Wisdom of the Crowd Voting: Truthful Aggregation of Voter Information and Preferences

Grant Schoenebeck* Biaoshuai Tao†

Abstract

We consider two-alternative elections where voters' preferences depend on a state variable that is not directly observable. Each voter receives a private signal that is correlated to the state variable. As a special case, our model captures the common scenario where voters can be categorized into three types: those who always prefer one alternative, those who always prefer the other, and those contingent voters whose preferences depends on the state. In this setting, even if every voter is a contingent voter, agents voting according to their private information need not result in the adoption of the universally preferred alternative, because the signals can be systematically biased.

We present a mechanism that elicits and aggregates the private signals from the voters, and outputs the alternative that is favored by the majority. In particular, voters truthfully reporting their signals forms a strong Bayes Nash equilibrium (where no coalition of voters can deviate and receive a better outcome).

1 Introduction

Social choice theory studies how to aggregate participants' heterogeneous opinions/preferences and output a collective decision from a set of alternatives. Typically, though not always, it is assumed that each participant has a clear preference over the alternatives, e.g., a preference order over all the alternatives, a valuation for each alternative, etc. However, even with only two alternatives, this is not the typical the case for *all* participants. In addition to the participants who have clear, *predetermined* preferences for one alternative over the other, typically there are also *contingent* participants who only have partial information on which alternative is "preferable for them" and yet would like to select the alternative that is "preferable for them."

A standard example would be an election with two candidates a and b coming from political parties A and B respectively. Voters are normally partitioned into three types: some partisans for Party A prefer candidate a based on his support for the platform of party A; other partisans for Party B prefer candidate b based on her support for the platform of party B; finally, there are swing voters who are largely indifferent between the party's platforms and would like to elect whichever candidate can make more progress on non-partisan issues. However, swing voters do not have perfect information about which candidate is better suited for addressing the non-partisan needs of the community. Instead, each voter has a hunch of which candidate will perform better on the non-partisan problems facing the community based on both public information and their private experiences and beliefs.

Additional examples where participant have preferences, but may or may not know what is "preferable for them" abound. In votes for corporate strategies, for hiring decisions, and for policy decisions typically some participants would like to select according to some truth they are collectively trying to discern (e.g., impact on future profits, suitability for the position, efficacy of policy, etc.) while others may have predetermined preferences, for example, because of the way they are uniquely affected (e.g. prominence of position in future corporate strategy, vision/skills of the job candidate, who in particular the policy benefits/harms).

*University of Michigan schoeneb@umich.edu

†John Hopcroft Center for Computer Science, Shanghai Jiao Tong University bstao@sjtu.edu.cn

1.1 Informal Setting

In this paper, we consider a two-alternative social choice setting where voters' preferences may depend on a state variable. In the above example, the state is which candidate will make more progress on non-partisan issues. In general, the state need not be binary. For example, we could generalize the above example so which candidate would preform better on non-partisan issues is on a scale of 1 to 10 (where 1 indicates candidate a is much better and 10 indicates candidate b is much better). Predetermined voters' preferences would still not depend on the state. However, contingent voters' may have different thresholds on the state where they would transfer their support from candidate a to candidate b.

The state variable is not directly observable in the election phase, as the performance of a new government official, a new hired employee, the effect of a new policy, etc., may not be revealed until many years after the vote. Instead, each voter receives a signal that is correlated to the hidden state which models the information voters received from difference sources.

Our goal is to select the alternative that would be preferred by the majority if they knew the state of the world. If some type of predetermined voters forms a majority, this is rather easy. However, in the case where the predetermined voters of neither alternative forms a majority, the mechanism needs to aggregate the information and preferences of the conditional voters and select the alternative which, for a majority of voters, is "preferable for them".

1.2 Imperfectly Informed Voters

1.2.1 Social Choice

The study of social choice theory with imperfectly informed voters dates back to Condorcet's jury theorem in 1785 [6], and has also been widely studied [20, 21, 30]. In Condorcet's setting, there are two alternatives one of which is "correct", and each voter votes for the correct alternative with probability p. Condorcet's jury theorem states that the probability that the majority voting scheme outputs the correct alternative goes to 1 as the number of voters increases when $p > 0.5$, and, conversely, this probability goes to 0 when $p < 0.5$.

Two unfortunate limitations for Condorcet's jury theorem are:

1. It fails to output the correct alternative in the case voters' beliefs are aligned to the incorrect alternative (i.e., $p < 0.5$).

