Binary Mechanisms under Privacy-Preserving Noise

Farzad Pourbabaee and Federico Echenique

May 2024

Abstract

We study mechanism design for public-good provision under a noisy privacy-preserving transformation of individual agents' reported preferences. The setting is a standard binary model with transfers and quasi-linear utility. Agents report their preferences for the public good, which are randomly "flipped," so that any individual report may be explained away as the outcome of noise. We study the tradeoffs between preserving the public decisions made in the presence of noise (noise sensitivity), pursuing efficiency, and mitigating the effect of noise on revenue.

Pourbabaee (far@caltech.edu) is at the Division of the Humanities and Social Sciences, Caltech. Echenique (fede@econ.berkeley.edu) is at the Department of Economics, UC Berkeley.

Contents

1 Introduction

The field of mechanism design considers agents who hold private information about their preferences. Agents are asked to surrender this information when properly incentivized; but, traditionally, mechanism design ignores any potential *privacy concerns* that may add to agents' reluctance to reveal their true preferences.

Privacy concerns may arise because agents have an intrinsic "non-instrumental" aversion to revealing their preferences, or because agents worry that such information can be used against them in future interactions. This may occur in the provision of both private and public goods. With private goods, an agent who surrenders their willingness to pay for a good to a seller will lose surplus in future interactions with this seller.

In public-goods settings, revealing willingness-to-pay today may imply higher future taxes for related public goods tomorrow. For example, revealing a high value for a playground today may reveal a high value for a public library in the future. Agents may also have non-instrumental preferences for privacy: public-health-related projects often involve sensitive information about agents' likelihood of being susceptible to disease (think, for example, of a cancer screening program). In such cases, we want to know when we can preserve individuals' privacy while minimally compromising the optimality of a public choice rule.

Our paper studies a specific mechanism design problem: a planner faces a standard binary public-good provision problem with quasilinear utility and monetary transfers. Our planner cares about individuals' preferences for the public good, and about the revenue they can collect from its provision (alternatively, the extent to which they need to subsidize the public good). In our version of the problem, privacy concerns are important, and dealt with using an embedded privacy-preserving operation.

There are a variety of proposals to protect individuals' privacy. Our approach in this paper involves adding random noise to the individuals' messages, using an *in-place* randomization device, before they reach the planner. In effect, the planner's mechanism takes as inputs the agents' perturbed reports — so it can only access the agents' reported types after they have been subject to a privacy-preserving random transformation. The idea follows the literature on *differential privacy* initiated by Dwork, McSherry, Nissim, and Smith (2006), (see also Dwork et al., 2006; Dwork, 2008; Dwork and Roth, 2014), by which individuals' privacy is preserved through the addition of random noise.

There are other privacy-preserving methods as well. For example, the method of de-identification, which involves removing names and personal identifying information from messages. However, as shown in a recent study by Evans and King (2023), de-identification

in the context of political surveys is vulnerable to re-identification attacks, compromising the privacy of survey respondents.[1]

Other methods of privacy protection require the availability of a trusted intermediary. This intermediary may use cryptographic tools. In this approach, agents report their types to the intermediary, who aggregates the information and passes along a privacy-preserving aggregate decision to the planner. The intermediary may be automated, and consist of a cryptographic algorithm, but still needs to be trusted by the agents involved. The downside of this method is that assuming the existence of a trusted intermediary can be quite restrictive. Additionally, such intermediaries may act strategically, and extract rents from the agents. For example, when agents reveal willingness to pay for a public good, a trusted and benevolent principal with imperfect commitment ability may extract future rent from some individual agents.

Aside from evaluating the pros and cons of other methods, our analysis in this paper focuses on the positive aspects. We believe that understanding the tradeoffs in a noisy environment is a valuable exercise in its own right.

In our paper, agents' types are binary and encode how much utility they receive from the public good: high or low. Types take the value $+1$ or -1. The provision outcome is also binary, i.e, $\{0, 1\}$-valued. Binary decisions are common in public goods environments because public goods are often about implementing a large indivisible project (a library, a bridge, a waste disposal facility, etc.). We focus on this binary setting, because the basic tradeoffs are captured by a yes/no decision. Therefore our model gets at the heart of the matter, while remaining tractable.

The random noise is then simply a "flip," which occurs with probability $\delta \in (0, 1/2)$. If an agent reports a type $x_i \in \{-1, +1\}$, then the mechanism receives x_i with probability $1 - \delta$, and a "flipped" report $-x_i$ with probability δ. As a consequence, an agent can always explain away any evidence about their type as the outcome of a random flip. Their explanation is more credible the larger the value of δ. Noise is then desirable because larger values of δ offer a better protection of privacy.

The problem with adding noise to the agents' reports—one might say the flip side—is, of course, that the quality of the planner's decision suffers. So we consider the probability that the planner's decision is affected by the noise we have added for reasons of privacy. A key concept is *noise sensitivity:* the probability that the planner's decision differs from what it

[1]De-identification has been ineffective in other areas as well. For example, in a study by Gymrek et al. (2013), researchers successfully re-identified individuals from the 1000 Genomes Project by cross-referencing their data with other publicly available online resources.

would have been, given the true and noise-free reports. In addition to standard considerations in mechanism design (such as revenue and surplus), our paper evaluates mechanisms on the basis of their noise sensitivity. We further show that noise sensitivity is closely tied to the distortions in social surplus caused by random transformations of the agents' reports.

Noise affects transfers, as well as the public-good provision decision. In consequence, the planner's expected revenue suffers. Standard ideas in mechanism design mean that an agent with a low value for the public good (a -1 type) pays less than an agent with a high value (a $+1$ type). With noise, it is possible that a low type has their truthful report flipped, and is thus subject to the higher payment designed for high types. This, in turn, affects the whole problem by means of the low types' participation constraints. The end result is lower revenue for the planner as a whole. In sum, as the level of noise δ increases, revenue and social surplus decrease. There is thus a tradeoff between the privacy protection afforded by noise, and the effectiveness of a given social choice function in terms of traditional economic objectives.

Our first main theorem concerns choosing a mechanism to minimize noise sensitivity (or equivalently, surplus distortion), given a target level of revenue and a fixed level of privacy-preserving noise. The resulting optimization is not convex, which presents a challenge, but we are able to characterize the optimal mechanisms asymptotically, as the number of agents grows. They take the form of linear threshold functions (basically implementing the public good once the number of "votes" in favor, or high types, exceeds those against by a certain margin). In our second main theorem, we characterize the mechanism that optimizes social surplus under a target revenue constraint, and a fixed noise level. The optimal provision rule is a linear threshold function and coincides with one of the optimal solutions from the earlier problem. We find the provision threshold as a function of the required revenue and the noise level. Moreover, it is shown that the lower the provision threshold, the higher the social surplus, and the smaller the revenue is. Together, these two theorems examine the *level* and *distortions* of social surplus under a revenue constraint across varying levels of privacy-preserving noise.

A key tradeoff in our paper involves noise sensitivity and revenue. A planner can make the mechanism more robust to noise (improve its noise sensitivity) at the cost of lower revenue. In sum, our paper describes a planner who balances several different objectives: privacy, efficiency, robustness and revenue.[2] The different tradeoffs involved are characterized through the theory developed in the paper.

Our model has two other interpretations, in addition to the emphasis on privacy that we

[2]We use the revenue terminology throughout the paper, but one may of course think of the objective as minimizing the amount of subsidy needed for the public good project.

have focused on so far. First, agents may be unable to perfectly communicate their preferences to the mechanism. Miscommunication has been documented experimentally (see Budish and Kessler, 2022), and pushed as an agenda by, for example, McFadden (2009). Second, agents may have imperfect information about their own preferences. Their reports are therefore only noisy versions of their underlying values for the public good. Imperfect knowledge of preferences has been considered a key motivation for studying information acquisition in mechanism design. Two recent examples are Gleyze and Pernoud (2022) and Thereze (2022). To summarize, the one formal framework that we introduce and study provides insights about three important environments.

Related Literature. We are not the first to study mechanism design together with a device for ensuring privacy. There is a literature on mechanism design and differential privacy. The first paper is McSherry and Talwar (2007), who shows that differential privacy can be a useful tool in obtaining incentive compatibility. By dampening the effect that any individual report has on the mechanism's decision, differential privacy can help ensure truthful behavior among agents. Nissim et al. (2012) develop these ideas in a construction that achieves approximately optimal virtual implementation. Their focus is therefore closer to the problem of full implementation, and not the standard mechanism design problem. Huang and Kannan (2012) propose mechanisms that are both incentive compatible and differentially private, but does not incorporate the analysis of the tradeoffs that are the focus of our paper. The works of Nissim et al. (2012), Xiao (2013), and Chen et al. (2016) all consider preferences over privacy explicitly in their mechanism design analysis. This is of course an important direction, but not the one we pursue here. Nissim and Xiao (2015) provide an overview of the literature on mechanism design and differential privacy.

Our paper is also related to recent works on monopolistic screening with privacy concerns (Eilat et al., 2021; Krähmer and Strausz, 2023). In the first paper, the privacy loss — measured by the Kullback-Leibler divergence between planner's prior and posterior belief about the buyer's type — is set as a constraint for the screening problem. Specifically, in this work the privacy is protected by selecting the message space as the partitions of the original type space (i.e., coarsening the type set). Hence, the message sent by the agent does not fully resolve the underlying type, thus protecting their privacy. In our binary setting, noisy flips are more natural than a partition of the type space, which is too blunt when there are only two types. The second paper reflects privacy concerns in the buyer's preference, much like the literature we discussed above. Neither of the papers address the tradeoffs that we focus on, or the issues regarding robustness.

The idea of adding noise as a means for privacy protection is very common in other areas as well (e.g., see Geng and Viswanath, 2015; He et al., 2018, for applications in communication and information theory). In political science Warner (1965) introduced the randomized response method as a survey technique, that asks respondents to use in-place randomization device to conceal their sensitive answers from the interviewer — Blair et al. (2015) summarizes the use of this method in this area. Since other methods of privacy protection (such as clean rooms and de-identification) have been shown to fail, differential privacy through the addition of calibrated noise gained traction in political science. In a sequence of studies by Evans et al. (2019), Evans et al. (2022) and Evans and King (2023) this method is shown to help social scientist to study the vast amount of user data owned by governments and companies while maintaining privacy issues. For example, the last US Census issued by the government is being released with noise.[3] Companies also use open source softwares that allow researchers to test their algorithms while concealing the private data of their users through the addition of statistical noise.[4]

Our model of public good provision with privacy-protection concerns is also formally equivalent to a setting in which agents cannot perfectly report their preferences to the planner. In that sense our paper is a theoretical contribution to a mostly empirical literature that documents preference *misrepresentation* in incentive compatible environments because of variety of reasons such as cognitive limitations or simply lack of perfect communication between participants and the planner. In his tribute to Hurwicz and Laffont, McFadden (2009) states that "in reality, mistakes that agents make in processing and drawing inferences from communications and information, and in exercising control and responding to incentives, can undermine the ideal efficiency of mechanisms, making it important to consider the robustness of mechanisms involving human agents."

A growing body of literature in applied mechanism design documents preference misrepresentation. For example, Hassidim et al. (2017) and Hassidim et al. (2021) show that students misreport their funding preferences when applying to graduate programs, despite the fact that the underlying matching mechanism is strategy-proof (in this case it is Deferred-Acceptance). In the context of residency matching mechanisms Rees-Jones (2018) and Rees-Jones and Skowronek (2018) present evidences that some students make futile attempts misrepresenting their preference ranking. In an experiment Budish and Kessler (2022) show that students fail to report their preferences accurately enough in a course scheduling mechanism.

[3]See https://www2.census.gov/about/policies/2019-11-paper-differential-privacy.pdf.
[4]See https://news.microsoft.com/on-the-issues/2020/08/27/statistical-noise-data-differential-privacy.

2 Model

2.1 Mechanisms

We consider the problem of providing a public good in an economy with n agents and quasilinear preferences. The decision is binary: a public good is either provided or not. Agents' types, which are denoted by $x_i \in \{-1, +1\}$, encode their value for the public good. An individual with a low (respectively, high) type has low (respectively, high) valuation for the public good. Ideally, a decision on whether to provide the public good is based on agents' realized types, but these are private information. We have access to monetary transfers that may be used to incentivize agents in reporting their types. The assumption of binary types is, of course, restrictive, but crucial for the methodology employed in our paper; it also offers the simplest framework for studying a public choice decision with heterogeneous and private preferences over the provision of the public good.

We focus on direct-revelation mechanisms. A (direct-revelation) public-good mechanism consists of an *allocation rule* $f : \{-1, +1\}^n \to \{0, 1\}$, and n *transfer rules*, denoted by $t_i : \{-1, +1\}^n \to \mathbb{R}$ for all $i \in [n]$. The allocation rule f takes in the $\{-1, +1\}$ messages sent by the individuals, and returns the provision decision, where an output of 1 means the public good is being provided, and a 0 output means otherwise. Often in the paper we call an allocation rule a *social choice function* (SCF).

2.2 Quasilinear Preferences

A profile of types $(x_1, \ldots, x_n)$ is drawn i.i.d. from the uniform distribution on $\{-1, +1\}$.[5] Individuals have quasilinear preferences over the final allocation and the transfer. Specifically, the utility of individual i, with type x_i, from $(f, t_i) \in \{0, 1\} \times \mathbb{R}$ is

$$u_i(f, t_i; x_i) = \left(\frac{b + x_i}{2}\right) f - t_i \,. \tag{2.1}$$

The parameter $b \in [0, 1]$ captures a possible bias in favor of the public good. For example, when $b = 1$, the efficient outcome is to always provide the public good, and when $b = 0$, the

[5] The measure does not need to be uniform. In fact, it is possible to change the type domain to any other bi-valued set with un-even probability — that just requires some scaling and normalization. For example, if the type space is $\{\ell, h\}$ with probabilities $\{p, 1-p\}$, we can can transform the preference bias from b to b' and the range of f from $\{0, 1\}$ to $\{0, M\}$ so that transformed problem becomes isomorphic to the original problem. We chose the convention of the uniform measure over $\{-1, +1\}$ because it is standard in the Boolean function literature.

preferences for the public good are *symmetric* around zero, and the efficient outcome coincides with a simple majority decision. The negative sign before t_i means that the transfers are from the individuals to the planner.

