_{-w}$. Since μ_0 has full support, it suffices to construct some $\boldsymbol{\theta}_{-w}$ for which the inequality is strict.

Consider any $\boldsymbol{\theta}_{-w}$ such that 1) Each $\theta_{\hat{w}}$ for $\hat{w} \neq w$ who is matched under m ranks $m(\hat{w})$ as the top choice, 2) $\theta_{f'}$ ranks w first, followed by $m(f')$, and 3) each $\theta_{\hat{f}}$ for $\hat{f} \neq f'$ ranks $m(\hat{f})$ at the top.[39] Let w' denote the least-preferred worker by f' in $m(f')$. Under $(\theta_w, \boldsymbol{\theta}_{-w})$, the firm-proposing DA proceeds as follows: in the first round, each firm $\hat{f} \neq f'$ proposes to $m(\hat{f})$ and get accepted immediately. f' proposes to $m(f') \cup \{w\} \backslash \{w'\}$ and get accepted by $m(f') \backslash \{w'\}$. In the second round, f' proposes to w' and gets accepted, resulting in matching m. Consequently, the matching m is realized under $(\theta_w, \boldsymbol{\theta}_{-w})$, yielding $P(m|\theta_w, \boldsymbol{\theta}_{-w}) = 1$.

Under $\hat{\theta}_w$, however, w now ranks f' first. Since f' also ranks w first, f' proposes to w and w accepts immediately. Therefore, w becomes matched to f' instead of f, so m does not occur. Thus, $P(m|\hat{\theta}_w, \boldsymbol{\theta}_{-w}) = 0$. This proves that $P(m|\theta_w) > P(m|\hat{\theta}_w)$ as required. *Q.E.D.*

B.1. A 3-by-3 Example

We consider a balanced one-to-one matching market with workers w_1, w_2, w_3 and firms f_1, f_2, f_3. Each firm has a unit demand. Every agent's preferences are i.i.d. uniform over the six strict orderings. We focus on worker w_1, and compute the probability of each matching conditional on w_1's type, under the worker- and firm-optimal stable matching mechanisms, respectively. We present the probabilities in the following matrices.[40] Rows index the six possible types of w_1 in lexicographic order. Columns index the six matchings (signals) writ-

[39] The ranking within $m(\hat{f})$ and unspecified rankings can be arbitrary.

[40] These matrices are derived by enumerating all preference profiles and counting the corresponding instances under the two DA mechanisms.

ten as "abc", meaning the matching $(w_1 \mapsto f_a, w_2 \mapsto f_b, w_3 \mapsto f_c)$. For ease of comparison, each matrix below is shown with a common denominator 432.

Worker-proposing DA (common denominator 432).

$W =$

	123	132	213	231	312	321
123	138	138	57	57	21	21
132	138	138	21	21	57	57
213	57	57	138	138	21	21
231	21	21	138	138	57	57
312	57	57	21	21	138	138
321	21	21	57	57	138	138

(each entry = numerator/432).

Firm-proposing DA (common denominator 432).

$F =$

	123	132	213	231	312	321
123	105	105	67	67	44	44
132	105	105	44	44	67	67
213	67	67	105	105	44	44
231	44	44	105	105	67	67
312	67	67	44	44	105	105
321	44	44	67	67	105	105

(each entry = numerator/432).

It is clear that W is not a garbling of F. To see that F is not a garbling of W, we construct a decision problem for which W has no value while F has a strictly positive value. Take the uniform prior on the six types of w_1. Consider the binary action problem where action a_0 gives a payoff of 0 over all states and action a_1 has a payoff vector given by:

$$u = (-5,\ 2,\ 5,\ -3,\ 5,\ -5).$$

The decision maker observes the signal (matching) and chooses a_1 if the posterior expectation $\mathbb{E}[u|m] > 0$. Because the prior is uniform, action a_1 is chosen if the inner product between u and the column in W or F corresponding to some matching m is positive. The following table lists these inner products (numerators over the common denominator 432), which the reader can check by direct multiplication of the columns of the displayed matri-

ces with u:

signal m	123	132	213	231	312	321
numerator under W	-12	-12	-147	-147	-57	-57
numerator under F	3	3	-152	-152	-67	-67

Under W, a_0 is strictly better after all signals, while under F, a_1 is strictly better after signals 123 and 132, and a_0 is strictly better after other signals. Therefore, F is strictly better for this decision problem; hence, F is not a garbling of W.

C. Proofs of Section 4

C.1. Duality Results

Proof of Theorem 3. We first argue that $U(\delta) \leq W(p)$. For all x such that $x \in \mathcal{D}$ and $L(x) \leq \delta$, by condition 1 we have that $\sum_\theta p(\theta, a)x(a|\theta)\mu_0(\theta) \leq 0$ for all $a \in A$. Therefore, we have

$$u(x) \leq u(x) - \sum_{\theta,a} p(\theta, a)x(a|\theta)\mu_0(\theta).$$

This implies that $U(\delta) \leq W(p)$.

To see x is optimal for Program ($\mathcal{P}_\delta$), note that

$$U(\delta) \leq W(p) = u(x) - \sum_{\theta,a} p(\theta, a)x(a|\theta)\mu_0(\theta) = u(x)$$

where the first equality follows from condition 2, and the second equality follows from condition 3. Since x is primal-feasible, we conclude that x is primal optimal. *Q.E.D.*

We establish strong duality for the problem ($\mathcal{P}_\delta$) under additional assumptions. Define the dual problem as:

$$\begin{aligned} W^*(\delta) := \inf_{p:\Theta\times A\to\mathbb{R}} & W(p) \\ s.t.\ & p(\cdot, a) \in P_\delta^{\mu_0},\ \forall a \in A. \end{aligned} \tag{$\mathcal{D}_\delta$}$$

We use ri to denote the relative interior.