2. It assumes voters vote truthfully and disregards voters' potential strategic behaviors. However, even in the case all voters have the same preference for the correct alternative, voting truthfully still may not be a Nash equilibrium [1].

To circumvent the first limitation, Feddersen and Pesendorfer [9] consider the scenario where voters play a Nash equilibrium strategy profile, while the voting rule is still the majority scheme. Feddersen and Pesendorfer [9] show that when the number of voters tends to infinity, the probability that a majority voting scheme outputs the correct alternative approaches 1 if voters play the equilibrium strategy profile, while this probability is bounded away from 1 if voters play the truthful strategy profile instead. Feddersen and Pesendorfer's model assigns each voter a preference parameter $x \in [-1, 1]$ describing his/her alignment to the two alternatives. In their model, the unique (Bayes) Nash equilibrium is characterized by two thresholds $x_0, x_1 \in [-1, 1]$ with $x_0 < x_1$ such that voters with preferences below x_0 always vote for one alternative, voters with preferences above x_1 always vote for the other, and voters with preferences between x_0 and x_1 vote truthfully. Although Feddersen and Pesendorfer's solution guarantees that the correct alternative is output with high probability, it requires sophisticated voters.

The voters need to calculate the values of x_0 and x_1 to decide their actions. The values x_0 and x_1 are each the zero point of a continuous monotone function involving a complicated Riemann integral. This is usually too demanding for voters in practice, especially those who do not have a mathematical background. Moreover, it needs to be a common knowledge that all agents can and will perform this computation.

In this paper, we take a different approach. Instead of asking voters to play the Nash equilibrium for majority voting, we seek to design a more sophisticated voting scheme, or a mechanism, than the majority voting scheme, such that voters are incentivized to vote *truthfully* under the mechanism, while guaranteeing

the correct alternative is output with high probability. Our social choice mechanism thus elicits truthful information from the voters and then aggregate it.

1.2.2 Information Aggregation

The information aggregation literature considers how to obtain a "correct answer" by aggregating individuals' partial information—the crowd's wisdom. The straightforward procedure of outputting the answer that is believed to be correct by the majority does not always work [3, 29]. An example where this fails is when the crowd has a strong prior belief over the incorrect answer while novel specialized knowledge is only shared among a minority of the agents. It is also known that further calibration based on collecting participants' confidences (the posterior of their beliefs) does not always solve this problem [11, 25]. In a seminal work by Prelec et al. [25], a new "surprising popular" approach was proposed: the participants' predictions over the other remaining participants' reported answers are collected, and the answer that is reported by more participants than predicted is output (we will review this in Sect. 3.1). They justified this approach both theoretically and through experimentation. In particular, they demonstrate the viability of approaches that require agents to predict reports of the other agents. Hosseini et al. [12] extend the surprising popular approach to the non-binary setting, where the goal is to learn the correct *ranking over many options* instead of the correct answer in two options.

The work by Prelec et al. [25] and Hosseini et al. [12] Firstly, the objective for an information aggregation mechanism is to output the correct answer. Participants who collaboratively contribute their knowledge/information do not have preferences on which answer is finally selected. This is fundamentally different from the social choice setting where the whole point of a social choice mechanism is to select an alternative favored by the majority. Secondly, as agents care about the outcome, agents may be strategic and manipulate their reports in order to make their preferred alternatives win, while Prelec et al. [25] do not put the problem in a game theory setting.

1.3 Our Results

In this paper, we study the social choice problem in a game theory setting with the existence of imperfectly informed voters who only have partial information regarding which alternative is more favorable. For various settings with two alternatives, we propose a mechanism that aggregates participants' private information and outputs the alternative favored by more than half of the participants. Our mechanisms are truthful, in the sense that the truthful strategy profile forms a *strong Bayes Nash Equilibrium*.

Our mechanisms borrows some ideas from Prelec et al.'s surprising popular mechanism and, in particular, require participants to predict the responses of other agents. On the other hand, we use a novel "median trick" that ensures our mechanisms have a group truthfulness property. Intuitively, by the *median voter theorem* [2, 13], the median voter's vote is favorable by the majority; by dedicated design, our mechanism ensures that the voters who are "below" the median have a conflict of interest to the voters who are "above" the median, which makes sure less than half of the voters have incentive to deviate and those voters can only change the outcome in the unfavorable direction by the property of median.