2.3 Noisy Reports

A key innovation in our paper is the presence of noisy preference reports. Specifically, we assume the message $m_i \in \{-1, +1\}$ sent by individual i is going to flip to $-m_i$ with probability $\delta \in (0, 1/2)$. We assume these flips are independent across all individuals, and refer to δ as the *noise probability*. Agents can explain away any information about their type as the result of these random flips. Such explanations are more credible the larger the value of δ. Noise in our model is a basic implementation of differential privacy (Dwork et al., 2006; Dwork and Roth, 2014). When δ is close to 1/2, each individual agent's report is approximately uniformly distributed on $\{-1, +1\}$, regardless of their actual report.

A basic inspiration for differential privacy is the model of "randomized response" used in survey studies in the social sciences, see Chapter 2 in Dwork and Roth (2014). In our setting, let m_i be the message sent by agent i, and y_i be the signal received by the planner. Then, following the language of differential privacy, this communication mechanism will be *ε-differentially private* if

$$\mathrm{e}^{\varepsilon} \geqslant \max\left\{\frac{\mathsf{P}\left(y_i = +1 | m_i = +1\right)}{\mathsf{P}\left(y_i = +1 | m_i = -1\right)}, \frac{\mathsf{P}\left(y_i = -1 | m_i = +1\right)}{\mathsf{P}\left(y_i = -1 | m_i = -1\right)}\right\} = \frac{1-\delta}{\delta}. \tag{2.2}$$

It essentially means that by observing the received signal, the planner cannot distinguish the transmitted message with high precision. Thus by setting $\delta = (1 + \mathrm{e}^{\varepsilon})^{-1}$ our model guarantees an ε-differentially private mechanism.

An alternative explanation for noisy preference reporting is that communication from the agents to the social planner can be lossy and imperfect; hence, random flips capture imperfect communication between the agents and the planner. If the privacy interpretation of our model makes sense when δ is large, the lossy communication interpretation makes most sense when δ is small.

2.4 Quantifications

In this section, we briefly discuss the implementability notions and the main quantities that we use in the paper. Subsequently, in the following sections, we delve deeper into their formal definitions

The first quantity that we introduce is *noise sensitivity*. In a Bayes-Nash incentive compatible (BN-IC) mechanism, each agent reports their true type $x_i \in \{-1, +1\}$, but as a result of noisy preference reporting, the planner receives $y_i \in \{-1, +1\}$, where $\mathsf{P}(x_i \neq y_i) = \delta$. Observe that the pairs (x_i, y_i) are i.i.d. over $i \in [n]$. In particular, $x = (x_1, \ldots, x_n) \sim \text{Unif}\big(\{-1, +1\}^n\big)$ and $y = (y_1, \ldots, y_n)$ is the noisy version of x received by the planner. This means that the implemented outcome that was supposed to be $f(x)$, now changes to $f(y)$. If $f(x)$ is the desired decision regarding the public good, we may be concerned that $f(y) \neq f(x)$. The probability that this occurs is termed the noise sensitivity of the SCF f.

Specifically, *noise sensitivity* is defined as

$$\mathsf{NS}_\delta[f] = \mathsf{P}\big(f(x) \neq f(y)\big)\,.$$

Noise sensitivity is a standard variable in the analysis of Boolean functions (see, e.g., the pioneering work by Benjamini et al. (1999) and the comprehensive treatment in O'Donnell (2014)). We believe that this quantity is important in and of itself. For one, the planner does not want to pick an allocation rule that frequently takes the individuals by surprise. This would affect the credibility and commitment power of the planner.

Second, in Section 5, we show that as the size of the economy grows, the distortions in social surplus caused by differential privacy noise can be closely approximated by the noise sensitivity (see, in particular, Proposition 3). Third, increasing the noise level δ adds to the privacy preservation power of the mechanism, at the expense of making the SCF more sensitive to the noise. Studying the dependence of noise sensitivity on δ quantifies the tradeoff between privacy and the ensuing distortion.

Next, we present the other two quantities: social surplus and revenue. Suppose that, by refusing to participate in the mechanism, any individual can guarantee themselves a utility of zero. A mechanism that respects the interim individual rationality constraint is referred to by IIR. We say that a SCF f is Bayes-Nash implementable if it is both BN-IC and IIR. For such a SCF we refer to its expected social surplus by $\mathsf{S}_\delta[f]$, and to the maximum expected revenue by $\mathsf{R}_\delta[f]$. The expectation and probability operators are with respect to the joint distribution of x and y.

3 Main Results

We introduced three main quantities in the previous section: expected social suplus $\mathsf{S}_\delta[f]$, maximum expected revenue $\mathsf{R}_\delta[f]$, and noise sensitivity $\mathsf{NS}_\delta[f]$. Understanding these quanti-

ties in a finite economy is very challenging, but we shall see that the problem is tractable in a large economy. We now proceed with the two main results in the paper, which examine the tradeoffs between these quantities as the number of individuals grows large (i.e., $n \to \infty$).

3.1 Revenue and Noise Sensitivity

The concerns for robustness in the presence of reporting noise motivates a natural optimization problem. Among the set of all implementable allocation rules that extract a target level of expected revenue (say R), which ones have the minimum noise sensitivity (or maximum noise robustness)? Formally, we seek the solution to the following optimization problem:

$$\begin{aligned} &\min_{f} \mathsf{NS}_{\delta}[f] \\ &\text{subject to: } \mathsf{R}_{\delta}[f] \geqslant R \text{ and } f \text{ being implementable}\,. \end{aligned} \tag{3.1}$$

The solution to problem (3.1) characterizes the tradeoff between privacy and expected revenue in public-good mechanisms. Specifically, raising the noise level δ provides higher privacy, but increases the noise sensitivity, and (as will be shown later) decreases revenue. Fixing the noise level δ, thereby guaranteeing a certain privacy preservation level, the above program outputs the SCF that raises the target revenue R and is maximally robust against the privacy-preserving noise induced through δ.

Obtaining a closed-form solution to Problem (3.1) is not tractable, but we can make progress under the assumption of large n.

Before stating the solution to Problem (3.1), we state some notational conventions.

Notation 1. We use $\varphi(\cdot)$ and $\Phi(\cdot)$ to respectively denote the density and cumulative distribution function of the standard Gaussian. Also, we denote the inverse function of the Gaussian density (taking values in $\mathbb{R}_+$) by φ^{-1}, and the inverse function of the Gaussian cumulative function by Φ^{-1}. We further denote the sum of individuals' types by $\nu_n(x) \coloneqq \sum_{i=1}^{n} x_i$, and sometimes drop x from the argument of $\nu_n(\cdot)$.

Notation 2. We normalize the target revenue R, and define $r \coloneqq R/(1-2\delta)\sqrt{n}$. The optimal value of the program in (3.1) is denoted by $\mathcal{V}_n(r)$.

Definition 1. A mapping from $\{-1,+1\}^n$ to $\{0,1\}$ is called a *linear threshold function* (LTF), if there exists some threshold τ, such that $f(x) = \mathbf{1}\{\nu_n(x) \geqslant \tau\}$. In the following, we mainly

work with two LTFs:

$$\bar{\ell}_n(x;r) \coloneqq \mathbf{1}\left\{\frac{\nu_n(x)}{\sqrt{n}} \geqslant \varphi^{-1}(r) + o(1)\right\}, \tag{3.2a}$$

$$\underline{\ell}_n(x;r) \coloneqq \mathbf{1}\left\{\frac{\nu_n(x)}{\sqrt{n}} \geqslant -\varphi^{-1}(r) + o(1)\right\}, \tag{3.2b}$$

where, as usual, $o(1)$ denotes a term that vanishes as $n \to \infty$.

Theorem 1. *The linear threshold functions* $\{\underline{\ell}_n(\cdot;r), \bar{\ell}_n(\cdot;r)\}$ *are asymptotically optimal choices for the revenue constrained noise sensitivity minimization problem in* (3.1)*. Formally,*

$$\mathcal{V}_n(r) \leqslant \mathsf{NS}_\delta[\ell_n] \leqslant \mathcal{V}_n(r) + o(1), \quad \textit{for } \ell_n \in \{\underline{\ell}_n(\cdot;r), \bar{\ell}_n(\cdot;r)\}. \tag{3.3}$$

In the following sections, we argue that the simple majority rule (a LTF with 0 threshold, or 50% of the votes) raises the maximum revenue, but if one wants to improve upon its noise sensitivity, then by Theorem 1 the optimal way, among all implementable Boolean functions, is to increase the 50% threshold of the majority function (or decrease it by a similar amount). The more one increases (or decreases) this threshold, the more noise robustness is gained and more expected revenue is lost. The optimal tradeoff is struck by the LTFs $\{\underline{\ell}_n(\cdot;r), \bar{\ell}_n(\cdot;r)\}$.

In Figure 1, we plot the maximum achievable noise robustness, i.e., $1 - \mathsf{NS}_\delta$, on the y-axis, given the revenue level on the x-axis. We use the normalized expected revenue (by $\sqrt{n}$ not $(1-2\delta)\sqrt{n}$). The figure indicates the asymptotic Pareto frontier, for three different noise levels, that are achieved by the LTFs in Theorem 1. The figure suggests that, as the noise *increases*, the frontier becomes *steeper*; meaning that giving up a fixed level of revenue can lead to greater robustness against noise, and this tradeoff is amplified in higher noise levels where the privacy protection is stronger.

Put differently, the maximum achievable noise robustness is decreasing with respect to both revenue and level of noise. However, these two variables act as *substitutes*. That is, lowering the required revenue is more effective for gaining noise robustness at higher levels of noise.

3.2 Revenue and Surplus

Our second main result deals with the tradeoff between revenue and surplus. Specifically, we ask and answer the following question: For a fixed level of privacy noise δ, and among all the implementable SCFs that raise a target expected revenue (say R), which SCF has the

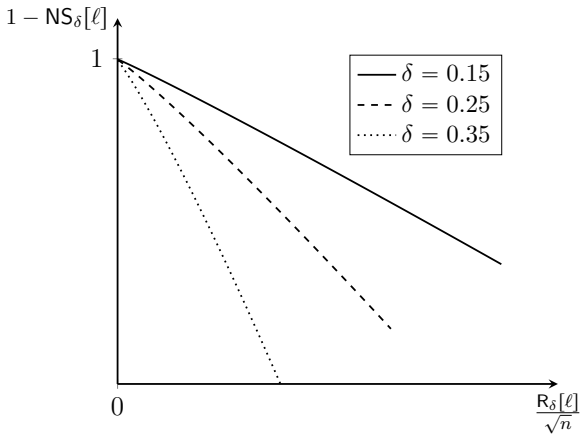

Figure 1: Asymptotic Pareto Frontier

highest social surplus? Formally, we study the following optimization problem:

$$\begin{aligned} &\max_f \mathsf{S}_\delta[f] \\ &\text{subject to: } \mathsf{R}_\delta[f] \geqslant R, \text{ and } f \text{ being implementable}. \end{aligned} \tag{3.4}$$

Theorem 2. *For large enough n, the optimal solution in the revenue/surplus tradeoff in* (3.4) *is the linear threshold function $\underline{\ell}_n(\cdot; r)$ expressed in* (3.2b).

In contrast with Theorem 1, Theorem 2 provides an allocation rule that is *exactly* optimal when n is large enough, whereas the candidates in Theorem 1 are *asymptotically* optimal, as stated in equation (3.3).

Theorem 2 states that, if one is willing to sacrifice some revenue (compared to the maximum achievable under the majority rule), the optimal approach to maximize expected social surplus is to lower the majority threshold below 50%. The lower the provision threshold, the higher the expected social surplus, and the smaller the expected revenue. Additionally, this tradeoff is optimally struck by the threshold function $(R, \delta) \mapsto -\varphi^{-1}(R/(1-2\delta)\sqrt{n})$, which is the provision threshold in $\underline{\ell}_n(\cdot; r)$.

Importantly, fixing a target revenue level R, one observes that securing the mechanism by increasing the noise level δ, raises the provision threshold (meaning the public-good is provided with smaller ex-ante probability, so less often), and thus lowers the expected social surplus. This theorem also quantifies the tradeoff between gaining privacy (by increasing δ) and losing social surplus (by raising the provision threshold) at a fixed revenue level.

Remark 1. In Theorem 2, we maximize the *level* of the social surplus subject to a revenue

constraint at a fixed noise level. In Section 5, we demonstrate that in large economies, the distortion in social surplus closely tracks the corresponding noise sensitivity. Consequently, Theorem 1 can be interpreted as a program that minimizes the distortions in social surplus while maintaining a revenue constraint at a fixed noise level. Together, these two theorems examine the *level* and *distortions* of social surplus under a revenue constraint across varying levels of privacy-preserving noise.

The plan in the following sections is to develop the necessary tools for proving the above claims. In the process, we will also provide additional insights and methods for examining the effects of noise on mechanisms' performance, offering a more comprehensive understanding of the topic.

In Section 4, we study the incentive compatibility and individual rationality of the mechanisms in the noisy environment. Then, in Section 5 we study the impact of noise in the allocation rule. Specifically, we present a self-contained introduction to the Fourier analysis of Boolean functions, and use it to study the comparative statics of social surplus and revenue with respect to the noise. Having introduced the required tools, we prove the theorems in Section 6. Finally in Section 7, we investigate the implications of our methodology in an environment where noise is used to represent agents' imperfect knowledge of their own preferences, rather than as a privacy protection measure.

4 Implementation under Noise

In this section, we study the incentive compatibility and individual rationality in the presence of noise. We show how the space of implementable SCFs vary with respect to the noise, and we offer a version of revenue equivalence theorem for any implementable SCF.

4.1 Incentive Compatibility

Suppose individual i reports message m_i to the planner. Denote the received message (that is subject to noise) by $y_i(m_i)$, so that $y_i(m_i) = m_i$ with probability $1-\delta$, and $y_i(m_i) = -m_i$ with probability δ. Also, let us denote the vector of reported types by $m = (m_1, \dots, m_n)$, and the vector of true types by $x = (x_1, \dots, x_n)$. A mechanism $(f, t_1, \dots, t_n)$ is Bayes-Nash

incentive compatible (BN-IC), if for every $i \in [n]$ and $x_i \in \{-1, +1\}$ one has

$$\begin{aligned}\left(\frac{b+x_i}{2}\right) \mathsf{E}\left[f\big(y_i(x_i), y_{-i}(x_{-i})\big)\big|x_i\right] - \mathsf{E}\left[t_i\big(y_i(x_i), y_{-i}(x_{-i})\big)\big|x_i\right] \geqslant \\ \left(\frac{b+x_i}{2}\right) \mathsf{E}\left[f\big(y_i(-x_i), y_{-i}(x_{-i})\big)\big|x_i\right] - \mathsf{E}\left[t_i\big(y_i(-x_i), y_{-i}(x_{-i})\big)\big|x_i\right],\end{aligned} \tag{4.1}$$

where the expectations are taken with respect to x_{-i} and their flips, namely $y_{-i}(x_{-i})$, as well as the noise in $y_i(x_i)$. To reduce clutter, we use y_j instead of $y_j(x_j)$, and similarly, y_{-j} instead of $y_{-j}(x_{-j})$. Also, as a shorthand, for every function $g : \{-1, +1\}^n \to \mathbb{R}$, define $\bar{g}_i(x_i) \coloneqq \mathsf{E}\left[g(x_i, x_{-i})\big|x_i\right]$.