Theorem C.1 (Strong Duality). *Suppose u is concave and $ri(\{x : u(x) > -\infty\}) \cap ri(\{x : L(x) \leq \delta\}) \neq \emptyset$. Then the following are true:*

1. *(Strong duality): $U(\delta) = W^*(\delta)$;*
2. *(Dual attainment): there exists some p that solves ($\mathcal{D}_\delta$).*

3. *(Complementary slackness): A primal-feasible x solves the primal problem if and only if there exists a dual solution p satisfying the conditions of Theorem 3.*

Proof. We prove the result for the uniform distribution μ_0. For other $\lambda \in \mathbb{R}^K_{++}$, the problem is equivalent by re-normalizing p.

We use Fenchel duality. Define two convex functions on $\mathbb{R}^{\Theta \times A}$:

$$f_1(x) = -u(x)$$
$$f_2(x) = \begin{cases} 0 & \text{if } L(x) \leq \delta \\ \infty & \text{otherwise} \end{cases}$$

The primal problem is equivalent to solving $-\inf_x\{f_1(x) + f_2(x)\}$. Since u is concave, f_1 is a convex function. Since d is quasi-convex, the set $\{x : L(x) \leq \delta\}$ is convex, so f_2 is also a convex function. The assumption that $\text{ri}(\{x : u(x) > -\infty\}) \cap \text{ri}(\{x : L(x) \leq \delta\}) \neq \emptyset$ is a Slater-type condition ensuring that $\text{ri}(\text{dom } f_1) \cap \text{ri}(\text{dom } f_2) \neq \emptyset$.

By the Fenchel-Rockafellar duality theorem (Rockafellar (1997), Theorem 16.4), strong duality holds:

$$\inf_x\{f_1(x) + f_2(x)\} = \sup_p\{-f_1^*(-p) - f_2^*(p)\},$$

and the supremum on the right is attained by some p.

Let's compute the conjugates. The inner product is denoted by $\langle p, x\rangle := \sum_{\theta,a} p(\theta, a)x(a|\theta)$.

$$f_1^*(-p) = \sup_x\{\langle -p, x\rangle - f_1(x)\} = \sup_x\{u(x) - \langle p, x\rangle\} = W(p).$$

$$f_2^*(p) = \sup_x\{\langle p, x\rangle - f_2(x)\} = \sup_{L(x)\leq\delta} \langle p, x\rangle = \sum_a \sup_{x(a|\cdot)\in C_\delta} \sum_\theta p(\theta, a)x(a|\theta).$$

This is the sum of support functions. $f_2^*(p)$ is 0 if $p(\cdot, a) \in P_\delta$ for all a, and ∞ otherwise.

Substituting back into the duality equation:

$$-U(\delta) = \sup_p\{-W(p) - f_2^*(p)\} = \sup_{p:p(\cdot,a)\in P_\delta}\{-W(p)\} = -\inf_{p:p(\cdot,a)\in P_\delta} W(p).$$

Therefore, $U(\delta) = \inf_{p:p(\cdot,a)\in P_\delta} W(p) = W^*(\delta)$. This proves (1) and (2).

For (3), the "if" direction is in Theorem 3. For the "only if" direction, let x^* be a solution to the primal problem and let p^* be a solution to the dual ($\mathcal{D}_\delta$). By strong duality, $u(x^*) =$

$W(p^*)$, so we must have

$$u(x^*) = \sup_x \{u(x) - \langle p^*, x\rangle\}.$$

This implies $u(x^*) \geq u(x^*) - \langle p^*, x^*\rangle$, which means $\langle p^*, x^*\rangle \geq 0$. However, x^* is primal feasible, so $x^*(a|\cdot) \in C_\delta$ for all a. And p^* is dual feasible, so $p^*(a|\cdot) \in P_\delta$ for all a. By definition of the polar cone, this implies $\sum_\theta p^*(\theta, a)x^*(a|\theta) \leq 0$ for each a. Summing over a gives $\langle p^*, x^*\rangle \leq 0$. The only way for both inequalities to hold is if $\langle p^*, x^*\rangle = 0$. Since each term in the sum over a is non-positive, each term must be zero. This proves Complementary Slackness. As a result, x^* attains the optimal value, so Priced Optimality is satisfied. Price Validity holds by definition of p^*. Thus, p^* is a valid certificate for x^*. *Q.E.D.*

C.2. Characterization of Polar Cone

Proof of Lemma 1. Denote $d = d_{\mu_0,\mathcal{H}} = \max_{i,j\in\mathcal{I}} \ell_{ij}$ for short and let $\nu_{\mu_0}(\theta) := \nu(\theta)\mu_0(\theta)$ be the μ_0-weighted likelihood vector. For a set $H \subset \Theta$, I abuse notation and denote $\nu_{\mu_0}(H) = \sum_{\theta\in H} \nu_{\mu_0}(\theta)$. We focus on Part 1 first.