From a high level, our work can be understood as a revelation principal applied to plurality voting.[1] However this view is not entirely accurate. First, our equilibrium concept is strong Bayes Nash equilibrium while we only know that plurality voting implements the desired majority preferred outcome in (Bayesian) Nash equilibrium [9]. Second, the revelation principal requires that agents report all their knowledge. In our case, this would include the entire prior, which is not realistic. In contrast, our mechanisms only require that agents report a preference and a prediction of other agents preferences. Such reporting requirements have previously been shown to be pragmatic [14, 25]. Third, our setting is different than prior work [9], and this makes our results incomparable. In particular, we deal with a discrete state space. This difference also allows us to achieve some of our results not just in the limit, but for finite sets of agents. Appendix A contains an additional comparison of the results.

[1] Loosely speaking, the revelation principal states that any outcome that can be implemented in equilibrium can also be truthfully implemented in equilibrium by having the mechanism play the equilibrium strategy on behalf of the truthful agents.

1.4 Additional Related Work

Our work is additionally related to the recent work on incentive compatible machine learning [23, 7, 4]. In these settings, the "social choice" being made is a machine learning predictor where agents benefit from there point having small error with respect to the chosen predictor. As in our setting, the information of the optimal model is distributed among the agents. Unlike our model, the private information and the preferences of the agents essentially coincide.

Information elicitation without verification, sometimes call peer prediction, is another very related line of research which shares some of the intuitions and techniques from information aggregation. The information elicitation literature has been well established in the past decades, starting from Prelec [24]'s Baysian Truth Serum and Miller et al. [22]'s peer-prediction method. These mechanisms cleverly design payments to the agents to guarantee that the truthful reporting of received information forms a Nash equilibrium.

A mass of recent work (see Faltings and Radanovic [8] for a survey) is dedicated to designing information elicitation mechanisms that work in more general settings (such as, supporting a small number of agents [5, 31, 26], allowing agents having information with different levels of sophistication [10, 17]), or achieving better truthful guarantees (such as, strict Nash equilibrium [27], informed Nash equilibrium [28], or even dominant strategy equilibrium [18, 15]) sometimes by studying more restrictive settings (e.g. multiple similar questions being asked simultaneously [5]).

Indeed, following Bayesian Truth Serum, many of these mechanism require the agents to predict other agents' reports [31, 16, 17, 18, 19, 26]

However, all these mechanisms rely on *payments* to the agents to incentivize truth-telling. In our social choice setting, on the other hand, we need to incentivize truth-telling solely based on choosing the winning alternative.

2 Model and Preliminaries

In our model, T agents are voting for two *alternatives*, $\mathbf{A}$ and $\mathbf{R}$ (corresponding to "accept" and "reject"). There is a set of N possible *worlds* (or *states*) $\mathcal{W} = \{1, \ldots, N\}$, where the higher the value the more $\mathbf{A}$ is preferred to $\mathbf{R}$. Agents do not know which world is the actual world that they are in. They have a prior common belief on the likelihood of each world. Let W be the actual world which is viewed as a random variable. Let $P_n = \Pr(W = n)$ be the prior over worlds. Each agent knows the values of $P_1, \ldots, P_n$ as prior believes. We further assume $P_n > 0$ for each n, for otherwise we can remove world n from $\mathcal{W}$ without loss of generality.

Each agent will then receives a *signal* from the set $\mathcal{S} = \{1, \ldots, M\}$. Let S_t be the random variable representing the signal that agent t receives. Given $W = n$, for any n, the signals agents receive have the same distribution and are conditionally independent. Let $P_{nm} = \Pr(S_t = m \mid W = n)$ be the probability that signal m will be received (by an arbitrary agent t) if the actual world is n. The set of values $\{P_{nm} : n = 1, \ldots, N; m = 1, \ldots, M\}$ is known by all the agents. Signals are positively correlated to the worlds:

$$\Pr(S_t \geq m \mid W = n_1) = \sum_{m'=m}^{M} P_{n_1 m'} > \Pr(S_t \geq m \mid W = n_2) = \sum_{m'=m}^{M} P_{n_2 m'} \tag{1}$$

for any worlds $n_1 > n_2$, any signal m, and any agent t.