In the following lemma, we characterize the space of all BN-IC direct mechanisms.

Lemma 1 (BN-IC)**.** *A mechanism consisting of the allocation rule f and the transfer functions $t = (t_1, \ldots, t_n)$ is Bayes-Nash incentive compatible if and only if for every $i \in [n]$,*

$$\left(\frac{b+1}{2}\right)\left(\bar{f}_i(+1) - \bar{f}_i(-1)\right) \geqslant \bar{t}_i(+1) - \bar{t}_i(-1) \geqslant \left(\frac{b-1}{2}\right)\left(\bar{f}_i(+1) - \bar{f}_i(-1)\right). \tag{4.2}$$

A corollary of the previous lemma is that the SCF f is implementable in the Bayes-Nash sense if and only if $\bar{f}_i(+1) - \bar{f}_i(-1) \geqslant 0$ for all $i \in [n]$. We call this property the *marginal monotonicity* of the allocation rule f. The concept of marginal monotonicity simply means on expectation the value of a function increases when the i-th input changes from -1 to $+1$.

Another important implication of the previous lemma is that the incentive compatibility of a mechanism does not depend on the noise level δ. In other words, a mechanism is BN-IC in the noisy environment if and only if it is BN-IC in the noise-free setting.

4.2 Individual Rationality and Expected Revenue

Suppose that by refusing to participate in the mechanism, any individual can ensure a utility of zero. The mechanism design problem then needs to incorporate interim individual rationality (IIR) constraints:

$$\left(\frac{b+x_i}{2}\right) \mathsf{E}\left[f\big(y_i(x_i), y_{-i}(x_{-i})\big)\big|x_i\right] - \mathsf{E}\left[t_i\big(y_i(x_i), y_{-i}(x_{-i})\big)\big|x_i\right] \geqslant 0.$$

Employing a similar approach to the one used for the BN-IC constraints, that is taking the expectation with respect to the others' types and noisy flips, one can verify that the above

equation reduces to

$$\left(\frac{b+1}{2}\right)\left((1-\delta)\bar{f}_i(+1)+\delta\bar{f}_i(-1)\right) \geqslant (1-\delta)\bar{t}_i(+1)+\delta\bar{t}_i(-1)\,, \tag{4.3a}$$

$$\left(\frac{b-1}{2}\right)\left(\delta\bar{f}_i(+1)+(1-\delta)\bar{f}_i(-1)\right) \geqslant \delta\bar{t}_i(+1)+(1-\delta)\bar{t}_i(-1)\,. \tag{4.3b}$$

The first (respectively, second) equation above expresses the IIR condition for the high (respectively, low) type.

By equations (4.3), one notices that the individual rationality constraints are in fact affected by the noise level δ. Since the mechanism can only rely on the noisy reports as the inputs, namely the y_i's, there is always a chance that the message sent by a low type individual flips, and they will end up paying the higher transfer $\bar{t}_i(+1)$ instead of $\bar{t}_i(-1)$ (in the BN sense). Therefore, they need to be compensated for this unexpected flip in order to participate, and this will induce a drag on the space of implementable mechanisms as the noise level increases.

We say a mechanism (f,t) is Bayes-Nash implementable if it is BN-IC and IIR. The next proposition shows that decreasing the noise level weakly *expands* the space of implementable mechanisms.

Proposition 1. *Suppose a mechanism (f,t) is Bayes-Nash implementable at the noise level δ. Then, it will remain Bayes-Nash implementable for all $\delta' < \delta$.*

Proof. We can express the IIR conditions in (4.3) as

$$\left(\frac{b+1}{2}\right)\bar{f}_i(+1)-\bar{t}_i(+1) \geqslant \delta\left[\left(\frac{b+1}{2}\right)\left(\bar{f}_i(+1)-\bar{f}_i(-1)\right)-\left(\bar{t}_i(+1)-\bar{t}_i(-1)\right)\right],$$

$$\left(\frac{b-1}{2}\right)\bar{f}_i(-1)-\bar{t}_i(-1) \geqslant \delta\left[\left(\bar{t}_i(+1)-\bar{t}_i(-1)\right)-\left(\frac{b-1}{2}\right)\left(\bar{f}_i(+1)-\bar{f}_i(-1)\right)\right].$$

The BN-IC constraints in equation (4.2) imply that the *rhs* to both of the above equations are non-negative. Therefore, decreasing δ relaxes the inequalities, and hence the claim follows. □

We say a SCF $f:\{-1,+1\}^n \to \{0,1\}$ is Bayes-Nash implementable if there exist transfer rules $t_i:\{-1,+1\}^n \to \mathbb{R}$ for $i \in [n]$, that make the mechanism (f,t) Bayes-Nash implementable. We now present a *revenue equivalence* type result for implementable SCFs in the current Boolean environment.

Proposition 2 (Revenue equivalence)**.** *A social choice function $f:\{-1,+1\}^n \to \{0,1\}$ is Bayes-Nash implementable if and only if it satisfies marginal monotonicity. In addition, the*

maximum expected revenue that the planner can raise from implementing f is

$$\mathsf{R}_\delta[f] \coloneqq (1-2\delta)\mathsf{E}\left[f(x)\sum_{i=1}^{n} x_i\right] + \left(\frac{b-1}{2}\right)\mathsf{E}[f(x)]\,. \tag{4.4}$$

Therefore, by marginal monotonicity of an implementable SCF, the first expectation term above is always non-negative, and hence the expected revenue becomes a linearly *decreasing* function in noise. We explore the response of the expected social surplus to the noise level as we introduce further tools in the next section. Finally, the above revenue equivalence representation implies the following result.

Corollary 1 (Maximum expected revenue)**.** In the space of all implementable Boolean SCFs, the majority rule asymptotically extracts the maximum expected revenue, where

$$f_{\text{maj}}(x) = \mathbf{1}\left\{\sum_{i=1}^{n} x_i \geqslant 0\right\}. \tag{4.5}$$

To see this, note that $\mathsf{R}_\delta[f]$ is linear in f, thus the following linear threshold function maximizes the expected revenue:

$$\hat{f}_n(x) = \mathbf{1}\left\{\sum_{i=1}^{n} x_i \geqslant \frac{1-b}{2(1-2\delta)}\right\}.$$

Let us denote the above threshold by $\tau \coloneqq \tau(b,\delta)$. The expected revenue associated with this SCF is

$$\mathsf{R}_\delta[\hat{f}_n] = (1-2\delta)\mathsf{E}\left[\sum_{i=1}^{n} x_i \cdot \mathbf{1}\left\{\sum_{i=1}^{n} x_i \geqslant \tau\right\}\right] + \left(\frac{b-1}{2}\right)\mathsf{P}\left(\sum_{i=1}^{n} x_i \geqslant \tau\right) = \frac{1-2\delta}{\sqrt{2\pi}}\sqrt{n}\big(1+o(1)\big)\,,$$

where the last equality follows from the application of the central limit theorem as $n \to \infty$ over the i.i.d. random variables $\{x_i : i \in [n]\}$. A similar approach shows that the expected revenue associated with f_{maj} is equal to $\frac{1-2\delta}{\sqrt{2\pi}}\sqrt{n}\big(1+o(1)\big)$, thus it asymptotically raises the maximum expected revenue. The above analysis is the reason we normalized the target revenue R (in Section 3) by $(1-2\delta)\sqrt{n}$.

5 Noise in the Allocation Rule

In Section 2.4, we introduced the noise sensitivity of a SCF and outlined three reasons for why this quantity is of interest. Here, we elaborate on the second one, which focuses on the distortions in the social surplus caused by the added noise. Specifically, we ask how does the level of noise affect the resulting social surplus of the economy? Denote by $S(x, f)$ the social surplus (namely the individuals' utility plus the revenue raised by the planner) when the true vector of types is x and the implemented outcome is $f \in \{0, 1\}$. Formally, it is equal to

$$S(x, f) = \sum_{i=1}^{n} \left(\frac{b + x_i}{2} \right) f \,. \tag{5.1}$$

We define the surplus distortion (denoted by $\mathsf{SD}_\delta[f]$) as the L^2 distance between what could have been achieved (i.e., $S(x, f(x))$) and what was ultimately realized (i.e., $S(x, f(y))$) as a result of noisy reports:

$$\mathsf{SD}_\delta[f] \coloneqq \mathsf{E}\left[\Big(S(x, f(y)) - S(x, f(x)) \Big)^2 \right] .$$

In the next proposition, we show asymptotically as $n \to \infty$, the surplus distortion closely follows the noise sensitivity, and thus providing additional support for minimizing the noise sensitivity in program (3.1). In particular, according to the following proposition, one can assert that in large economies, selecting a SCF that minimizes the surplus distortion is equivalent to minimizing the noise sensitivity.

Proposition 3. $\lim_{n\to\infty} \left| \frac{1}{n^2} \mathsf{SD}_\delta[f] - \frac{b^2}{4} \mathsf{NS}_\delta[f] \right| = 0$, *uniformly over all* $f : \{-1, +1\}^n \to \{0, 1\}$.

Motivated by the need to study the comparative statics with respect to the noise level δ, and further studying the notion of noise sensitivity, in the next part, we briefly present a self-contained introduction to the Fourier analysis of Boolean functions. A tool that can be applied extensively to many questions in the Boolean environments (e.g., see its application in social choice Kalai, 2002).[6]

5.1 Fourier Analysis of Boolean Functions

Let the n-dimensional Boolean hypercube $\{-1, +1\}^n$ be equipped with the uniform probability measure. The space of $\mathbb{R}$-valued and square integrable functions on this hypercube, denoted

[6] The interested reader is encouraged to refer to the book by O'Donnell (2014) and read further topics in this area.

by $H \coloneqq L^2\left(\{-1,+1\}^n\right)$, is in fact a separable Hilbert space with the inner product operator:

$$\langle f, g\rangle = \mathsf{E}\left[f(x)g(x)\right] = \frac{1}{2^n}\sum_{x\in\{-1,+1\}^n} f(x)g(x), \qquad \forall f, g \in H\,.$$

For every subset $S \subseteq [n]$, define $\chi_S(x) \coloneqq \prod_{i\in S} x_i$. It can be readily checked that the collection of functions $\{\chi_S(\cdot) : S \subseteq [n]\}$ constitutes an *orthonormal* basis for H. In particular, for $S = \varnothing$, one has $\chi_\varnothing(\cdot) \equiv 1$. Every function $f \in H$ thus has a *unique* Fourier expansion in terms of these basis elements, namely

$$f(x) = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x)\,, \tag{5.2}$$

in that $\hat{f}(S)$ is called a Fourier coefficient of f, and is the projection f onto χ_S, that is

$$\hat{f}(S) = \langle f, \chi_S\rangle = \mathsf{E}\left[f(x)\chi_S(x)\right]\,.$$

In particular, $\hat{f}(\varnothing)$ is equal to the mean value of f (i.e., $\mathsf{E}[f]$), and $\hat{f}(\{i\}) = \mathsf{E}\left[f(x)x_i\right] = \big(\bar{f}_i(+1) - \bar{f}_i(-1)\big)/2$ is called a *degree*-1 Fourier coefficient.

Example 1. Let $f(x) = \max\{x_1, x_2\}$, then one can write $f(x)$ as

$$f(x) = \frac{1}{2} + \frac{1}{2}x_1 + \frac{1}{2}x_2 - \frac{1}{2}x_1x_2\,,$$

therefore, $\hat{f}(\varnothing) = \hat{f}(\{1\}) = \hat{f}(\{2\}) = 1/2$, $\hat{f}(\{1,2\}) = -1/2$ and all other Fourier coefficients are zero.

Next, we introduce the concept of *noise stability* that proves very useful in the analysis of noise sensitivity.

Definition 2 (Noise stability)**.** Let $f : \{-1,+1\}^n \to \mathbb{R}$ belong to H. Suppose y is the δ-noisy version of the vector $x \sim \mathrm{Unif}\left(\{-1,+1\}^n\right)$. That is, each y_i is independently distributed from other y_j's and $\mathsf{P}(y_i \neq x_i) = \delta$. Then, the noise stability of the function f is defined as

$$\mathsf{Stab}_\delta[f] \coloneqq \mathsf{E}\left[f(x)f(y)\right]\,. \tag{5.3}$$

From the Fourier expansion in equation (5.2) one has

$$\begin{aligned}\mathsf{E}\left[f(y)|x\right] &= \sum_{S\subseteq[n]} \hat{f}(S) \prod_{i\in S} \mathsf{E}\left[y_i|x_i\right] \\ &= \sum_{S\subseteq[n]} \hat{f}(S) \prod_{i\in S} (1-2\delta)x_i = \sum_{S\subseteq[n]} \hat{f}(S)(1-2\delta)^{|S|}\chi_S(x)\,,\end{aligned} \tag{5.4}$$

where $|S|$ refers to the cardinality of the set S. Therefore, an equivalent representation for the noise stability (in terms of the Fourier coefficients) would be

$$\mathsf{Stab}_\delta[f] = \sum_{S\subseteq[n]} (1-2\delta)^{|S|}\hat{f}(S)^2\,.$$

This representation suggests that higher-degree Fourier weights have a relatively smaller impact on the stability of any SCF due to their effect being subject to a geometric discounting.

Using the concepts introduced above, in the next part we explore the comparative statics of the social surplus and the revenue with respect to the noise level δ.

5.2 Impact of Noise on Revenue and Surplus

We saw in Proposition 2 that increasing the noise level δ decreases the expected revenue in every implementable SCF. Now we see how the ideas from spectral analysis offered in the previous section may be directly applied to study the comparative statics of expected social surplus with respect to the noise.

Equation (5.1) expresses the *realized* social surplus, when the individuals' true type is x, and the implemented outcome is $f \in \{0,1\}$. Therefore, in the noisy setting where $f(y)$ is directed instead of $f(x)$, the expected social surplus is equal to

$$\mathsf{S}_\delta[f] \coloneqq \mathsf{E}\left[S(x,f(y))\right] = \sum_{i=1}^{n} \mathsf{E}\left[\left(\frac{b+x_i}{2}\right) f(y)\right]\,. \tag{5.5}$$

In the next proposition, we offer the comparative statics of $(\mathsf{R}_\delta, \mathsf{S}_\delta)$ with respect to the noise level δ. Before that, we highlight an important connection between implementability and Fourier coefficients.