(Part 1: "If" direction) Suppose the condition holds. For any $\nu \in C_\delta$, we have

$$p \cdot \nu_{\mu_0} \leq \sum_{i\in I} \left(\sum_{\theta\in H_i} p(\theta)\nu_{\mu_0}(\theta) \right) \leq \sum_{i\in I} p_i \nu_{\mu_0}(H_i) = \sum_{i\in I} p_i^+ \nu_{\mu_0}(H_i) - \sum_{i\in I} p_i^- \nu_{\mu_0}(H_i),$$

where the first inequality holds because $p(\theta) \leq 0$ for $\theta \notin \cup_i H_i$. Let

$$\nu_i := \frac{\nu_{\mu_0}(H_i)}{\mu_0(H_i)} = \sum_{\theta\in H_i} \nu(\theta)\mu_0(\theta|H_i).$$

The condition $d(\nu) \leq \delta$ implies $\nu_i \leq e^\delta \nu_j$ for all i, j, and thus $\max_i \nu_i \leq e^\delta \min_i \nu_i$. Let $\nu_{\max} = \max_i \nu_i$ and $\nu_{\min} = \min_i \nu_i$. We can bound the sum:

$$\sum_{i\in I} p_i^+ \nu_i \mu_0(H_i) - \sum_{i\in I} p_i^- \nu_i \mu_0(H_i) \leq \nu_{\max} \sum_i p_i^+ \mu_0(H_i) - \nu_{\min} \sum_i p_i^- \mu_0(H_i)$$

$$\leq e^\delta \nu_{\min} \sum_i p_i^+ \mu_0(H_i) - \nu_{\min} \sum_i p_i^- \mu_0(H_i) = \nu_{\min} \left(e^\delta \sum_i p_i^+ \mu_0(H_i) - \sum_i p_i^- \mu_0(H_i) \right).$$

Since the term in the parenthesis is non-positive by assumption, we have $p \cdot \nu_{\mu_0} \leq 0$.

(Part 1: "Only if" direction) We prove the contrapositive. First, suppose $p(\theta) > 0$ for some $\theta \notin \cup_i H_i$. Let ν be a vector with $\nu(\theta) > 0$ and zero otherwise. Then $d(\nu) = 0 \leq \delta$, but $p \cdot \nu_{\mu_0} = p(\theta)\nu(\theta)\mu_0(\theta) > 0$, so $p \notin P_\delta^{\mu_0}$. Next, suppose $\sum_i p_i^- \mu_0(H_i) < e^\delta \sum_i p_i^+ \mu_0(H_i)$. For

each i, let $\theta_i \in H_i$ be a state such that $p(\theta_i) = p_i$. Construct a vector ν such that $\nu(\theta_i) = 1$ if $p_i < 0$, and $\nu_{\mu_0}(\theta_i) = e^\delta$ if $p_i \geq 0$, and zero otherwise. By construction, $d(\nu) \leq \delta$. The inner product is:

$$p \cdot \nu_{\mu_0} = \sum_{i:p_i<0} p_i \mu_0(H_i) + \sum_{i:p_i \geq 0} p_i e^\delta \mu_0(H_i) = -\sum_i p_i^- \mu_0(H_i) + e^\delta \sum_i p_i^+ \mu_0(H_i) > 0.$$

This shows $p \notin P_\delta^{\mu_0}$, completing the proof.

(Part 2) Let $C_\delta^{\mu_0} := \{(\nu(\theta)\mu_0(\theta))_{\theta\in\Theta} : d(\nu) \leq \delta, \nu \in \mathbb{R}_+^K\}$. Then $C_\delta^{\mu_0} = \cap_j C_{\delta,j}^{\mu_0}$ and $P_{\delta,j}^{\mu_0}$ is the polar cone of $C_{\delta,j}^{\mu_0}$ for each j. Corollary 16.4.2 of Rockafellar (1997) implies that the polar cone of $\cap_j \text{cl}(C_{\delta,j}^{\mu_0})$ is $\text{cl}(\sum_j P_{\delta,j}^{\mu_0})$, where cl denotes closure. This implies that $\sum_j P_{\delta,j}^{\mu_0} \subset P_\delta^{\mu_0}$. *Q.E.D.*

C.3. Data Minimization

The data minimization problem can be formulated as

$$\begin{aligned} L^*(\underline{U}) := \inf_{x:\Theta\to\Delta(A)} \quad & L(x) \\ \text{s.t.} \quad & u(x) \geq \underline{U}, \end{aligned} \tag{$\mathcal{M}_{\underline{U}}$}$$

where $\underline{U}$ is the minimum acceptable utility.

The following proposition clarifies the relationship between Program ($\mathcal{M}_{\underline{U}}$) and ($\mathcal{P}_\delta$). Let $U_+^{-1}(y) := \inf\{\delta : U(\delta) \geq y\}$ and $(L^*)_-^{-1}(y) := \sup\{U : L^*(U) \leq y\}$ denote the generalized inverses. Then:

Proposition C.1. *Suppose u is upper-semicontinuous. Then $L^*(\underline{U}) = U_+^{-1}(\underline{U})$ and $U(\delta) = (L^*)_-^{-1}(\delta)$. Moreover, if x^* solves $\mathcal{P}_{L^*(\underline{U})}$, then x^* also solves $\mathcal{M}_{\underline{U}}$.*

Proof. First, we show $L^*(\underline{U}) = U_+^{-1}(\underline{U})$. By definition, $U_+^{-1}(\underline{U}) = \inf\{\delta : U(\delta) \geq \underline{U}\}$. The condition $U(\delta) \geq \underline{U}$ means $\sup\{u(x) : L(x) \leq \delta\} \geq \underline{U}$. Since u is upper-semicontinuous and L is lower-semicontinuous, we know that whenever the inequality holds, there exists x such that $L(x) \leq \delta$ and $u(x) \geq \underline{U}$. So, $U_+^{-1}(\underline{U})$ is the infimum δ for which such a mechanism exists. This is equivalent to finding the infimum of $L(x)$ over all mechanisms x that satisfy $u(x) \geq \underline{U}$. This is precisely the definition of $L^*(\underline{U})$. The proof for $U(\delta) = (L^*)_-^{-1}(\delta)$ is analogous.