Each agent t has a *utility function* $v_t : \mathcal{W} \times \{\mathbf{A}, \mathbf{R}\} \to \{0, 1, \ldots, B\}$. As mentioned earlier, higher value of W indicates $\mathbf{A}$ is more preferable: $v_t(n_1, \mathbf{A}) > v_t(n_2, \mathbf{A})$ and $v_t(n_1, \mathbf{R}) < v_t(n_2, \mathbf{R})$ for any $n_1, n_2 \in \mathcal{W}$ with $n_1 > n_2$. Since we can always rescale agent's utility, for simplicity, we assume without loss of generality that agents' utilities are integers and bounded by $B \in \mathbb{Z}^+$. Agents, with their prior believes and receiving signals, will have posterior believes about the distribution of W and react to the mechanism in a way maximizing their expected utilities.

We assume $v_t(n, \mathbf{A}) \neq v_t(n, \mathbf{R})$ for each agent t and each $n \in \mathcal{W}$, so that agents always strictly prefer one alternative over the other. Based on his/her utility function v_t, each agent has a threshold θ_t on the worlds such that $v_t(n, \mathbf{A}) > v_t(n, \mathbf{R})$ for all $n > \theta_t$ and $v_t(n, \mathbf{A}) < v_t(n, \mathbf{R})$ for all $n < \theta_t$. In other words, θ_t is the borderline that separates those "good" worlds where agent t prefers $\mathbf{A}$ to $\mathbf{R}$ and those "bad" worlds where agent t prefers $\mathbf{R}$ to $\mathbf{A}$. We will avoid $n = \theta_t$ by assuming $\theta_t \in \{0.5, 1.5, 2.5, \ldots, N + 0.5\}$ without loss

of generality. For example, if agent t prefers $\mathbf{A}$ to $\mathbf{R}$ for $W \geq 4$ and prefers $\mathbf{R}$ to $\mathbf{A}$ for $W \leq 3$, we will say $\theta_t = 3.5$. In particular, $\theta_t = 0.5$ implies that agent t always prefers $\mathbf{A}$ to $\mathbf{R}$, and $\theta_t = N + 0.5$ implies that agent t always prefers $\mathbf{R}$ to $\mathbf{A}$.

For each $\vartheta = 0.5, 1.5, 2.5, \ldots, N + 0.5$, let α_ϑ be the fraction of agents whose thresholds are ϑ. Naturally, $\sum_{\vartheta=0.5}^{N+0.5} \alpha_\vartheta = 1$. In most parts of this paper, we assume that $\alpha_{0.5}, \alpha_{1.5}, \ldots, \alpha_{N+0.5}$ are known by all agents. However, in Sect. 4, we discuss the scenario where this is not the case.

The goal is to output the alternative that is preferred by at least half of the agents, which is exactly the most common *majority vote* in the social choice literature. Equivalently, if $W = n^*$ is the actual world, alternative $\mathbf{A}$ should be output if $\sum_{\vartheta<n^*} \alpha_\vartheta \geq 0.5$, and alternative $\mathbf{R}$ should be output if otherwise.

We assume that the number of agents T is sufficiently large, and the parameters B, $\{P_n : n = 1, \ldots, N\}$, $\{P_{nm} : n = 1, \ldots, N; m = 1, \ldots, M\}$ and $\{\alpha_\vartheta : \vartheta = 0.5, 1.5, \ldots, N + 0.5\}$ are constants that do not depend on T. It is helpful to consider $T \to \infty$ throughout this paper. We further assume that T is an odd number to avoid ties.

2.1 The Running Example

In this section, we describe a running example that will be used throughout this paper.

Suppose a department of T faculty members, or agents, need to decide whether or not to hire a new faculty candidate. $\mathbf{A}$ and $\mathbf{R}$ naturally correspond to the two decisions: to accept or to reject this candidate.

The worlds $\mathcal{W} = \{1, \ldots, N\}$ describe the quality of the candidate, with 1 being the worst and N being the best. Those T agents do not know the true quality of the candidate, or the actual world they are in. They first see the CV of this candidate, and have a prior belief about the quality of this candidate, which corresponds to the knowledge of $P_1, \ldots, P_N$ in our model.

After this, each faculty member t will interview this candidate by an individual meeting, and obtain a signal $S_t \in \{1, \ldots, M\}$ which corresponds to the impression of this candidate, with $S_t = 1$ being the worst impression and $S_t = M$ being the best impression. It is natural to assume that S_t's are positively correlated to W, which agrees with our model.