Remark 2. A SCF f is implementable if and only if all of its degree-1 Fourier coefficients are non-negative. This is the case because (Bayes-Nash) implementability is equivalent to marginal monotonicity, and that in turn means $\bar{f}_i(+1) - \bar{f}_i(-1) \geqslant 0$ for all $i \in [n]$. The former difference is simply equal to $2\hat{f}(\{i\})$, and thus the claim follows.

Proposition 4 (Comparative statics)**.** *For every implementable SCF f, as the noise level $\delta \in (0, 1/2)$ increases, the expected revenue $\mathsf{R}_\delta[f]$ and the expected social surplus $\mathsf{S}_\delta[f]$ decrease linearly in δ.*

Proof. It was previously shown in the revenue equivalence expression (4.4) that R_δ is a linearly decreasing function of δ. Next, using the expression (5.5) and the expansion for the conditional expectation in (5.4), one obtains the following representation for S_δ:

$$\begin{aligned}\mathsf{S}_\delta[f] = \sum_{i=1}^{n} \mathsf{E}\left[\left(\frac{b+x_i}{2}\right) f(y)\right] &= \sum_{i=1}^{n} \mathsf{E}\left[\left(\frac{b+x_i}{2}\right) \sum_{S\subseteq[n]} \hat{f}(S)(1-2\delta)^{|S|}\chi_S(x)\right] \\ &= \frac{b\,n}{2}\,\hat{f}(\varnothing) + \frac{1-2\delta}{2}\sum_{i=1}^{n}\hat{f}(\{i\})\,.\end{aligned}$$

The last equality holds because $\mathsf{E}\left[\chi_S(x)\right] = 0$ for all $S \neq \varnothing$, and $\mathsf{E}\left[x_i\chi_S(x)\right] = 1$ if $S = \{i\}$ and otherwise is equal to zero. Since f is implementable, then all of its degree-1 Fourier coefficients are non-negative, and hence S_δ becomes a linearly decreasing function in δ. □

The intuition behind this result is rather simple. As it relates to the expected revenue, a low type agent must be compensated enough to participate, because there is always a chance that their message flips and they end up paying the high type transfer, even though they derive no utility from the public good. The higher the noise level, the more a low type agent ought to be compensated. On the other hand, a positive transfer from the planner to a low type agent seems alluring to a high type individual. Therefore, to deter them from misreporting their type, the planner has to reduce the transfer *paid* by a high type agent. Both of these two effects create a negative pressure on the expected revenue as the noise level increases. For the expected social surplus, observe the complementarity between the outcome f and the agent's type in the utility function (equation (2.1)). Introducing the noise breaks the optimal assortative allocation with some positive probability and thus lowers the expected social surplus.

Proposition 4 also underscores the cost of protecting privacy. While adding noise to individuals' messages protects the full revelation of their private types, it comes at the cost of decreasing the expected revenue and social surplus associated with each implementable SCF.

In terms of differential privacy, Proposition 4 describes how privacy guarantees translate into efficiency and revenue losses. If we desire an ε-differentially private mechanism, then the relation $\delta = (1 + \mathrm{e}^{\varepsilon})^{-1}$ (followed from (2.2)), and the linear dependence of revenue and surplus on δ, quantify the economic consequences of a given privacy guarantee.

To develop some intuition, in the next section, we provide insights about the revenue and the noise sensitivity of the majority rule. Using them as a stepping stone, we provide the solution to the optimization problem of (3.1) and (3.4) in Section 6.

5.3 Majority Rule

In Corollary 1, we showed that asymptotically as $n \to \infty$, the majority rule extracts the maximum expected revenue. In the following proposition, we provide representations for its revenue and noise sensitivity.

Proposition 5 (Majority rule)**.** *The expected revenue and the noise sensitivity of the majority rule are as follows:*

$$\mathsf{R}_\delta[f_{maj}] = \frac{(1-2\delta)}{\sqrt{2\pi}}\sqrt{n}\big(1+o(1)\big)\,,$$
$$\mathsf{NS}_\delta[f_{maj}] = \frac{\arccos(1-2\delta)}{\pi}\big(1+o(1)\big)\,.$$

The curve in Figure 2 traces the asymptotic values for the *normalized* expected revenue (on the x-axis) and the noise sensitivity (on the y-axis) of the majority rule as $n \to \infty$, while the noise parameter δ varies from 0 to 0.5. As previously mentioned, higher levels of noise are associated with better privacy protection, higher noise sensitivity, and lower expected revenue for every implementable SCF (and here in particular for f_{maj}).

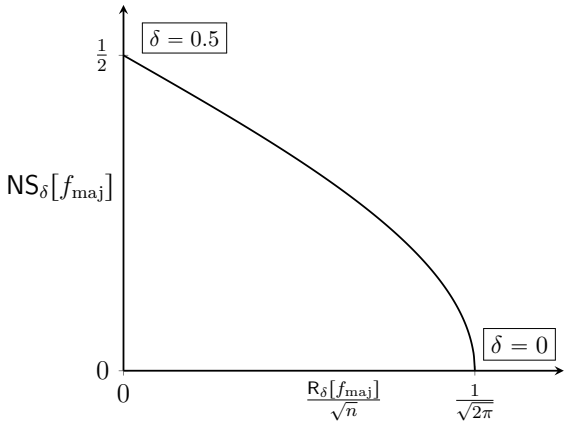

Figure 2: Revenue and Noise Sensitivity of the Majority Rule

A small increase in δ relative to the noise-free environment changes the expected revenue by a little, but significantly raises the noise sensitivity. This is owed to the fact that expected

revenue changes linearly in δ, but the noise sensitivity of the majority rule has "*infinite*" derivative at $\delta = 0$.

Recall that using the language of differential privacy, δ is connected to the privacy guarantee of ε. So our results provide a quantitative relation between the promised level of privacy, the resulting noise sensitivity, and revenue loss for the majority function.

A natural question to ask is: given a fixed level of δ, how much noise robustness can be gained if one is willing to sacrifice some revenue relative to the majority function? This question is the basis of the program in (3.1), which we investigate in the next section.

6 Proof of the Main Results

In this section, we provide the proofs of the main results in Section 3.

6.1 Proof of Theorem 1

In this section, we find the asymptotically optimal solution to the optimization problem of (3.1). Specifically, we ask whether one can find a curve which consistently stays below the one in Figure 2. That is, for a certain level of expected revenue, is there any implementable Boolean function that achieves a smaller noise sensitivity than the majority rule? We answer this question affirmatively and prove that there are two LTFs, whose thresholds are symmetric around 50%, which are asymptotically optimal for optimization (3.1). The one with the smaller provision threshold has the additional advantage of maximizing the expected social surplus $\mathsf{S}_\delta[\cdot]$ given a target revenue level (this is the content of Theorem 2).

Our proof follows three main steps: (i) simplifying the objective and the constraint set; (ii) relaxing the constraint set; and (iii) identifying the asymptotically optimal solutions in the relaxed region and demonstrating that they also belong to the original constraint set.

Step (i)

Observe that since the allocation rules in (3.1) are $\{0,1\}$-valued, then

$$\mathsf{NS}_\delta[f] = \mathsf{P}\big(f(x) \neq f(y)\big) = \mathsf{E}\left[\big(f(x) - f(y)\big)^2\right] = 2\mathsf{E}\left[f(x)^2\right] - 2\mathsf{E}\left[f(x)f(y)\right] .$$

Therefore, we can express the noise sensitivity in terms of the noise stability defined in (5.3), namely $\mathsf{NS}_\delta[f] = 2\left(\mathsf{E}[f] - \mathsf{Stab}_\delta[f]\right)$.[7] Also, note that since the range of all allocation rules

[7] Here, we used the fact that $\mathsf{E}[f^2] = \mathsf{E}[f]$ because f is $\{0,1\}$-valued.

is the binary set $\{0,1\}$, the program in (3.1) always has a solution.

Remark 3. Since the optimization problem (3.1) involves no ex ante heterogeneity across the input coordinates $\{x_i : i \in [n]\}$, there is always a solution that respects the *anonymity* of the type vector x. Formally, the optimal solution only depends on the number of $+1$'s (or equivalently -1's) in the input vector. Therefore, without any loss, we can restrict the constraint set in this optimization problem to all functions that also satisfy the anonymity condition. Henceforth, with some abuse of notation, we refer to $f(x)$ by $f(\nu_n(x))$, $f(\nu)$, or sometimes f.

Following the remarks in Corollary 1 and Notation 2, we normalize the target revenue, $r \coloneqq R/(1-2\delta)\sqrt{n}$. We further assume $r < 1/\sqrt{2\pi}$, as otherwise when $n \to \infty$, there is no SCF (other than the majority rule) that extracts such a high expected revenue.

From revenue equivalence in Proposition 2, we know one can always find a set of transfers, that extract the maximum expected revenue from an implementable SCF f. Hence, thanks to the anonymity condition the expression (4.4) simplifies to:

$$\mathsf{R}_\delta[f] = (1-2\delta)\mathsf{E}\left[f(\nu_n)\,\nu_n\right] + \left(\frac{b-1}{2}\right)\mathsf{E}[f(\nu_n)]\,.$$

Finally, recall that a SCF is implementable if and only if it is marginally monotone. Putting the previous derivations together, we can now express an *equivalent* optimization problem to the one in (3.1):

$$\begin{aligned} &\min_{f} \quad 2\left(\mathsf{E}[f] - \mathsf{Stab}_\delta[f]\right) \\ \text{subject to:} \quad &\frac{1}{\sqrt{n}}\,\mathsf{E}\left[f(\nu_n)\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}}\,\mathsf{E}\left[f(\nu_n)\right] \geqslant r, \\ &\text{and } f \text{ being marginally monotone}\,. \end{aligned} \tag{6.1}$$

We denote the optimal value of the above minimization problem by $\mathcal{V}_n(r)$, that is equal to the minimum noise sensitivity of the implementable SCFs that raise the normalized expected revenue of r.

Remark 4. Even if one is willing to convexify the constraint set in (6.1), by allowing f to take values in the range $[0,1]$, the objective function is still not concave in f, and hence the extreme point theory (commonly used in mechanism design literature) cannot be applied.

Step (ii)

In this part of the proof, we relax the constraint set in (6.1). Toward this, we *index* the above program with the bias (or the mean) of the SCFs. Specifically, we find the minimum bias of the SCFs that satisfy the above revenue constraint:

$$\alpha_n(r) \coloneqq \inf \left\{ \mathsf{E}[f] : \frac{1}{\sqrt{n}} \mathsf{E}\left[f(\nu_n)\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}} \mathsf{E}\left[f(\nu_n)\right] \geqslant r,\ \ f \in H_{[0,1]} \right\}, \tag{6.2}$$

where $H_{[0,1]}$ is the closed subset of L^2 functions from $\{-1,+1\}^n$ to $[0,1]$. Since this is a compact subset, and the revenue constraint induces a closed region, the above infimum is achieved.

Lemma 2 (Bias indexing)**.** *The optimal solution to the minimization problem of* (6.2) *is obtained by* $\bar{\ell}_n(\cdot; r)$ *in* (3.2a)*, and the minimum value* $\alpha_n(r)$ *satisfies*

$$\lim_{n\to\infty} \alpha_n(r) = \Phi\left(-\varphi^{-1}(r)\right). \tag{6.3}$$

This result tells us among all SCFs that raise a target revenue the *linear threshold functions* have the smallest mean. In addition, the associated threshold depends on the normalized revenue r. Higher levels of normalized revenue corresponds to smaller thresholds, thus getting closer to the majority rule.

Remark 5. Taking $r < 1/\sqrt{2\pi}$, equation (6.3) implies that $\lim_{n\to\infty} \alpha_n(r) < 1/2$, so for all n greater than a certain level, one has $\alpha_n(r) < 1/2$. Therefore, we can define the *mirrored* optimization problem to the one in (6.2) as follows:

$$\sup \left\{ \frac{1}{\sqrt{n}} \mathsf{E}\left[f\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}} \mathsf{E}\left[f\right] : \mathsf{E}[f] \geqslant 1-\alpha,\ \ f \in H_{[0,1]} \right\}. \tag{6.4}$$

Since the distribution of ν_n is symmetric around 0, one can see that there exists an $o(1)$ sequence such that, replacing α with $\alpha_n(r) + o(1)$ in the above constraint leads to a supremum of r. That is

$$\mathsf{E}[f] \geqslant 1 - \left(\alpha_n(r) + o(1)\right) \text{ implies } \frac{1}{\sqrt{n}} \mathsf{E}\left[f\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}} \mathsf{E}\left[f\right] \leqslant r\,. \tag{6.5}$$

And specifically, using the same techniques as in Lemma 2, one can show the existence of

a LTF with the following description (as previously expressed in equation (3.2b)),

$$\underline{\ell}_n(x;r) \coloneqq \mathbf{1}\left\{\frac{\nu_n(x)}{\sqrt{n}} \geqslant -\varphi^{-1}(r) + o(1)\right\},$$

that exactly achieves the normalized revenue r, and its mean, i.e., $\mathsf{E}\left[\underline{\ell}_n(x;r)\right]$ equals $1 - \big(\alpha_n(r) + o(1)\big)$.

Our next step is to use the idea of bias indexing to relax the constraint set in (6.1). Observe that the definition of $\alpha_n(\cdot)$ in (6.2) and the condition (6.5) jointly imply the following set inclusion:

$$\begin{aligned}\left\{f \in H_{[0,1]} : \frac{1}{\sqrt{n}}\mathsf{E}\left[f\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}}\mathsf{E}\left[f\right] \geqslant r \text{ and } f \text{ being marginally monotone}\right\} \\ \subseteq \left\{f \in H_{[0,1]} : \alpha_n(r) \leqslant \mathsf{E}\left[f\right] \leqslant 1 - \big(\alpha_n(r) + o(1)\big)\right\}.\end{aligned}$$

Consequently, we can relax the constraint set of the original problem and arrive to the following program:

$$\begin{aligned}&\min_f \; 2\left(\mathsf{E}[f] - \mathsf{Stab}_\delta[f]\right) \\ &\text{subject to: } \; \alpha_n(r) \leqslant \mathsf{E}\left[f\right] \leqslant 1 - \big(\alpha_n(r) + o(1)\big) \text{ and } f \in H_{[0,1]}.\end{aligned} \tag{6.6}$$

We denote the value to this minimization problem by $\mathcal{V}_n^{\mathrm{rel}}(r)$, and importantly we have $\mathcal{V}_n^{\mathrm{rel}}(r) \leqslant \mathcal{V}_n(r)$.