Second, let x^* solve $\mathcal{P}_{L^*(\underline{U})}$. This means $L(x^*) \leq L^*(\underline{U})$, and $u(x^*) = U(L^*(\underline{U}))$. From the first part, since U is non-decreasing, $U(L^*(\underline{U})) = U(U_+^{-1}(\underline{U})) \geq \underline{U}$. Thus, $u(x^*) \geq \underline{U}$, which means x^* is feasible for $\mathcal{M}_{\underline{U}}$. The privacy loss of x^* is $L(x^*)$. Since x^* is feasible for $\mathcal{M}_{\underline{U}}$, by

definition of the infimum, $L(x^*) \geq L^*(\underline{U})$. Combining $L(x^*) \leq L^*(\underline{U})$ and $L(x^*) \geq L^*(\underline{U})$, we must have $L(x^*) = L^*(\underline{U})$. This shows that x^* achieves the infimum loss for $\mathcal{M}_{\underline{U}}$ and is feasible, so it is a solution. *Q.E.D.*

C.4. Voting

Proof of Proposition 3. The proof proceeds by constructing a set of shadow prices and an optimal voting rule, and then using Theorem 3 (with $\lambda := \mu_0$) to certify that this rule is indeed a solution to the planner's problem. In the end, we argue that the optimal voting rule is unique up to randomization at the cutoffs. We observe that when the privacy constraint is not binding, the majority rule is the unique optimal solution, which takes the form in the proposition. Therefore, in what follows, we assume the privacy constraint is binding.

(Step 1: Constructing the Shadow Prices) The privacy constraint in this problem is an LDP constraint applied to each voter i. This can be viewed as an intersection of individual privacy constraints. We start by conjecturing a simple, symmetric form for the shadow prices associated with each individual voter's privacy. For each voter i, let the price for revealing their signal be given by a vector $p^i(\theta, a)$. We conjecture that $p^i(\cdot, a)$ is measurable with respect to s^i and write it as $p^i(s^i, a)$ for short. By symmetry, we conjecture that for some price level $\hat{p} > 0$:

$$p^i(s_1, a_1) = p^i(s_0, a_0) = \hat{p}, \quad p^i(s_1, a_0) = p^i(s_0, a_1) = -e^{\delta}\hat{p}.$$

By Lemma 1, this price vector is in the polar cone of the individual privacy constraint for voter i, since $\mu_0(s_1^i) = \mu_0(s_0^i) = \frac{1}{2}$.

The total shadow price for a given signal profile θ and action a is the sum of these individual prices. Recall that $m(\theta)$ denotes the number of s_1 signals in profile θ. For action a_1, this gives:

$$p(\theta, a_1) := \sum_{i=1}^{n} p^i(s^i(\theta), a_1) = m(\theta)\hat{p} - (n - m(\theta))e^{\delta}\hat{p}, \tag{15}$$

and similarly, $p(\theta, a_0) := (n - m(\theta))\hat{p} - m(\theta)e^{\delta}\hat{p}$. By Lemma 1, since each individual price vector p^i is in the polar cone for the i-th constraint, their sum p is in the polar cone for the intersection of these constraints. Thus, the Price Validity condition of Theorem 3 is satisfied.

(Step 2: Solving the Priced Problem) Let $\Delta v(\theta) := v(\theta, a_1) - v(\theta, a_0)$. The planner's priced

problem from ($\mathcal{W}_p$) is equivalent to:

$$\max_{x(a_1|\theta)\in[0,1]} \quad \sum_{\theta} \left(\Delta v(\theta) - (p(\theta, a_1) - p(\theta, a_0))\right) x(a_1|\theta)\mu_0(\theta). \tag{16}$$

Substituting the prices from (15), the term in the parenthesis becomes $\Delta v(\theta) - (2m(\theta) - n)(1 + e^{\delta})\hat{p}$. The optimal rule is to set $x(a_1|\theta) = 1$ when this term is positive, $x(a_1|\theta) = 0$ when it is negative, and to randomize when it is zero.

Note that $\Delta v(\theta)$, which is given by Equation 3, depends only on $m(\theta)$, the number of s_1 signals. The function $\Delta v(m)$ is S-shaped: it is convex for $m < n/2$ and concave for $m > n/2$. The net privacy cost, $(2m - n)(1 + e^{\delta})\hat{p}$, is linear in m. For $\hat{p} > 0$, the S-shaped benefit and the linear cost can intersect at most three times. By symmetry, these intersections occur at some points $n - \hat{m}, n/2, \hat{m}$. This implies that the optimal rule must take the cutoff structure described in the proposition.

(Step 3: Pinning Down the Price and Uniqueness) The price level $\hat{p}$ (and thus the cutoff $\hat{m}$) is determined by the Complementary Slackness condition of Theorem 3. For each action a, this requires $\sum_{\theta} p(\theta, a)x(a|\theta)\mu_0(\theta) = 0$. By symmetry, it is sufficient to check this for one action, say a_1. The condition is $\sum_{\theta}\sum_{i} p^i(s^i(\theta), a_1)x(a_1|\theta)\mu_0(\theta) = 0$. Rearranging the summation gives $\sum_{i}\sum_{\theta} p^i(s^i(\theta), a_1)x(a_1|\theta)\mu_0(\theta) = 0$. Since the prices p^i and the rule are symmetric across voters, this holds if it holds for a single voter i. This is equivalent to:

$$p^i(s_1, a)P_x(a|s^i = s_1)\mu_0(s^i = s_1) + p^i(s_0, a)P_x(a|s^i = s_0)\mu_0(s^i = s_0) = 0,$$