Naturally, different agents may have different thresholds on the quality of the candidate. For example, if a theory candidate is interviewed in a computer science department, the theory faculty members may have low thresholds for this candidate, however, the AI, software and hardware faculty members may have high thresholds. Since the number of theory, AI, software and hardware faculty members are public information, it is natural to assume $\alpha_{0.5}, \alpha_{1.5}, \ldots, \alpha_{N+0.5}$ are known, as it is in our model. Admittedly, agents may not have full information on these values in some other scenarios. We will discuss a model where agents only have partial information on $\alpha_{0.5}, \alpha_{1.5}, \ldots, \alpha_{N+0.5}$ in Sect. 4.

The mechanism should aggregate the information from the T (strategic) agents and output a decision that is favored by the majority.

2.2 Strategy and ε-Strong Bayes Nash Equilibrium

A *mechanism* collects a *report* from each agent, and then outputs an alternative which is either $\mathbf{A}$ or $\mathbf{R}$. The mechanism specifies the content of the report by specifying questions for the agents. Examples of those questions include asking each agent for the signal he/she receives, asking each agent to predict the other agents' reports, etc.

Let $\mathcal{R}$ be the space of all possible reports, which depends on the design of the mechanism. A *pure strategy* of an agent is given by a function $\sigma : \mathcal{S} \to \mathcal{R}$ that maps a signal received by this agent to a report. The *mixed strategy* is defined in a way standard in game theory, so that σ can be a random function.

An agent's strategy is *truthful* if it always specifies the correct answer to each question in the report, to the best of the agent's knowledge after receiving the signal. For example, if the mechanism asks for the agent's signal, an agent playing the truthful strategy should report the signal he/she receives; if the mechanism asks the agents to predict the fraction of agents who will receive signal m, an agent playing the truthful strategy should report his/her posterior belief on this computed by the Bayes rule (Sect. 2.3 discusses the computation of posterior belief).

Given a strategy profile $\Sigma = (\sigma_1, \ldots, \sigma_T)$, let $u_t(\Sigma)$ be the expected utility of agent t, where the expectation is taken over the sampling of agents' signals. Notice that we use u_t to denote the *ex-ante* utility (as

defined just now) and we have used v_t to denote the *ex-post* utility (see the third paragraph in Sect. 2). Most parts of this paper will focus on the *ex-ante* utility, especially when we are talking about any equilibrium solution concept.

Since we are in a social choice setting with a large number of agents, a single agent's behavior may not have much effect. Thus, instead of the typical Bayes Nash Equilibrium, we consider a much stronger goal—the Strong Bayes Nash Equilibrium.

Definition 2.1. A strategy profile $(\sigma_1, \ldots, \sigma_T)$ is an *ε-Strong Bayes Nash Equilibrium* if there does not exist a subset of agents D and a strategy profile $(\sigma'_1, \ldots, \sigma'_T)$ such that

1. $\sigma_t = \sigma'_t$ for each $t \notin D$,
2. $u_t(\sigma'_1, \ldots, \sigma'_T) \geq u_t(\sigma_1, \ldots, \sigma_T)$ for each $t \in D$, and
3. there exist $t \in D$ such that $u_t(\sigma'_1, \ldots, \sigma'_T) > u_t(\sigma_1, \ldots, \sigma_T) + \varepsilon$.

Notice that a 0-Strong Bayes Nash Equilibrium is the conventional Strong Bayes Nash Equilibrium. When a strategy profile $(\sigma_1, \ldots, \sigma_n)$ is not an ε-Strong Bayes Nash Equilibrium, we will call the subset of the agents D in Definition 2.1 the *deviating agents* or the *deviating coalition*. Note that the larger ε is, the harder it is to find a deviating coalition, and so the the larger the set of ε-Strong Bayes Nash Equilibrium.

2.3 Posterior Update by Bayes Rule

Upon receiving a signal $S_t \in \{1, \ldots, M\}$, agent t updates his/her posterior belief (about the probability that (s)he is in world n for each $n = 1, \ldots, N$, the fraction of agents that will receive signal m for each $m = 1, \ldots, M$, etc.) based on Bayes rule. Below we consider an arbitrary $m \in \mathcal{S}$ and an arbitrary $n \in \mathcal{W}$.