Step (iii)

In this part, we characterize the asymptotically optimal solutions for the relaxed program, and further show they satisfy the original constraint in (6.1), thereby proving their asymptotic optimality for the main program.

Essentially, we will show that the LTFs $\bar{\ell}_n(\cdot;r)$ (in equation (3.2a)) and $\underline{\ell}_n(\cdot;r)$ (in equation (3.2b)) are approximately optimal for the relaxed program. The former achieves the bias lower bound in (6.6), and the latter achieves the bias upper bound. Additionally, since both functions satisfy the constraints of the original optimization problem in (6.1)—namely raising precisely r and being marginally monotone—they will remain asymptotically optimal for the original program. The approximation error due to choosing them as suboptimal solutions for (6.1) converges to zero as $n \to \infty$.

To justify the previous claims, we borrow from a seminal result in the analysis of Boolean

functions, that goes under the name of "majority is the stablest", and its proof mainly relies on the Gaussian isoperimetric inequality (first proved by Borell (1985)). In the following lemma we present a version of this result that suits our need, and we provide a rough sketch of its proof in the appendix.[8] Before that we need to define a notation for the two dimensional CDF of correlated Gaussians.

Definition 3. Let (Z_1, Z_2) be two standard Gaussian random variables, that are ρ-correlated, namely $\mathsf{E}[Z_1 Z_2] = \rho$. We define $\Phi_\rho : \mathbb{R}^2 \to [0,1]$ as $\Phi_\rho(t_1, t_2) \coloneqq \mathsf{P}_\rho(Z_1 \leqslant t_1, Z_2 \leqslant t_2)$. In particular, when $t_1 = t_2 = t$, with some abuse of notation we use $\Phi_\rho(t) \equiv \Phi_\rho(t,t)$.

Lemma 3 ("Majority is the stablest")**.** *Let $f : \{-1,+1\}^n \to [0,1]$ be an anonymous function, and $\delta \in (0, 1/2)$, then*

$$\mathsf{Stab}_\delta[f] \leqslant \Phi_{1-2\delta}\big(\Phi^{-1}(\mathsf{E}[f])\big) + o(1)\,,$$

where the $o(1)$ approximation term is uniform across all anonymous functions.

By using the above inequality, we demonstrate that the two LTFs mentioned earlier are asymptotically optimal for the original program. This establishes the bounds in (3.3), thereby concluding the proof of Theorem 1.

By the previous lemma, the objective function in the relaxed program of (6.6) is lower bounded by

$$\mathsf{E}[f] - \mathsf{Stab}_\delta[f] \geqslant \mathsf{E}[f] - \Phi_{1-2\delta}\big(\Phi^{-1}(\mathsf{E}[f])\big) + o(1)\,,$$

where the $o(1)$ term is uniform across all anonymous SCFs. The expression on the *rhs* above—up to the exclusion of the $o(1)$ term—is symmetric around $\mathsf{E}[f] = 1/2$. In particular, it is increasing (respectively, decreasing) on the region where $\mathsf{E}[f] \leqslant 1/2$ (respectively, $\mathsf{E}[f] \geqslant 1/2$).[9] Therefore, for any anonymous f that belongs to the constraint set of the relaxed problem in (6.6), one has

$$\mathsf{E}[f] - \Phi_{1-2\delta}\big(\Phi^{-1}(\mathsf{E}[f])\big) \geqslant \alpha_n(r) - \Phi_{1-2\delta}\big(\Phi^{-1}(\alpha_n(r)\big) + o(1)\,,$$

where the inequality binds for $f \in \{\underline{\ell}_n(\cdot\,; r), \bar{\ell}_n(\cdot\,; r)\}$, because as constructed in step (ii) above, we have $\mathsf{E}\big[\bar{\ell}_n(\nu_n; r)\big] = \alpha_n(r)$ and $\mathsf{E}\big[\underline{\ell}_n(\nu_n; r)\big] = 1 - \big(\alpha_n(r) + o(1)\big)$. This in turn implies that

$$\mathsf{NS}_\delta[\ell_n] \leqslant \mathcal{V}_n^{\mathrm{rel}}(r) + o(1), \quad \text{for } \ell_n \in \{\underline{\ell}_n(\cdot\,; r), \bar{\ell}_n(\cdot\,; r)\},$$

[8] The original proof is rather long, and has several steps. The curious reader should consult Mossel et al. (2010) or chapter 11.7 of O'Donnell (2014) for the complete proof.

[9] Let us define $\Psi(x) \coloneqq x - \Phi_\rho\big(\Phi^{-1}(x)\big)$ for $x \in [0,1]$. In Appendix A.6, we show that for any $\rho \in [-1,1]$, the mapping Ψ is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$.

and hence the second inequality in equation (3.3) follows because $\mathcal{V}_n^{\mathrm{rel}}(r) \leqslant \mathcal{V}_n(r)$. The first inequality readily holds because $\{\underline{\ell}_n(\cdot;r), \bar{\ell}_n(\cdot;r)\}$ also belong to the constraint set of the original problem in (6.1), as they both raise the normalized expected revenue of r and are monotone functions. This completes the justification of (3.3), and hence the proof of Theorem 1. □

6.2 Proof of Theorem 2

By borrowing the expression found for the expected social surplus in the proof of Proposition 4, and following the approach in the previous section to simplify the revenue constraint, we can recast the optimization problem of (3.4) as:

$$\begin{aligned} &\max \left\{ \frac{b}{2}\,\mathsf{E}[f(\nu_n)] + \frac{1-2\delta}{2n}\,\mathsf{E}\left[f(\nu_n)\nu_n\right] \right\} \\ \text{subject to: } &\frac{1}{\sqrt{n}}\mathsf{E}\left[f(\nu_n)\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}}\,\mathsf{E}\left[f(\nu_n)\right] \geqslant r\,, \\ &\text{and } f \text{ being implementable}\,. \end{aligned} \tag{6.7}$$

This problem falls under the class of linear programs, in that one needs to assign the optimal value to $f(\nu)$ for every $\nu \in \{-n, -n+2, \ldots, n\}$. The corresponding Lagrangian for the relaxed problem, where we skip the implementability condition, is

$$\begin{aligned} \mathcal{L} = &\frac{b}{2}\mathsf{E}[f(\nu_n)] + \frac{1-2\delta}{2n}\,\mathsf{E}\left[f(\nu_n)\nu_n\right] \\ +\lambda &\left(\frac{1}{\sqrt{n}}\,\mathsf{E}\left[f(\nu_n)\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}}\,\mathsf{E}\left[f(\nu_n)\right] - r \right)\,. \end{aligned}$$

Since the Lagrange multiplier λ is non-negative, then $\lambda + \frac{1-2\delta}{2\sqrt{n}} > 0$, and the candidate solution takes the following form

$$f(\nu_n) = \mathbf{1}\left\{ \frac{\nu_n}{\sqrt{n}} \geqslant \frac{-\left(b + \frac{\lambda(b-1)}{(1-2\delta)\sqrt{n}}\right)}{2\lambda + \frac{1-2\delta}{\sqrt{n}}} \right\}\,.$$

One can easily check that increasing λ in the above expression, raises the provision threshold, thus asymptotically (as $n \to \infty$) decreases the expected social surplus, while increasing the expected revenue. This is so because the second term in $\mathsf{S}_\delta[f]$ is of order $O(1/\sqrt{n})$ and asymptotically vanishes compared to the first term, which in turn is decreasing in the provision threshold. Therefore, we should find the minimum λ that satisfies the revenue

constraint. For this, let us denote the threshold by

$$\xi_n \equiv \xi_n(b,\delta,\lambda) \coloneqq \frac{-\left(b+\frac{\lambda(b-1)}{(1-2\delta)\sqrt{n}}\right)}{2\lambda+\frac{1-2\delta}{\sqrt{n}}}\,.$$

Hence, we seek the minimum λ satisfying the following inequality:

$$\mathsf{E}\left[\frac{\nu_n}{\sqrt{n}}\cdot\mathbf{1}\left\{\frac{\nu_n}{\sqrt{n}}\geqslant\xi_n(b,\delta,\lambda)\right\}\right]+\frac{(b-1)}{2(1-2\delta)\sqrt{n}}\,\mathsf{P}\left(\frac{\nu_n}{\sqrt{n}}\geqslant\xi_n(b,\delta,\lambda)\right)\geqslant r\,.$$

As $n\to\infty$, the normalized sum $\nu_n/\sqrt{n}$ converges in distribution to the standard Gaussian, thus the *lhs* in the above inequality converges. Specifically, the first term is asymptotically equal to $\varphi(\xi_n)+o(1)$, and the second term is also of $o(1)$. Therefore, the λ in ξ_n must be chosen so that

$$\varphi\left(\xi_n\right)+o(1)=r\,.$$

Since the provision threshold ξ_n is negative, then the above condition implies that the optimal threshold is $-\varphi^{-1}(r)+o(1)$. Specifically, letting this $o(1)$ sequence be equal to the one in the threshold of $\underline{\ell}_n$ raises precisely the normalized revenue of r, thereby verifying the optimality of the LTF in (3.2b). □

6.3 Additional Remarks

Remark 6. The normalized expected revenue raised by $\{\underline{\ell}_n(\cdot;r),\bar{\ell}_n(\cdot;r)\}$ is equal to r, hence the (unnormalized) expected revenue is $(1-2\delta)\sqrt{n}r$. Furthermore, since $\mathsf{E}\left[\ell_n\right]=\Phi\big(-\varphi^{-1}(r)\big)+o(1)$, for $\ell_n\in\{\underline{\ell}_n(\cdot;\bar{r}),\bar{\ell}_n(\cdot;\bar{r})\}$, then the noise sensitivity takes the following form:

$$\begin{aligned}\mathsf{NS}_\delta[\ell_n]&=2\Big\{\mathsf{E}\left[\ell_n\right]-\Phi_{1-2\delta}\big(\Phi^{-1}(\mathsf{E}\left[\ell_n\right])\big)\Big\}\\&=2\Big\{\Phi\big(-\varphi^{-1}(r)\big)-\Phi_{1-2\delta}\big(-\varphi^{-1}(r)\big)\Big\}+o(1)\,.\end{aligned}$$

We used the above expression to plot the asymptotic Pareto frontiers in Figure 1.

Remark 7. In public good mechanisms, one could envision three main objectives: revenue, surplus, and noise robustness (equivalently, privacy). In our two main theorems, we studied the tradeoffs between each of the last two with the revenue. However, one may question the interaction between social surplus and noise robustness. In fact, in the absence of any revenue constraint, there will be no tradeoff between those two, because the SCF that always

provides the public good, achieves the maximum social surplus and zero noise sensitivity. This is so because the per-capita surplus (as stated in the objective function of (6.7)) consists of two terms. The second component is of $O(1/\sqrt{n})$, and thus negligible compared to the first term. Therefore, by always providing the public good, the first term of the objective function achieves its maximum and the noise sensitivity is zero.

7 Imperfect Knowledge of Preferences

So far, we have studied a setting in which individuals perfectly know their preferences, and the noisy flips take place when they send their messages to the planner. We associated two interpretations with this setting: (i) noise is deliberately added for privacy-preserving concerns; (ii) miscommunication between individuals and the planner is inevitable and reported types could alter as a result.

In this section, we turn to an interpretation of our model where individuals simply do not know their own preferences and observe a noisy signal instead. The idea that agents have imperfect knowledge of their own preferences has received some attention in the mechanism design literature, including the recent work of Gleyze and Pernoud (2022) and Thereze (2022). Now, individuals' reported preferences may differ from their true underlying types, not necessarily due to strategic considerations, but because they lack perfect knowledge about their type. Formally, let x_i be uniformly distributed on $\{-1,+1\}$, representing the true type of agent i, that is hidden from the agent. Instead, they receive a noisy signal $y_i \in \{-1,+1\}$ that is correlated with their type, in the sense that $\mathsf{P}(y_i = x_i) = 1-\delta$ for $\delta \in (0,1/2)$. This means the probability that the agent's signal (information) matches their true type is higher than the probability that it differs. As before, we assume the pairs $\{(x_i,y_i) : i \in [n]\}$ are independently distributed and each has the same distribution explained before.

A mechanism (f,t) is Bayes-Nash incentive compatible in this setting, when each agent reports their signal (i.e., y_i) truthfully, while taking expectations with respect to the others' types. Let $y=(y_1,\ldots,y_n)$ be the vector of signals transmitted by the individuals. Then, in a BN-IC mechanism the planner outputs $f(y)$ and charges agent i by the amount $t_i(y)$ for each $i \in [n]$. The interim incentive constraint for agent i with signal y_i is:

$$\begin{aligned} &\mathsf{E}\left[\left(\frac{b+x_i}{2}\right) f(y_i,y_{-i})\middle|y_i\right] - \mathsf{E}\left[t_i(y_i,y_{-i})\middle|y_i\right] \geqslant \\ &\mathsf{E}\left[\left(\frac{b+x_i}{2}\right) f(-y_i,y_{-i})\middle|y_i\right] - \mathsf{E}\left[t_i(-y_i,y_{-i})\middle|y_i\right]. \end{aligned} \tag{7.1}$$

Lemma 4. *In the present setting, where agents do not have perfect knowledge about their types, a mechanism (f,t) is BN-IC if and only if for every $i \in [n]$,*

$$\left(\frac{b+1}{2}-\delta\right)\left(\bar{f}_i(+1)-\bar{f}_i(-1)\right) \geqslant \bar{t}_i(+1)-\bar{t}_i(-1) \geqslant \left(\frac{b-1}{2}+\delta\right)\left(\bar{f}_i(+1)-\bar{f}_i(-1)\right). \quad (7.2)$$

We skip the proof of this lemma. It follows directly from equation (7.1), observing that because of the independence, the conditional distribution of y_{-i} given y_i is the same as the unconditional distribution of x_{-i}. The first (respectively, second) inequality in (7.2) refers to the interim IC constraint when $y_i = +1$ (respectively, $y_i = -1$). In a sharp contrast with the previous setting, where noise came around in the communication stage, the incentive constraints are now affected by the noise level δ. This is so because in the former case, the noise could flip the individual's message and alter their expected transfer, but in the current setting when the agent sends their signal y_i, the transfer they expect, namely $\bar{t}_i(y_i)$, is not further modified by the noise. Finally, equation (7.2) also confirms that as the noise level δ increases the space of BN-IC mechanisms shrinks.