which gives:

$$\log \frac{P_x(a_1|s^i = s_1)}{P_x(a_1|s^i = s_0)} = \delta.$$

Let $C : [\frac{n+1}{2}, n] \to \mathbb{R}$ be the correspondence mapping a cutoff $\hat{m}$ to the set of possible values of the log-likelihood ratio. It is a set when $\hat{m}$ is an integer so that $x(a_1|\hat{m})$ and $x(a_1|n - \hat{m})$ can take any value in $[0, 1]$. The correspondence is continuous because the conditional probabilities $\Pr(a_1|s^i)$ are continuous functions of the randomization probabilities at the integer cutoffs. It is increasing in the sense that if $\hat{m}' > \hat{m}$, then any selection from $C(\hat{m}')$ is greater than or equal to any selection from $C(\hat{m})$. This is because increasing the cutoff $\hat{m}$ expands the set of signal profiles with $m(\theta) > \frac{n}{2}$ for which a_1 is chosen and shrinks the set of signal profiles with $m(\theta) < \frac{n}{2}$ for which a_1 is chosen. Since we assume the privacy constraint is binding, we have $\min C(\frac{n+1}{2}) < 0 \leq \delta < \max C(n)$. By the intermediate value theorem for correspondences, there must exist a cutoff $\hat{m} \in [\frac{n+1}{2}, n]$ and a randomization rule that solves the binding constraint equation.

To see the uniqueness, note that any optimal solution $\hat{x}$ must also solve the priced problem (16). This is because if $\hat{x}$ is feasible and $v(\hat{x}) - \sum_{\theta,a} p(\theta,a)\hat{x}(a|\theta)\mu_0(\theta) < W(p) = v(x)$, then we must have $v(\hat{x}) < v(x)$ since by Price Validity we have $\sum_{\theta,a} p(\theta,a)\hat{x}(a|\theta)\mu_0(\theta) \leq 0$.[41] Therefore, the multiplicity can only arise because of different randomizations at $m(\theta) = \hat{m}$ or $m(\theta) = n - \hat{m}$. *Q.E.D.*

Proof of Corollary 1. (First statement) Let $\hat{m}(n)$ be the optimal cutoff and x_n the optimal voting rule when the number of voters is n. Let P_n denote the probability distribution induced by x_n. We first establish that as $n \to \infty$, the influence of a single voter's signal on the outcome probability, conditional on the state, vanishes. For any a, ω, s^i, we have:

$$\left|P_n(a|\omega, s^i) - P_n(a|\omega)\right| \leq \left|P_n(a|\omega, s^i = s_1) - P_n(a|\omega, s^i = s_0)\right|.$$

This is derived by unpacking $P_n(a|\omega)$ using law of total probability. The term on the right is the probability that voter i is pivotal, which under the cutoff rule x_n only occurs if the number of s_1 signals among the other $n-1$ voters is within one of the cutoffs. As $n \to \infty$, the probability of this event converges to zero. Therefore, we have $\lim_{n\to\infty} |P_n(a|\omega, s^i) - P_n(a|\omega)| = 0$.

Let $c := \liminf_n P_n(a_1|\omega_1)$. The planner's payoff converges to 1 if and only if $c = 1$. Let (n_k) be a subsequence such that $P_{n_k}(a_1|\omega_1) \to c$. First, consider the case where $\log \frac{q}{1-q} > \delta$. Suppose for contradiction that $c = 1$. Then, using the result above, we have:

$$\frac{P_{n_k}(a_1|s^i = s_1)}{P_{n_k}(a_1|s^i = s_0)} = \frac{qP_{n_k}(a_1|\omega_1, s^i = s_1) + (1-q)P_{n_k}(a_1|\omega_0, s^i = s_1)}{(1-q)P_{n_k}(a_1|\omega_1, s^i = s_0) + qP_{n_k}(a_1|\omega_0, s^i = s_0)}$$

$$\to \frac{qc + (1-q)(1-c)}{(1-q)c + q(1-c)} = \frac{q}{1-q}.$$

Since $\log \frac{q}{1-q} > \delta$, this implies that for all large k, the privacy constraint is violated, a contradiction. Therefore, we must have $c < 1$.

Next, consider the case where $\log \frac{q}{1-q} \leq \delta$. Suppose for contradiction that $c < 1$. This implies that for all large k, the voting rule is not first-best, which means the privacy constraint must be binding. However, the same calculation shows that the log-likelihood ratio converges to $\log \frac{qc+(1-q)(1-c)}{(1-q)c+q(1-c)} < \log \frac{q}{1-q} \leq \delta$. This means that for all large k, the privacy constraint is slack, a contradiction. Therefore, we must have $c = 1$.

(Second statement) When $\delta = 0$, the privacy constraint is always binding. We take the price vector p constructed in the proof of Proposition 3, and let x be an optimal cutoff voting rule

[41] We slightly abuse notation and use $v(x)$ to denote the expected payoff under x.

characterized in Proposition 3. From the proof of Proposition 3, we know that x solves the priced problem for the price vector p defined in (15), and that the Complementary Slackness condition holds, i.e., $\sum_{\theta,a} p(\theta,a)x(a|\theta)\mu_0(\theta) = 0$.