Suppose agent t receives signal $S_t = m$. (S)he believes that the actual world is n with the following probability:

$$\Pr(W = n \mid S_t = m) = \frac{\Pr(W = n) \cdot \Pr(S_t = m \mid W = n)}{\Pr(S_t = m)} = \frac{P_n P_{nm}}{P_1 P_{1m} + P_2 P_{2m} + \cdots + P_N P_{Nm}},$$

Furthermore, this agent believes that another arbitrary agent will receive signal m' with probability

$$\begin{aligned} T_{mm'} &= P_{1m'} \cdot \Pr(W = 1 \mid S_t = m) + P_{2m'} \cdot \Pr(W = 2 \mid S_t = m) + \cdots + P_{Nm'} \cdot \Pr(W = N \mid S_t = m) \\ &= \frac{P_1 P_{1m} P_{1m'} + P_2 P_{2m} P_{2m'} + \cdots + P_N P_{Nm} P_{Nm'}}{P_1 P_{1m} + P_2 P_{2m} + \cdots + P_N P_{Nm}}. \end{aligned} \tag{2}$$

In this paper, we will use $T_{m_1 m_2}$ to denote the probability that an agent who receives signal m_1 believes that another agent will receive signal m_2.

2.4 Additional Notations

Given a strategy profile $\Sigma = \{\sigma_1, \ldots, \sigma_T\}$ and a mechanism $\mathcal{M}$, let $\lambda_n^{\mathcal{M}}(\Sigma)$ be the probability that alternative **A** is announced as the winner given the actual world is n, then $1 - \lambda_n^{\mathcal{M}}(\Sigma)$ is the probability that alternative **R** wins given the actual world is 1. We will omit the superscript $\mathcal{M}$ when it is clear what mechanism we are discussing.

All the agents' *ex-ante* utilities depend exclusively on $\lambda_1(\Sigma), \ldots, \lambda_N(\Sigma)$, and each agent t's utility is given by

$$u_t(\Sigma) = \sum_{n=1}^{N} P_n \left(\lambda_n(\Sigma) v_t(n, \mathbf{A}) + (1 - \lambda_n(\Sigma)) v_t(n, \mathbf{R})\right) \tag{3}$$

$$= \sum_{n=1}^{N} P_n v_t(n, \mathbf{R}) + \sum_{n=1}^{N} P_n \lambda_n(\Sigma)(v_t(n, \mathbf{A}) - v_t(n, \mathbf{R})). \tag{4}$$

Specifically, when we consider the setting with two worlds $N = 2$, the (*ex-ante*) utility of an arbitrarily agent t is

$$u_t(\Sigma) = P_1 \left(\lambda_1(\Sigma)v_t(1, \mathbf{A}) + (1 - \lambda_1(\Sigma))v_t(1, \mathbf{R})\right) + P_2 \left(\lambda_2(\Sigma)v_t(2, \mathbf{A}) + (1 - \lambda_2(\Sigma)v_t(2, \mathbf{R}))\right), \quad (5)$$

which can also be rewritten as

$$u_t(\Sigma) = P_1 v_t(1, \mathbf{A}) + P_2 v_t(2, \mathbf{R}) + P_1(1 - \lambda_1(\Sigma))(v_t(1, \mathbf{R}) - v_t(1, \mathbf{A})) + P_2\lambda_2(\Sigma)(v_t(2, \mathbf{A}) - v_t(2, \mathbf{R})), \quad (6)$$

and

$$u_t(\Sigma) = P_1 v_t(1, \mathbf{R}) + P_2 v_t(2, \mathbf{R}) + P_1\lambda_1(\Sigma)(v_t(1, \mathbf{A}) - v_t(1, \mathbf{R})) + P_2\lambda_2(\Sigma)(v_t(2, \mathbf{A}) - v_t(2, \mathbf{R})). \quad (7)$$

We will always use $\Sigma^* = \{\sigma_1^*, \ldots, \sigma_T^*\}$ to denote the truthful strategy profile.

3 Two Worlds and Two Signals

In this section, we discuss the scenario with two worlds $N = 2$ and two signals $M = 2$. This captures many essential ideas behind our mechanism for general cases. Moreover, in many scenarios, it is reasonable and even more practical to assume that the worlds and the signals can only be either "good" or "bad". For example, it is much easier for an agent to distinguish good and bad than to rank the quality using numerical numbers from 1 to M. In addition, numerical ranking causes more subjective systematic bias, and the heterogeneity of the bias among the agents makes agents' reports more noisy. Finally, our mechanism can be much easier to implement under the setting with $N = 2$ and $M = 2$ (we will discuss more on this later), which makes our mechanism more appealing in practice.