Next, we express the interim individual rationality constraint for the agent i who received the signal y_i, and has an outside option of zero:

$$\mathsf{E}\left[\left(\frac{b+x_i}{2}\right) f(y_i,y_{-i})\middle|y_i\right] - \mathsf{E}\left[t_i(y_i,y_{-i})\middle|y_i\right] \geqslant 0\,.$$

One can simplify this constraint into two inequalities, that respectively indicate the IR conditions for the high (i.e., $y_i = +1$) and low (i.e., $y_i = -1$) signals:

$$\left(\frac{b+1}{2}-\delta\right)\bar{f}_i(+1) \geqslant \bar{t}_i(+1)\,, \quad (7.3a)$$

$$\left(\frac{b-1}{2}+\delta\right)\bar{f}_i(-1) \geqslant \bar{t}_i(-1)\,. \quad (7.3b)$$

We now state the counterpart of Proposition 2 in the current setting.

Proposition 6. *In the present setting, where agents do not have perfect knowledge about their types, a SCF $f : \{-1,+1\}^n \to \{0,1\}$ is implementable if and only if it satisfies marginal monotonicity, namely $\bar{f}_i(+1) - \bar{f}_i(-1) \geqslant 0$ for all $i \in [n]$. In addition, the maximum expected revenue that the planner can collect from implementing f is*

$$\widetilde{\mathsf{R}}_\delta[f] \coloneqq (1-2\delta)\mathsf{E}\left[f(x)\sum_{i=1}^n x_i\right] + \left(\frac{b-1}{2}+\delta\right)\mathsf{E}[f(x)]\,, \quad (7.4)$$

where the expectation is taken with respect to the uniform measure on $\{-1,+1\}^n$.

We continue by studying the revenue/surplus tradeoff when agents have imperfect knowledge of their preferences. Observe that, in the new setting the implemented outcome is $f(y)$ while the agents' true vector of types is x. Therefore, the expected social surplus follows the same expression of equation (5.5). Hence, the revenue/surplus tradeoff is pinned down by the following program:

$$\begin{gathered}\max \mathsf{S}_\delta[f] \\ \text{subject to: } \widetilde{\mathsf{R}}_\delta[f] \geqslant R, \text{ and } f \text{ being implementable}.\end{gathered}$$

As before, we normalize the lower bound on the expected revenue by $r = R/(1-2\delta)\sqrt{n}$. Then, using the same apparatus as in the proof of Theorem 2, one can show that the same LTF, namely $\underline{\ell}_n(\cdot; r)$, solves the above problem.

Suppose the required revenue R remains fixed, and one looks at the response of the constrained efficient allocation rule in the above problem to the noise. As the agents' information about their preferences deteriorate (corresponding to an increase in δ), the normalized revenue r increases, and correspondingly the provision threshold gets closer to the simple 50% majority rule from *below*. Conversely, an improvement in the agents' knowledge about their types, decreases the threshold and thus increases the chances of provision. This means in the societies where agents have better knowledge about their preferences for public good, the expected likelihood of provision in the efficient allocation rule is higher.

Next, we study the revenue/noise robustness tradeoff. Specifically, we ask the similar question expressed in the optimization problem of (3.1), in that one seeks the SCF with the minimum noise sensitivity subject to raising a target level of expected revenue, in the present setting where agents have imperfect knowledge of their types:

$$\begin{gathered}\min \mathsf{NS}_\delta[f] \\ \text{subject to: } \widetilde{\mathsf{R}}_\delta[f] \geqslant R \text{ and } f \text{ being implementable}.\end{gathered} \tag{7.5}$$

A quick inspection on the expressions for expected revenue in these two settings, namely equations (4.4) and (7.4), implies that

$$\frac{1}{(1-2\delta)\sqrt{n}}\left|\widetilde{\mathsf{R}}_\delta[f] - \mathsf{R}_\delta[f]\right| = o\left(\frac{1}{\sqrt{n}}\right).$$

Therefore, one can follow the same steps taken in Section 6.1, and show that the two LTFs

with approximate thresholds (up to $o(1)$ variations) at $-\varphi^{-1}(r)$ and $\varphi^{-1}(r)$ are asymptotically optimal for the above problem. Hence the following proposition—which is the analogue of Theorem 1 in this setting—follows:

Proposition 7. *In the present setting, where agents do not have perfect knowledge about their types, the following LTFs are asymptotically optimal for the program in* (7.5)*:*

$$g_n(x;r) \coloneqq \mathbf{1}\left\{\frac{\nu_n(x)}{\sqrt{n}} \geqslant -\varphi^{-1}(r) + o(1)\right\}, \text{ and } h_n(x;r) \coloneqq \mathbf{1}\left\{\frac{\nu_n(x)}{\sqrt{n}} \geqslant \varphi^{-1}(r) + o(1)\right\}.$$

Quite naturally, the noise sensitivity of the optimal SCF increases as the agents' knowledge of their preferences deteriorate. But more importantly, similar to the interpretation we attached to Figure 1, the worse are the agents' knowledge about their preferences (equivalently the higher is δ), the *smaller* expected revenue the planner has to give up in order to gain a certain level of noise robustness.

8 Conclusion

We have studied the tradeoffs between privacy preservation, the standard economic objectives of efficiency and revenue, and the stability of the public-good decision rule. Privacy preservation compromises the pursuit of other objectives, but in a large economy we are able to characterize the asymptotically optimal decision rules, and uncover the underlying quantitative tradeoffs.

Our model is standard, but stylized, assuming binary types and a yes/no decision on the provision of a public good. This structure is essential to our results, mainly because we rely on the Fourier analysis of Boolean functions, that is tailored for binary structures. There are some developments about the spectral analysis and noise sensitivity in more general domains (e.g., Chapter 8 of O'Donnell (2014)). However, they are not as crisp and sharp as the binary setting, and we think much more work required in this area.

It is natural to ask the questions of our paper in other environments. Preservation of privacy is an overarching concern, and one can imagine private goods models, as well as public-good settings that are richer than the ones we have focused on here, in which to analyze the effect of privacy-preserving noise. We can only hope that our paper proves a useful starting point for further work.

A Proofs

A.1 Proof of Lemma 1

When the true type of agent i is x_i, the incentive constraint in equation (4.1) reduces to

$$\left(\frac{b+x_i}{2}\right)\mathsf{E}\left[f\big(y_i(x_i),y_{-i}\big)\big|x_i\right]-\mathsf{E}\left[t_i\big(y_i(x_i),y_{-i}\big)\big|x_i\right]\geqslant$$
$$\left(\frac{b+x_i}{2}\right)\mathsf{E}\left[f\big(y_i(-x_i),y_{-i}\big)\big|x_i\right]-\mathsf{E}\left[t_i\big(y_i(-x_i),y_{-i}\big)\big|x_i\right].$$

Since the flips are independent across the individuals, the joint distribution of $\big(y_i(x_i),y_{-i}\big)$ is the same as $\big(y_i(x_i),x_{-i}\big)$. Therefore, one can summarize the previous condition as

$$\left(\frac{b+x_i}{2}\right)\mathsf{E}\left[\bar{f}_i\big(y_i(x_i)\big)\big|x_i\right]-\mathsf{E}\left[\bar{t}_i\big(y_i(x_i)\big)\big|x_i\right]\geqslant$$
$$\left(\frac{b+x_i}{2}\right)\mathsf{E}\left[\bar{f}_i\big(y_i(-x_i)\big)\big|x_i\right]-\mathsf{E}\left[\bar{t}_i\big(y_i(-x_i)\big)\big|x_i\right],$$

in that the expectation operators only refer to the noisy flips. When $x_i=+1$, we expand this expression and cancel the appearing term $1-2\delta$ from both sides, thereby showing the first inequality constraint in equation (4.2). Similarly, when $x_i=-1$, the incentive constraint reduces to the second inequality in (4.2). □

A.2 Proof of Proposition 2

We divide the proof into two parts: (i) showing the equivalence between marginal monotonicity and Bayes-Nash implementability; (ii) proving the revenue equivalence representation in equation (4.4).

Part (i): As a rather immediate corollary of incentive constraints in (4.2), one can observe that the marginal monotonicity of SCF is necessary for every BN-IC mechanism (f,t). It is further sufficient, because if $\bar{f}_i(+1)-\bar{f}_i(-1)\geqslant 0$ for all $i\in[n]$, one can always find a set of transfer functions, $t=(t_1,\ldots,t_n)$, such that their induced marginals $\big(\bar{t}_i(-1),\bar{t}_i(+1)\big)$ satisfy the BN-IC condition in equation (4.2), and the two IIR conditions in (4.3) for each $i\in[n]$. To justify this claim, let $\big(\bar{t}_i(-1),\bar{t}_i(+1)\big)=(\beta_{-1},\beta_{+1})$ be any pair that satisfies the BN-IC condition of equation (4.2) and the IIR conditions of (4.3), induced by the marginally monotone pair $\big(\bar{f}_i(-1),\bar{f}_i(+1)\big)$. We want to show that there exists a function $t:\{-1,+1\}^n\to\mathbb{R}$, whose marginals on the i-th coordinate (averaging out other coordinates)

match (β_{-1}, β_{+1}). To find such a function, we restrict the search to the smaller space of "anonymous" functions, whose value only depend on the number of $+1$'s in the input vector, namely on

$$m(x) \coloneqq \#\{i : x_i = +1\} .$$

Therefore, we denote $t(x)$ by $t(m(x))$. Hence, it is required that

$$\begin{aligned}
\beta_{-1} &= \sum_{m=0}^{n-1} t(m) \binom{n-1}{m} \frac{1}{2^{n-1}} , \\
\beta_{+1} &= \sum_{m=0}^{n-1} t(m+1) \binom{n-1}{m} \frac{1}{2^{n-1}} .
\end{aligned}$$

Let us denote the anonymous function $t(\cdot)$ by the vector $\boldsymbol{t} \equiv \big(t(0), t(1), \ldots, t(n)\big)$, and $\binom{n}{k}$ by $C_{n,k}$. Then, the above linear system is expressed by

$$\begin{bmatrix} C_{n-1,0} & C_{n-1,1} & \cdots & C_{n-1,n-1} & 0 \\ 0 & C_{n-1,0} & C_{n-1,1} & \cdots & C_{n-1,n-1} \end{bmatrix} \boldsymbol{t} = 2^{n-1} \begin{bmatrix} \beta_{-1} \\ \beta_{+1} \end{bmatrix} .$$

Since, the first and last columns of the coefficient matrix are linearly independent, then there always exists a solution to the above system. Therefore, one can always find an anonymous transfer function $t_i(\cdot)$ that implements the marginally monotone pair $\big(\bar{f}_i(-1), \bar{f}_i(+1)\big)$.

Part (ii): For any implementable SCF f, the planner receives the expected transfer

$$\frac{1}{2}\big(\bar{t}_i(+1) + \bar{t}_i(-1)\big) , \tag{A.1}$$

from individual i. Therefore, one should maximize this expression, subject to the BN-IC and IIR conditions, to achieve the maximum expected transfer obtained from the SCF f. To solve this program, we first show the IIR condition for the low type (namely equation (4.3b)) together with the BN-IC condition for the high type (namely the first inequality in (4.2)) imply the IIR condition for the high type (i.e., equation (4.3a)). From the low type IIR condition one obtains

$$\bar{t}_i(-1) \leqslant -\frac{\delta}{1-\delta}\, \bar{t}_i(+1) + \left(\frac{b-1}{2}\right) \left(\frac{\delta}{1-\delta}\, \bar{f}_i(+1) + \bar{f}_i(-1)\right) , \tag{A.2}$$

and the high type BN-IC condition implies

$$\bar{t}_i(+1) \leqslant \bar{t}_i(-1) + \left(\frac{b+1}{2}\right)\left(\bar{f}_i(+1) - \bar{f}_i(-1)\right) .$$

Replacing the former upper bound on $\bar{t}_i(-1)$ in the above inequality and applying some rearrangements imply that

$$\bar{t}_i(+1) \leqslant \left(\frac{b+1}{2} - \delta\right)\bar{f}_i(+1) - (1-\delta)\bar{f}_i(-1) . \tag{A.3}$$

Next, let us investigate the validity of the high type IIR condition (equation (4.3a)). We use equations (A.2) and (A.3) to obtain the following upper bound on the expected transfer paid by the high type, namely the *rhs* of equation (4.3a):

$$\begin{aligned}(1-\delta)\bar{t}_i(+1) + \delta\bar{t}_i(-1) &\leqslant \left(\frac{1-2\delta}{1-\delta}\right)\bar{t}_i(+1) + \delta\left(\frac{b-1}{2}\right)\left(\frac{\delta}{1-\delta}\,\bar{f}_i(+1) + \bar{f}_i(-1)\right) \\ &\leqslant \left(\frac{b+1-\delta(b+3)}{2}\right)\bar{f}_i(+1) + \left(\frac{\delta(b+3)}{2} - 1\right)\bar{f}_i(-1) \\ &= \left(\frac{b+1}{2}\right)\left((1-\delta)\bar{f}_i(+1) + \delta\bar{f}_i(-1)\right) - \left(\delta\bar{f}_i(+1) + (1-\delta)\bar{f}_i(-1)\right) .\end{aligned}$$

This implies that equation (4.3a), which is the high type IIR condition, falls out of the high type BN-IC constraint and the low type IIR constraint.