Now consider the constant voting rule $\overline{x}$ such that $\overline{x}(a_1|\theta) = 1/2$ for all θ, which yields a payoff of $1/2$. If $v^*(0;n) = 1/2$, then $\overline{x}$ must also be an optimal solution. We now argue that this cannot be the case. Note that for the price vector p, the constant rule $\overline{x}$ also satisfies the Complementary Slackness condition by definition of p when $\delta = 0$: $\sum_\theta p(\theta,a)\overline{x}(a|\theta)\mu_0(\theta) = 0$ for both actions. The value of the priced problem for x is $v(x) - \sum p(\theta,a)x(a|\theta)\mu_0(\theta) = v(x)$. The value of the priced problem for $\overline{x}$ is $v(\overline{x}) - \sum p(\theta,a)\overline{x}(a|\theta)\mu_0(\theta) = v(\overline{x}) = 1/2$. However, the function $\Delta v(m)$ is strictly S-shaped for $n \geq 3$, while the net privacy cost is linear. Therefore, a constant rule $\overline{x}$ cannot be a solution to the priced problem (16). This means that the value of the priced problem for x must be strictly greater than for $\overline{x}$:

$$v(x) > v(\overline{x}) = 1/2.$$

This shows that the constant rule is suboptimal, and thus $v^*(0;n) > 1/2$. *Q.E.D.*

C.5. Credit Screening

Proof of Proposition 4. I first show that the mechanism is optimal when the threshold q is taken to be the solution to

$$\sum_\theta \mu_0(\theta)\Big(v(F_\theta^{-1}(q),\theta)\cdot \mathbb{1}(q^*_\theta < q) + v(F_\theta^{-1}(e^\delta q),\theta)\cdot e^\delta \mathbb{1}(q^*_\theta > e^\delta q)\Big) = 0. \tag{17}$$

Note that the left-hand side of Equation 17 is strictly increasing and continuous in q on $[0,1]$ when $\delta \leq \delta^*$. Moreover, it is strictly negative when $q = 0$ and strictly positive when $q = 1$, so it admits a unique solution.

To certify the optimality of the proposed mechanism, we use Theorem 3. First note that since we assumed that $q^*_\theta \leq \frac{1}{2}$, we have $q_\theta \leq \frac{1}{2}$, and thus $\max_{\theta,\theta'} \log \frac{q_\theta}{q_{\theta'}} \leq \delta$ implies that $\max_{\theta,\theta'} \log \frac{1-q_\theta}{1-q_{\theta'}} \leq \delta$, so the proposed solution is feasible.

Define the shadow prices for the rejection action a_0 as $p(\theta,a_0) := -v(F_\theta^{-1}(q_\theta),\theta)$ and for the approval action as $p(\theta,a_1) := 0$. We verify the three conditions of Theorem 3, with $\lambda := \mu_0$.

(Priced Optimality) The priced objective for the designer is to choose quantiles $(\hat{q}_\theta)_{\theta\in\Theta}$ to maximize

$$\sum_\theta \mu_0(\theta)\left(\int_{\hat{q}_\theta}^1 v(F_\theta^{-1}(s),\theta)ds - p(\theta,a_0)\hat{q}_\theta\right).$$

The first-order condition with respect to $\hat{q}_\theta$ is $-v(F_\theta^{-1}(\hat{q}_\theta), \theta) - p(\theta, a_0) = 0$. By our choice of prices, this condition is satisfied at $\hat{q}_\theta = q_\theta$. Thus, the mechanism is a solution to the priced problem ($\mathcal{W}_p$). Moreover, any other choice of $\hat{q}_\theta$ will yield a strictly lower value, so the $(q_\theta)_{\theta\in\Theta}$ would be the unique solution once we certify its optimality.

(Price Validity) The prices $p(\cdot, a_1)$ are all zero and thus trivially in any polar cone. For $p(\cdot, a_0)$, we use Lemma 1. We must verify that $\sum_\theta p(\theta, a_0)^- \mu_0(\theta) \geq e^\delta \sum_\theta p(\theta, a_0)^+ \mu_0(\theta)$. By definition of q_θ and since $v(F_\theta^{-1}(t_\theta^*), \theta) = 0$:

- If $q_\theta^* < q$, then $q_\theta = q > q_\theta^*$. This implies $p(\theta, a_0) = -v(F_\theta^{-1}(q), \theta) < 0$.
- If $q_\theta^* \in [q, e^\delta q]$, then $q_\theta = q_\theta^*$. This implies $p(\theta, a_0) = -v(F_\theta^{-1}(t_\theta^*), \theta) = 0$.
- If $q_\theta^* > e^\delta q$, then $q_\theta = e^\delta q < q_\theta^*$. This implies $p(\theta, a_0) = -v(F_\theta^{-1}(e^\delta q), \theta) > 0$.

The Price Validity condition becomes

$$\sum_{\theta: q_\theta^* < q} v(F_\theta^{-1}(q), \theta)\mu_0(\theta) \geq e^\delta \sum_{\theta: q_\theta^* > e^\delta q} -v(F_\theta^{-1}(e^\delta q), \theta)\mu_0(\theta).$$

Rearranging this gives exactly Equation 17, which holds by definition of q. Thus, the prices are valid.

(Complementary Slackness) The condition for a_1 is trivially satisfied since $p(\cdot, a_1) = 0$. For a_0, we must show $\sum_\theta p(\theta, a_0)x(a_0|\theta)\mu_0(\theta) = 0$. This is $\sum_\theta p(\theta, a_0)q_\theta\mu_0(\theta) = 0$. Substituting the prices and quantiles gives:

$$-\sum_{\theta: q_\theta^* < q} \mu_0(\theta)v(F_\theta^{-1}(q), \theta)q - \sum_{\theta: q_\theta^* > e^\delta q} \mu_0(\theta)v(F_\theta^{-1}(e^\delta q), \theta)e^\delta q = 0$$

Factoring out q and we get Equation 17, which is zero. Thus, complementary slackness holds.