The above analysis implies that one needs to only maximize the expected transfer on the constrained set induced by the incentive constraints (i.e., equation (4.2)) and the low type IIR condition. Therefore, at the optimum the low type IIR condition as well as one of the incentive constraints must bind. One can show that since $\delta < 1/2$, the extreme point associated with the meet of the low type IIR and high type BN-IC achieves a higher expected revenue than the meet of the low type IIR and low type BN-IC. Hence, the following profile of interim transfers pins down the optimum:

$$\begin{aligned}\bar{t}_i(-1) &= -\delta\bar{f}_i(+1) + \left(\frac{b-1}{2} + \delta\right)\bar{f}_i(-1) , \\ \bar{t}_i(+1) &= \left(\frac{b+1}{2} - \delta\right)\bar{f}_i(+1) - (1-\delta)\bar{f}_i(-1) .\end{aligned}$$

Therefore, the maximum expected transfer from individual i is equal to

$$\frac{\bar{t}_i(+1)+\bar{t}_i(-1)}{2}=\left(\frac{b+1}{4}-\delta\right)\bar{f}_i(+1)+\left(\frac{b-3}{4}+\delta\right)\bar{f}_i(-1)\,.$$

Since the types are distributed uniformly on $\{-1,+1\}^n$, one has

$$\begin{aligned}\bar{f}_i(+1)&=\mathsf{E}\left[f\right]+\mathsf{E}\left[f(x)x_i\right]\,,\\ \bar{f}_i(-1)&=\mathsf{E}\left[f\right]-\mathsf{E}\left[f(x)x_i\right]\,.\end{aligned}$$

Hence, the maximum expected revenue from implementing f follows:

$$\mathsf{R}_\delta[f]=\sum_{i\in[n]}\frac{\bar{t}_i(+1)+\bar{t}_i(-1)}{2}=(1-2\delta)\mathsf{E}\left[f(x)\sum_{i=1}^n x_i\right]+\left(\frac{b-1}{2}\right)\mathsf{E}[f(x)]\,,$$

thereby establishing the representation in (4.4). □

A.3 Proof of Proposition 3

The distortion function follows:

$$\begin{aligned}\mathsf{SD}_\delta[f]&=\mathsf{E}\left[\Big(S(x,f(y))-S(x,f(x))\Big)^2\right]=\mathsf{E}\left[\left(\sum_{i=1}^n\left(\frac{b+x_i}{2}\right)\right)^2\big(f(x)-f(y)\big)^2\right]\,.\\ &=\frac{n^2b^2}{4}\,\mathsf{E}\left[\big(f(x)-f(y)\big)^2\right]+\frac{nb}{2}\,\mathsf{E}\left[\left(\sum_{i=1}^n x_i\right)\big(f(x)-f(y)\big)^2\right]\\ &\qquad+\frac{1}{4}\,\mathsf{E}\left[\left(\sum_{i=1}^n x_i\right)^2\big(f(x)-f(y)\big)^2\right]\,.\end{aligned}$$

Since f is $\{0,1\}$-valued, it holds that

$$\mathsf{E}\left[\big(f(x)-f(y)\big)^2\right]=\mathsf{P}\left(f(x)\neq f(y)\right)=\mathsf{NS}_\delta[f]\,.$$

Recall that $\nu_n=\sum_{i=1}^n x_i$. Therefore,

$$\begin{aligned}\left|\frac{1}{n^2}\,\mathsf{SD}_\delta[f]-\frac{b^2}{4}\,\mathsf{NS}_\delta[f]\right|&\leqslant\frac{b}{2}\,\mathsf{E}\left[\frac{|\nu_n|}{n}\big(f(x)-f(y)\big)^2\right]+\mathsf{E}\left[\frac{\nu_n^2}{4n^2}\big(f(x)-f(y)\big)^2\right]\\ &\leqslant\mathsf{E}\left[\frac{b|\nu_n|}{2n}+\frac{\nu_n^2}{4n^2}\right]\,,\end{aligned}$$

where the second inequality holds because $\big(f(x) - f(y)\big)^2 \leqslant 1$ for all f. Next, observe that $\nu_n/n \to 0$ almost surely, and $|\nu_n/n| \leqslant 1$ because $x_i \in \{-1, +1\}$. Therefore, by Lebesgue dominated convergence theorem, we have:

$$\lim_{n\to\infty} \left| \frac{1}{n^2}\, \mathsf{SD}_\delta[f] - \frac{b^2}{4}\, \mathsf{NS}_\delta[f] \right| \leqslant \lim_{n\to\infty} \mathsf{E}\left[\frac{b|\nu_n|}{2n} + \frac{\nu_n^2}{4n^2} \right] = 0\,,$$

that concludes the proof of uniform convergence. □

A.4 Proof of Proposition 5

The expected revenue extracted by the majority rule (followed by equation (4.4)) is equal to

$$\mathsf{R}_\delta[f_{\text{maj}}] = (1-2\delta)\mathsf{E}\left[\nu_n \cdot \mathbf{1}\left\{\nu_n \geqslant 0\right\}\right] + \left(\frac{b-1}{2}\right) \mathsf{P}\left(\nu_n \geqslant 0\right)\,.$$

Since $\{x_i : i \in [n]\}$ are i.i.d. and uniformly distributed $\{-1, +1\}$-valued random variables, then by the central limit theorem $\frac{1}{\sqrt{n}}\nu_n$ converges in distribution to the standard Gaussian, i.e., $Z \sim \mathcal{N}(0,1)$. Therefore, using the Lebesgue dominated convergence theorem one has

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}}\, \mathsf{E}\left[\nu_n^+\right] = \mathsf{E}\left[Z^+\right] = \frac{1}{\sqrt{2\pi}}, \quad \text{and} \quad \lim_{n\to\infty} \mathsf{P}\left(\nu_n \geqslant 0\right) = \frac{1}{2}\,,$$

thus implying $\mathsf{R}_\delta[f_{\text{maj}}] = \frac{(1-2\delta)}{\sqrt{2\pi}}\sqrt{n}\big(1 + o(1)\big)$.

Next, we examine the noise sensitivity of the majority rule.[10] Let $\operatorname{sgn}(\cdot)$ denote the sign function. For the vector of true types x, and its noisy variant y, one has

$$\mathsf{NS}_\delta[f_{\text{maj}}] = \mathsf{P}\left(\operatorname{sgn}\Big(\sum_{i=1}^n x_i\Big) \neq \operatorname{sgn}\Big(\sum_{i=1}^n y_i\Big) \right),$$

that in turn, due to the symmetry between x and y, is equal to twice the following expression

$$\mathsf{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \geqslant 0 \text{ and } \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i < 0 \right). \tag{A.4}$$

Observe that $\mathsf{E}\left[x_i y_j\right] = 1 - 2\delta$ when $i = j$ and zero otherwise. Then, because of the multi-dimensional version of central limit theorem, the following weak convergence result holds as

[10] Chapter 5 of O'Donnell (2014) includes a comprehensive study of the spectral properties of the majority function.

$n \to \infty$:

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \right) \Rightarrow \left(Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2 \right) ,$$

where $\rho \coloneqq 1 - 2\delta$, and (Z_1, Z_2) are independent standard Gaussians. Hence, the probability in (A.4) converges to

$$\mathsf{P}\left(Z_1 \geqslant 0 \text{ and } \rho Z_1 + \sqrt{1-\rho^2} Z_2 < 0 \right) ,$$

which by the rotational symmetry of (Z_1, Z_2) is equal to $\frac{\arccos \rho}{2\pi}$. Therefore,

$$\mathsf{NS}_\delta[f_{\text{maj}}] = \frac{\arccos(1-2\delta)}{\pi} \big(1 + o(1)\big) .$$

□

A.5 Proof of Lemma 2

The minimization problem in (6.2) clearly falls under the class of linear programs. Therefore, one can express the Lagrangian for this problem as follows:

$$\mathcal{L} = \mathsf{E}\left[f(\nu_n)\right] + \lambda \left(r - \frac{1}{\sqrt{n}} \mathsf{E}\left[f(\nu_n)\nu_n\right] - \frac{b-1}{2(1-2\delta)\sqrt{n}} \mathsf{E}\left[f(\nu_n)\right] \right) .$$

The optimal solution thus takes the following form

$$f(\nu_n) = \mathbf{1}\left\{ \frac{\nu_n}{\sqrt{n}} \geqslant \frac{1}{\lambda} - \frac{b-1}{2(1-2\delta)\sqrt{n}} \right\} . \tag{A.5}$$

Denote the threshold in the above function by $\eta_n \equiv \eta_n(b, \delta, \lambda)$. Since a linear threshold function with the above from is pointwise increasing in λ, and we want to actually minimize $\mathsf{E}\left[f(\nu_n)\right]$, then one needs to choose the minimum λ that satisfies the revenue constraint, namely:

$$\frac{1}{\sqrt{n}} \mathsf{E}\left[f(\nu_n)\nu_n\right] + \frac{b-1}{2(1-2\delta)\sqrt{n}} \mathsf{E}\left[f(\nu_n)\right] \geqslant r .$$

Plugging the optimal form—presented in equation (A.5)—in the above inequality amounts to:

$$\mathsf{E}\left[\frac{\nu_n}{\sqrt{n}} \cdot \mathbf{1}\left\{ \frac{\nu_n}{\sqrt{n}} \geqslant \eta_n(b,\delta,\lambda) \right\} \right] + \frac{b-1}{2(1-2\delta)\sqrt{n}} \mathsf{P}\left(\frac{\nu_n}{\sqrt{n}} \geqslant \eta_n(b,\delta,\lambda) \right) \geqslant r .$$

Applying the central limit theorem followed by monotone convergence theorem imply that as $n \to \infty$, the *lhs* in the above inequality becomes equal to $\varphi(\eta_n) + o(1)$. Therefore, the

optimal threshold in equation (A.5) satisfies:

$$\eta_n = \varphi^{-1}(r) + o(1)\,.$$

This verifies the expression for the optimal solution in equation (3.2a). Next, one can plug the above finding in equation (A.5) and obtain an expression for the optimal value of the minimization problem, namely $\alpha_n(r)$:

$$\alpha_n(r) = \mathsf{P}\left(\frac{\nu_n}{\sqrt{n}} \geqslant \varphi^{-1}(r) + o(1)\right) = \Phi\Big(-\varphi^{-1}(r)\Big) + o(1)\,.$$

The second equality above follows directly from the central limit theorem and thus justifying equation (6.3). □

A.6 The Mapping Ψ

Let $\Lambda_\rho(x, y) \coloneqq \Phi_\rho(\Phi^{-1}(x), \Phi^{-1}(y))$. Then, we have $\Psi(x) = x - \Lambda_\rho(x, x)$. We claim that

$$\frac{\partial \Lambda_\rho(x,y)}{\partial x} = \Phi\left(\frac{\Phi^{-1}(y) - \rho\Phi^{-1}(x)}{\sqrt{1-\rho^2}}\right)\,. \tag{A.6}$$

To see this, let $u = \Phi^{-1}(x)$, and $v = \Phi^{-1}(y)$. Then,

$$\frac{\partial \Lambda_\rho(x,y)}{\partial x} = \frac{\partial \Phi_\rho(u,v)}{\partial u}\,\frac{\partial u}{\partial x} = \frac{\partial \Phi_\rho(u,v)}{\partial u}\,\frac{1}{\varphi(u)}\,. \tag{A.7}$$

Let $\varphi_\rho(u, v)$ denote the density associated with $\Phi_\rho(u, v)$. Given u, the distribution of v is a Gaussian with mean ρu and variance $1 - \rho^2$, so

$$\varphi_\rho(u,v) = \varphi(u)\varphi\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right)\,.$$

Therefore,

$$\Phi_\rho(u,v) = \int_{-\infty}^{u} \varphi(s) \int_{-\infty}^{v} \varphi\left(\frac{t-\rho s}{\sqrt{1-\rho^2}}\right) \mathrm{d}t\,\mathrm{d}s\,,$$

that in turn means

$$\frac{\partial \Phi_\rho(u,v)}{\partial u} = \varphi(u)\int_{-\infty}^{v} \varphi\left(\frac{t-\rho u}{\sqrt{1-\rho^2}}\right)\mathrm{d}t = \varphi(u)\Phi\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right)\,.$$

Substituting the above derivation in (A.7) implies (A.6). Therefore, by the symmetry between x and y, it follows that

$$\Psi'(x) = 1 - 2\Phi\left(\frac{(1-\rho)\Phi^{-1}(x)}{\sqrt{1-\rho^2}}\right) = 1 - 2\Phi\left(\sqrt{\frac{1-\rho}{1+\rho}}\,\Phi^{-1}(x)\right).$$

The above function is decreasing in x, and is equal to 0 at $x = 1/2$, therefore, $\Psi(x)$ is increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$.

A.7 Proof Sketch for Proposition 6

Following the similar steps of the the proof of Proposition 2, we can show that marginal monotonicity is a necessary and sufficient condition for the Bayes-Nash implementability of the SCF f. Next, observe that the expected transfer from agent i to the planner is $\big(\bar{t}_i(-1) + \bar{t}_i(+1)\big)/2$. It is then straightforward to show that in the optimum the BN-IC constraint for the high type (namely the first inequality in (7.2)) and the IR condition for the low type (i.e., equation (7.3b)) bind. Hence, the optimum transfers are:

$$\begin{aligned}
\bar{t}_i(-1) &= \left(\frac{b-1}{2} + \delta\right)\bar{f}_i(-1)\,,\\
\bar{t}_i(+1) &= \left(\frac{b+1}{2} - \delta\right)\bar{f}_i(+1) - (1-2\delta)\bar{f}_i(-1)\,.
\end{aligned}$$

Given that $\bar{f}_i(z) = \mathsf{E}[f] + z\,\mathsf{E}\left[f(x)x_i\right]$ for $z \in \{-1, +1\}$, summing these expressions over i and then dividing by two results in the representation (7.4).

B Intuitive Proof of Lemma 3

We present a very high level sketch of the proof, explaining the pillars and the main ideas. The are a handful of different methods for proving this theorem (as recent as Eldan et al. (2022)), but we rely on the approach offered in Mossel et al. (2010).

The proof relies on two main ideas: (i) Borell's Gaussian isoperimetric inequality; (ii) *Invariance principle*. We first present some preliminaries that discipline the reading of how these two ideas come together and shape the proof.

B.1 Preliminaries

We start with the definition of the noise operator acting on the Hilbert space $H = L^2(\{-1,+1\}^n)$ with the uniform measure on the hypercube.

Definition 4 (Noise operator)**.** Let $\rho \in (0,1)$ and define $\mathsf{T}_\rho : H \to H$ as

$$\mathsf{T}_\rho f(x) = \mathsf{E}\left[f(y)|x\right] ,$$

where $x = (x_1, \ldots, x_n)$ is a point uniformly drawn from the hypercube, and y is its ρ-correlated version, such that $\mathsf{E}\left[y_i x_i\right] = \rho$ for each coordinate $i \in [n]$.

For every basis element $\chi_S \in H$, one has $\mathsf{T}_\rho \chi_S(x) = \rho^{|S|}\chi_S(x)$. Since, the noise operator is linear, applying that on the Fourier expansion in equation (5.2) implies

$$\mathsf{T}_\rho f(x) = \sum_{S\subseteq[n]} \rho^{|S|}\hat{f}(S)\chi_S(x) .$$

In addition, the noise operator is commutative and has the semi-group property, that is for $\rho_1, \rho_2 \in (0,1)$, one has $\mathsf{T}_{\rho_1}\mathsf{T}_{\rho_2} = \mathsf{T}_{\rho_2}\mathsf{T}_{\rho_1} = \mathsf{T}_{\rho_1\rho_2}$. Furthermore, the above Fourier representation of the noise operator implies that for every $f, g \in H$, it holds that $\langle f, \mathsf{T}_\rho g\rangle = \langle \mathsf{T}_\rho f, g\rangle$.