Finally, one directly checks that the solution to Equation 17, $q(\delta)$, is strictly decreasing in δ, while $e^\delta q(\delta)$ is strictly increasing in δ. *Q.E.D.*

D. Proofs of Section 5

Proof of Proposition 5. Proposition 5 follows directly from Theorem A.1 by taking $\mathcal{H}_{n,t_{-n}}$ as the aspects, where $\mathcal{H}_{n,t_{-n}}$ contains singletons $\{\theta\}$ such that $\theta_{-n} = t_{-n}$. *Q.E.D.*

Proof of Proposition 6. The proof proceeds by constructing a set of shadow prices for the DP constraint and showing that the proposed voting rule satisfies the conditions of Theorem 3.

(Step 1: Constructing the Shadow Prices) The DP privacy constraint is an intersection of individual constraints of the form $\log \frac{x(a|\theta')}{x(a|\theta)} \leq \delta$ for all neighboring pairs (θ, θ'). We construct the total shadow price $p(\theta, a)$ as the sum of the prices for each of these individual constraints.

Consider a pair of neighboring profiles (θ, θ') where $m(\theta') = m(\theta) + 1$. For this pair, we define a price vector $\hat{p}_{\theta,\theta'}$ that is non-zero only for states θ and θ'. We focus on the case where $m(\theta) \geq (n-1)/2$. The case of $n - m(\theta) \geq (n-1)/2$ can be symmetrically constructed by relabeling a_0 and a_1. For action a_0, we set:

$$\hat{p}_{\theta,\theta'}(\theta', a_0) = -e^{\delta}\hat{p}_{m(\theta')}/\mu_0(\theta'), \quad \hat{p}_{\theta,\theta'}(\theta, a_0) = \hat{p}_{m(\theta')}/\mu_0(\theta),$$

and zero for all other states and for $a = a_1$. The price levels $\hat{p}_m \geq 0$ will be determined later. By Lemma 1, this price vector is in the polar cone for the constraint between θ and θ'.

The total shadow price $p(\theta, a)$ is the sum of these pairwise prices over all neighbors of θ: $p(\theta, a) = \sum_{\theta' \mathcal{N} \theta} \hat{p}_{\theta,\theta'}(\theta, a)$. Given a profile θ with $m(\theta) \geq (n+1)/2$ signals of s_1, it has $m(\theta)$ neighbors with $m(\theta) - 1$ signals of s_1 and $n - m(\theta)$ neighbors with $m(\theta) + 1$ signals of s_1. The total price for action a_0 is:[42]

$$p(\theta, a_0) = ((n - m(\theta))\hat{p}_{m(\theta)+1} - m(\theta)e^{\delta}\hat{p}_{m(\theta)})/\mu_0(\theta).$$

$p(\theta, a_1)$ is defined symmetrically by relabeling a_0 and a_1. By Lemma 1, this total price vector p is in the polar cone of the intersection of all pairwise constraints, so the Price Validity condition is satisfied.

(Step 2: Solving the Priced Problem) The proposed solution x has full support for every θ. For such a rule to be optimal in the priced problem (16), the net benefit of choosing an action must be exactly offset by its net privacy cost. This gives the condition:

$$\Delta v(m(\theta)) = p(\theta, a_1) - p(\theta, a_0). \tag{18}$$

Substituting the expressions for the prices, this yields a system of recursive equations for the price levels $\hat{p}_m$. For $m > (n+1)/2$, we have $p(\theta, a_1) = 0$, so this gives:

$$\Delta v(m) = (me^{\delta}\hat{p}_m - (n - m)\hat{p}_{m+1})/\mu_0(m),$$

where $\mu_0(m)$ denotes the prior probability of any signal profile with m signals of s_1. These

[42] We adopt the convention that $\hat{p}_{(n-1)/2} = 0$.

equations uniquely pin down the price levels:

$$\hat{p}_n = \frac{\Delta v(n)\mu_0(n)}{ne^\delta}, \quad \hat{p}_m = \frac{\Delta v(m)\mu_0(m) + (n-m)\hat{p}_{m+1}}{me^\delta} \text{ for } m > \frac{n+1}{2}.$$

Finally, at $m = (n+1)/2$, Equation 18 pins down $\hat{p}_{(n+1)/2}$. By construction, the value of the priced problem is independent of the choice of x, so any rule with full support, including our proposed one, is a solution.

(Step 3: Verifying Complementary Slackness and Uniqueness) The Complementary Slackness condition requires $\sum_\theta x(a|\theta)p(\theta,a)\mu_0(\theta) = 0$. Since p is a sum of pairwise prices, it is sufficient to check the condition for each component. For a pair (θ,θ') with $m(\theta) \geq (n-1)/2$ and $m(\theta') = m(\theta)+1$, the condition for a_0 is:

$$\sum_{\hat{\theta}} x(a_0|\hat{\theta})\hat{p}_{\theta,\theta'}(\hat{\theta},a_0)\mu_0(\hat{\theta}) = x(a_0|\theta)\hat{p}_{m(\theta')} - x(a_0|\theta')e^\delta\hat{p}_{m(\theta')} = 0. \tag{19}$$

This holds if and only if $x(a_0|\theta') = e^{-\delta}x(a_0|\theta)$, which is precisely the form of the proposed solution. The case for $m(\theta) \leq (n+1)/2$ can be argued symmetrically.