Looking back at the definition of the noise stability in equation (5.3), one observes that

$$\mathsf{Stab}_\delta[f] = \langle f, \mathsf{T}_{1-2\delta} f\rangle = \langle \mathsf{T}_{\sqrt{1-2\delta}} f, \mathsf{T}_{\sqrt{1-2\delta}} f\rangle = \mathsf{E}\left[\left(\mathsf{T}_{\sqrt{1-2\delta}} f(x)\right)^2\right] . \tag{B.1}$$

Next, we present the passing from the Boolean to Gaussian environment. Let γ be the standard Gaussian measure on $\mathbb{R}^n$, and $L^2(\mathbb{R}^n;\gamma)$ be the Hilbert space of square integrable functions with respect to γ, equipped with its natural inner product.

Definition 5 (Gaussian evaluation)**.** Let $z \in \mathbb{R}^n$ be distributed according to the standard Gaussian measure γ. For a Boolean function $f \in H$, we abuse the notation and define its Gaussian evaluation as

$$f(z) = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(z) .$$

Since $f \in H$, then

$$\mathsf{E}_\gamma\left[f(z)^2\right] = \sum_{S\subseteq[n]} \hat{f}(S)^2 = \mathsf{E}\left[f(x)^2\right] < \infty ,$$

and hence the Gaussian passing of f belongs to $L^2(\mathbb{R}^n;\gamma)$.

Remark 8. Inspired by the previous definition, one can extend the domain of other operators, such Stab_δ and T_ρ, to $L^2(\mathbb{R}^n;\gamma)$. For example, let z and z' be two n-dimensional standard Gaussian vectors, where their corresponding coordinates are ρ-correlated, then:

$$\mathsf{T}_\rho f(z) = \mathsf{E}\left[f(z')|z\right] = \sum_{S\subseteq[n]} \rho^{|S|}\hat{f}(S)\chi_S(z)\,,$$
$$\mathsf{Stab}_\delta[f] = \left\langle f, \mathsf{T}_{1-2\delta}f\right\rangle = \left\langle \mathsf{T}_{\sqrt{1-2\delta}}f, \mathsf{T}_{\sqrt{1-2\delta}}f\right\rangle = \mathsf{E}\left[\left(\mathsf{T}_{\sqrt{1-2\delta}}f(z)\right)^2\right].$$

B.2 Borell's Isoperimetric Inequality

At this point it is recommended for the reader to refresh their memory with the definitions of Gaussian functions in the remarks 1 and 3.

Theorem 3 (Borell (1985))**.** *Fix $\delta \in (0,1/2)$. Then, for any $f \in L^2(\mathbb{R}^n;\gamma)$ with the range $[0,1]$, and $\mathsf{E}[f] = \mu$, it holds that*

$$\mathsf{Stab}_\delta[f] \leqslant \Phi_{1-2\delta}\big(\Phi^{-1}(\mu)\big)\,. \tag{B.2}$$

The *rhs* to the above inequality is *equal* to the noise stability of the indicator function of any half-space $H \subseteq \mathbb{R}^n$ with the Gaussian volume of $\mathrm{Vol}_\gamma(H) = \mu$.

An interesting corollary to this theorem is that among all measurable subsets of $\mathbb{R}^n$, with a fixed Gaussian volume, the half-spaces have the minimum sensitivity to noise. Formally, let us denote the n-dimensional standard Gaussian probability measure by P_γ. Consider any measurable subset A with $\mathrm{Vol}_\gamma(A) = \mu > 0$, and any half-space H with the same volume μ. Then, inequality (B.2) implies that

$$\mathsf{P}_\gamma\left(x \in A, y \in A\right) \leqslant \mathsf{P}_\gamma\left(x \in H, y \in H\right)\,,$$

where $x \sim \gamma$ and y is its δ-noisy version, that is $\mathsf{E}[y_i|x_i] = (1-2\delta)x_i$ for each $i \in [n]$. Canceling $\mathsf{P}_\gamma(x \in A)$ from both sides amounts to

$$\mathsf{P}_\gamma\left(y \in A \middle| x \in A\right) \leqslant \mathsf{P}_\gamma\left(y \in H \middle| x \in H\right)\,.$$

This means if one starts at a random point x inside the subset A, then the chances of leaving this region due to adding noise is minimal for half-spaces.

B.3 Invariance Principle

In this part, we offer an intuitive statement of the invariance principle. For that, we need to define the concept of *influence*.

Let $x^{i\mapsto+1}$ be the vector x, where its i-th coordinate is replaced with $+1$. Similarly, define $x^{i\mapsto-1}$. Then, holding all other coordinates constant, one can define the *derivative* operator $\mathsf{D}_i : H \to \mathbb{R}$ as

$$\mathsf{D}_i[f](x) = \frac{f(x^{i\mapsto+1}) - f(x^{i\mapsto-1})}{2}\,.$$

Definition 6 (Coordinate influence)**.** For $f : \{-1,+1\}^n \to \mathbb{R}$ and $i \in [n]$ define

$$\mathsf{Inf}_i[f] = \sum_{S\ni i} \hat{f}(S)^2\,.$$

That is the influence of coordinate i on f is the sum of f's squared Fourier weights containing i. One can immediately see that $\mathsf{Inf}_i[f] = \mathsf{E}\left[\mathsf{D}_i[f](x)^2\right]$. Hence, the influence of input i should be interpreted as the expected change that it makes on the function f.

Next, we explain what it means for a function $F : \{-1,+1\}^n \to \mathbb{R}$ to be *invariant*. For any x not belonging to the hypercube, we identify $F(x)$ by the evaluation of its Fourier representation at x. Hence, with some abuse of notation one can extend the domain of F to the entire $\mathbb{R}^n$.

Let $x = (x_1, \ldots, x_n)$ and $z = (z_1, \ldots, z_n)$ be two vectors with i.i.d. elements, such that their first few moments match, namely $\mathsf{E}[x_i] = \mathsf{E}[x_i^3] = \mathsf{E}[z_i] = \mathsf{E}[z_i^3] = 0$, $\mathsf{E}[x_i^2] = \mathsf{E}[z_i^2] = 1$ for all $i \in [n]$, and the fourth moment is finite. For example, x can be drawn uniformly from the hypercube $\{-1,+1\}^n$ and z from the n-dimensional standard Gaussian distribution on $\mathbb{R}^n$. Suppose the previously mentioned function F has *small* influence with respect to all of its input coordinates, that is there is no single coordinate that can determine the outcome with high probability.[11] Then, the invariance principle claims that for any sufficiently smooth function $\Upsilon : \mathbb{R} \to \mathbb{R}$, as $n \to \infty$ one has

$$\left|\mathsf{E}\left[\Upsilon\big(F(x)\big)\right] - \mathsf{E}\left[\Upsilon\big(F(z)\big)\right]\right| = o(1)\,. \tag{B.3}$$

The approximation error $o(1)$ becomes *uniform* over all F's, that put vanishingly small influence on every single coordinate.

[11] Observe that we intentionally state these results qualitatively, as their quantitative versions require many approximation steps, which are carried out in Mossel et al. (2010).

B.4 Proof Sketch

The reader should now have good senses on how to put the previous two ideas together and reach to the conclusion. First, observe that in our setup, where f is supposed to be an anonymous SCF, the small influence condition automatically holds, because f treats all its input coordinates symmetrically, thus one can safely apply the invariance principle. Second, equation (B.1) hints at choosing Υ to be the quadratic function, i.e., $t \mapsto t^2$—that is "sufficiently smooth"—and to assign $F(x) = \mathsf{T}_{\sqrt{1-2\delta}} f(x)$. Then, the invariance principle in equation (B.3) implies that

$$\left| \mathsf{E}\left[\left(\mathsf{T}_{\sqrt{1-2\delta}} f(x) \right)^2 \right] - \mathsf{E}\left[\left(\mathsf{T}_{\sqrt{1-2\delta}} f(z) \right)^2 \right] \right| = o(1)\,.$$

Third, the Gaussian noise stability is upper bounded by the Borell's isoperimetric inequality in equation (B.2). Hence, the previous equation implies that for every anonymous SCF f:

$$\mathsf{Stab}_\delta[f] = \mathsf{E}\left[\left(\mathsf{T}_{\sqrt{1-2\delta}} f(x) \right)^2 \right] \leqslant \Phi_{1-2\delta}\big(\Phi^{-1}(\mathsf{E}[f])\big) + o(1)\,,$$

thereby verifying the claim of Lemma 3.

References

[1] Itai Benjamini, Gil Kalai, and Oded Schramm (1999). "Noise Sensitivity of Boolean Functions and Applications to Percolation," *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 90: 5–43.

[2] Graeme Blair, Kosuke Imai, and Yang-Yang Zhou (2015). "Design and Analysis of the Randomized Response Technique," *Journal of the American Statistical Association*, 110(511): 1304–1319.

[3] Christer Borell (1985). "Geometric Bounds on the Ornstein-Uhlenbeck Velocity Process," *Probability Theory and Related Fields*, 70(1): 1–13.

[4] Eric Budish and Judd B. Kessler (2022). "Can Market Participants Report their Preferences Accurately (Enough)?" *Management Science*, 68(2): 1107–1130.

[5] Yiling Chen, Stephen Chong, Ian A. Kash, Tal Moran, and Salil Vadhan (2016). "Truthful Mechanisms for Agents that Value Privacy," *ACM Transactions on Economics and Computation (TEAC)*, 4(3): 1–30.

[6] Cynthia Dwork (2008). "Differential Privacy: A Survey of Results," "International Conference on Theory and Applications of Models of Computation," Springer, 1–19.

[7] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor (2006). "Our data, Ourselves: Privacy via Distributed Noise Generation," "Annual International Conference on the Theory and Applications of Cryptographic Techniques," Springer, 486–503.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith (2006). "Calibrating Noise to Sensitivity in Private Data Analysis," Shai Halevi and Tal Rabin (Editors), "Theory of Cryptography," Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-32732-5, 265–284.

[9] Cynthia Dwork and Aaron Roth (2014). "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407, ISSN 1551-305X, doi:10.1561/0400000042, URL http://dx.doi.org/10.1561/0400000042.

[10] Ran Eilat, Kfir Eliaz, and Xiaosheng Mu (2021). "Bayesian Privacy," *Theoretical Economics*, 16(4): 1557–1603.

[11] Ronen Eldan, Dan Mikulincer, and Prasad Raghavendra (2022). "Noise Stability on the Boolean Hypercube via a Renormalized Brownian Motion," *arXiv preprint arXiv:2208.06508*.

[12] Georgina Evans and Gary King (2023). "Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook Urls Dataset," *Political Analysis*, 31(1): 1–21.

[13] Georgina Evans, Gary King, Margaret Schwenzfeier, and Abhradeep Thakurta (2019). "Statistically valid Inferences from Privacy Protected Data," *American Political Science Review*.

[14] Georgina Evans, Gary King, Adam D. Smith, and Abhradeep Thakurta (2022). "Differentially Private Survey Research," *American Journal of Political Science*, 27: 703–709.

[15] Quan Geng and Pramod Viswanath (2015). "The Optimal Noise-Adding Mechanism in Differential Privacy," *IEEE Transactions on Information Theory*, 62(2): 925–951.

[16] Simon Gleyze and Agathe Pernoud (2022). "How Competition Shapes Information in Auctions," Mimeo, Stanford University.

[17] Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich (2013). "Identifying Personal Genomes by Surname Inference," *Science*, 339(6117): 321–324.

[18] Avinatan Hassidim, Déborah Marciano, Assaf Romm, and Ran I. Shorrer (2017). "The Mechanism is Truthful, Why aren't You?" *American Economic Review*, 107(5): 220–24.

[19] Avinatan Hassidim, Assaf Romm, and Ran I. Shorrer (2021). "The Limits of Incentives in Economic Matching Procedures," *Management Science*, 67(2): 951–963.

[20] Jianping He, Lin Cai, and Xinping Guan (2018). "Preserving Data-Privacy with Added Noises: Optimal Estimation and Privacy Analysis," *IEEE Transactions on Information Theory*, 64(8): 5677–5690.

[21] Zhiyi Huang and Sampath Kannan (2012). "The Exponential Mechanism for Social Welfare: Private, Truthful, and Nearly Optimal," "2012 IEEE 53rd Annual Symposium on Foundations of Computer Science," IEEE, 140–149.

[22] Gil Kalai (2002). "A Fourier-Theoretic Perspective on the Condorcet Paradox and Arrow's Theorem," *Advances in Applied Mathematics*, 29(3): 412–426.

[23] Daniel Krähmer and Roland Strausz (2023). "Optimal Nonlinear Pricing with Data-Sensitive Consumers," *American Economic Journal: Microeconomics*, 15(2): 80–108.

[24] Daniel McFadden (2009). "The Human Side of Mechanism Design: a Tribute to Leo Hurwicz and Jean-Jacque Laffont," *Review of Economic Design*, 13(1): 77–100.

[25] Frank McSherry and Kunal Talwar (2007). "Mechanism Design via Differential Privacy," "48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)," IEEE, 94–103.

[26] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz (2010). "Noise Stability of Functions with Low Influences: Invariance and Optimality," *Annals of Mathematics*, 171: 295–341.

[27] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz (2012). "Approximately Optimal Mechanism Design via Differential Privacy," "Proceedings of the 3rd Innovations in Theoretical Computer Science conference," ITCS '12, Association for Computing Machinery, New York, NY, USA, 203–213.

[28] Kobbi Nissim and David Xiao (2015). *Mechanism Design and Differential Privacy*, Springer Berlin Heidelberg, New York, NY, URL http://link.springer.com/referenceworkentry/10.1007/978-3-642-27848-8_548-1.

[29] Ryan O'Donnell (2014). *Analysis of Boolean Functions*, Cambridge University Press.

[30] Alex Rees-Jones (2018). "Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match," *Games and Economic Behavior*, 108: 317–330.

[31] Alex Rees-Jones and Samuel Skowronek (2018). "An Experimental Investigation of Preference Misrepresentation in the Residency Match," *Proceedings of the National Academy of Sciences*, 115(45): 11471–11476.

[32] João Thereze (2022). "Adverse Selection and Endogenous Information," Mimeo, Princeton University.

[33] Stanley L. Warner (1965). "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60(309): 63–69.

[34] David Xiao (2013). "Is Privacy Compatible with Truthfulness?" "Proceedings of the 4th Conference on Innovations in Theoretical Computer Science," ITCS '13, Association for Computing Machinery, 67–86.