For uniqueness, following the same argument as in the proof of Proposition 3, any optimal solution must solve the priced problem under our constructed price vector p and satisfy Complementary Slackness. When $\delta > 0$, the set of relations given by (19) for all neighboring pairs, combined with the fact that probabilities must sum to one, forms a system of equations whose unique solution is the voting rule proposed in the proposition. In particular, the solution must satisfy the pivot conditions $x(a_1|(n+1)/2) = e^\delta x(a_1|(n-1)/2)$ and $x(a_0|(n-1)/2) = e^\delta x(a_0|(n+1)/2)$, and then decay exponentially towards the extremes. This structure admits a unique solution. When $\delta = 0$, any constant voting rule satisfies Complementary Slackness and is thus optimal.

Finally, we consider the welfare implications. As n gets large, we have that $m/n \to q$ a.s. conditional on ω_1, and $m/n \to 1-q$ a.s. conditional on ω_0. The optimal solution therefore implies that $P_x(a_1|\omega_1) \to 1$ and $P_x(a_0|\omega_0) \to 1$ as $n \to \infty$ when $\delta > 0$. This means that $v^*(\delta;n) \to 1$. When $\delta = 0$, any uninformative rule is optimal, so $v^*(0;n) = 1/2$. *Q.E.D.*

Proof of Proposition 7. (First statement) We verify that L satisfies the properties of a worst-case privacy measure that is $\cup_{\gamma\in\Gamma}\mathcal{C}_\gamma$-monotonic. First, we have $L \geq 0$ and $L(x) = 0$ when x is uninformative. If x dominates x' for $\cup_{\gamma\in\Gamma}\mathcal{C}_\gamma$, then x dominates x' for each $\mathcal{C}_\gamma$. Therefore, we have $L_\gamma(x) \geq L_\gamma(x')$ for all $\gamma \in \Gamma$ and thus $L(x) \geq L(x')$. Let $x = \alpha x' + (1-\alpha)x''$ where

$\alpha \in (0,1)$ and $\text{Supp}(x') \cap \text{Supp}(x'') = \emptyset$. Then we have

$$L(x) = \sup_{\gamma \in \Gamma} L_\gamma(x) = \sup_{\gamma \in \Gamma} \max\{L_\gamma(x'), L_\gamma(x'')\}$$

$$= \max\left\{\sup_{\gamma \in \Gamma} L_\gamma(x'), \sup_{\gamma \in \Gamma} L_\gamma(x'')\right\} = \max\{L(x'), L(x'')\}.$$

(Second statement) We use a constructive approach. We denote $\hat{x} \succeq_\gamma x$ if $\hat{x}$ dominates x in $\mathcal{C}_\gamma$. Define

$$L_\gamma(x) := \inf_{\hat{x}: \hat{x} \succeq_\gamma x} L(\hat{x}).$$

By definition, $L_\gamma \geq 0$, $L_\gamma(x) = 0$ when x is uninformative, and L_γ satisfies $\mathcal{C}_\gamma$-Monotonicity.

Next, we show L_γ satisfies Worst-Case Protection. Let $x = \alpha x' + (1-\alpha)x''$ for some $\text{Supp}(x') \cap \text{Supp}(x'') = \emptyset$ and $\alpha \in (0,1)$. We first show $L_\gamma(x) \leq \max\{L_\gamma(x'), L_\gamma(x'')\}$. Take any $\hat{x}', \hat{x}''$ such that $\hat{x}' \succeq_\gamma x'$ and $\hat{x}'' \succeq_\gamma x''$. We only need to show $L_\gamma(x) \leq \max\{L(\hat{x}'), L(\hat{x}'')\}$. To see this, note that

$$\hat{x} := \alpha\hat{x}' + (1-\alpha)\hat{x}'' \succeq_\gamma \alpha x' + (1-\alpha)x'' = x.$$

This implies $L_\gamma(x) \leq L(\hat{x}) = \max\{L(\hat{x}'), L(\hat{x}'')\}$. For the other direction, take any $\hat{x}$ such that $\hat{x} \succeq_\gamma x$. Since $\text{Supp}\,\tau_x = \text{Supp}\,\tau_{x'} \cup \text{Supp}\,\tau_{x''}$, we have $\hat{x} \succeq_\gamma x'$ and $\hat{x} \succeq_\gamma x''$. This implies that $L(\hat{x}) \geq \max\{L_\gamma(x'), L_\gamma(x'')\}$. Taking infimum across all such $\hat{x}$, we conclude that $L_\gamma(x) \geq \max\{L_\gamma(x'), L_\gamma(x'')\}$.

Finally, we show that $L(x) = \max_{\gamma \in \Gamma} L_\gamma(x)$ for all $x \in \mathcal{X}_f$. By definition, we have $L_\gamma(x) \leq L(x)$, so $\max_{\gamma \in \Gamma} L_\gamma(x) \leq L(x)$. For the other direction, to the contrary, suppose $L(x) > \max_{\gamma \in \Gamma} L_\gamma(x)$ for some $x \in \mathcal{X}_f$. This means that for each $\gamma \in \Gamma$, there exists $x_\gamma \in \mathcal{X}_f$ such that $x_\gamma \succeq_\gamma x$ and $L(x_\gamma) < L(x)$. Define:

$$\hat{x} := \sum_{\gamma \in \Gamma} \frac{1}{|\Gamma|} x_\gamma.$$

Since L satisfies Worst-Case Protection, we have $L(\hat{x}) = \max_{\gamma \in \Gamma} L(x_\gamma) < L(x)$. However, by definition, we have $\hat{x} \succeq_\gamma x$ for all $\gamma \in \Gamma$. By $\cup_{\gamma \in \Gamma} \mathcal{C}_\gamma$-Monotonicity, we have $L(\hat{x}) \geq L(x)$. This is a contradiction, so we must have $L = \max_{\gamma \in \Gamma} L_\gamma$. *Q.E.D.*