In our faculty hiring running example, the two worlds $W = 1$ and $W = 2$ correspond to the candidate being "bad" and "good" respectively. Similarly, the two signals $S_t = 1$ and $S_t = 2$ correspond to the impression of agent t on the candidate being "bad" and "good" respectively.

Throughout this section, we will define the agents having threshold 0.5 as the *candidate-friendly* agents, the agents having threshold 1.5 as the *contingent* agents, and the agents having threshold 2.5 as the *candidate-unfriendly* agents. Those candidate-friendly agents will always prefer to accept the candidate, those candidate-unfriendly will always prefer to reject, and those contingent agents will prefer to accept the candidate if and only if the candidate is good (i.e., $W = 2$).

The classification of agents into three categories is natural in many settings, for instance, the election example in introduction.

We will first review Prelec et al.'s *Surprising Popular* algorithm [25], which works under a setting similar to ours but with non-strategic agents. Some part of the intuition behind our mechanism is based on Prelec et al.'s work.

3.1 Prelec et al.'s Surprising Popular Algorithm

In this section, we review the Surprising Popular algorithm proposed by Prelec et al. [25], which works under the setting similar to ours but with non-strategic agents. For the purpose of this paper, we will describe the algorithm with two worlds and two signals.

The algorithm asks each agent t the signal (s)he receives, and his/her belief on the fraction of agents who have received signal 1 (or signal 2). In our notation, each agent reports the realization of S_t and, assuming $S_t = m$, the value of T_{m1} (or T_{m2}, which equals to $1 - T_{m1}$). Since agents are assumed to be non-strategic, those who receive signal 1 will report $(1, T_{11})$ and those who receive signal 2 will report $(2, T_{21})$. The algorithm then computes the fraction of agents who report signal 1, and the average value of all the reported T_{m1}'s. If the former is greater than the latter, 1 is considered as being "surprisingly popular" and the algorithm will conclude that 1 is the actual world. Otherwise, 2 will be considered as being "surprisingly popular" and 2 will be concluded as being the actual world.

The correctness of this algorithm is based on the following simple yet important observation in Theorem 3.1. In particular, the average of agents' reported predictions (those T_{m1}'s) will be between T_{21} and T_{11}. When the number of agents T is sufficiently large, the actual fraction of agents who vote for 1 will be either approximately P_{11} (if 1 is the actual world) or approximately P_{21} (if 2 is the actual world). Theorem 3.1 then implies the correctness of the Surprisingly Popular algorithm.

Theorem 3.1. $P_{21} < T_{21} < T_{11} < P_{11}$.

The intuition is straightforward. Recall P_{21} and P_{11} are the expected fraction of agents receiving signal 1 in worlds 2 and 1 respectively. By assumption $P_{21} < P_{11}$ (equation (1)). Agents have some posterior belief over the world they are in. Thus their expected value of the number of 1's agents receive is a weighted average of P_{21} and P_{11}, and therefore in between both values. Agents that receive a 1 signal believe that world 1 is more likely than agents that receive a 2 signal. Thus, $T_{21} < T_{11}$.

Proof. By taking $n_1 = 2$, $n_2 = 1$, $m = 2$ in (1), we have $P_{22} > P_{12}$, so $P_{21} = 1 - P_{22} < 1 - P_{12} = P_{11}$. By (2) and $P_{21} < P_{11}$, we have

$$T_{11} = \frac{P_1 P_{11}^2 + P_2 P_{21}^2}{P_1 P_{11} + P_2 P_{21}} < \frac{P_1 P_{11}^2 + P_2 P_{21} P_{11}}{P_1 P_{11} + P_2 P_{21}} = P_{11}$$

and

$$T_{21} = \frac{P_1 P_{11} P_{12} + P_2 P_{21} P_{22}}{P_1 P_{12} + P_2 P_{22}} > \frac{P_1 P_{21} P_{12} + P_2 P_{21} P_{22}}{P_1 P_{12} + P_2 P_{22}} = P_{21}.$$

Finally, to show $T_{11} > T_{21}$, it suffices to show that

$$\lambda_1 := \frac{P_1 P_{11}}{P_1 P_{11} + P_2 P_{21}} > \lambda_2 := \frac{P_1 P_{12}}{P_1 P_{12} + P_2 P_{22}},$$

since $T_{11} = \lambda_1 P_{11} + (1 - \lambda_1) P_{21}$, $T_{21} = \lambda_2 P_{11} + (1 - \lambda_2) P_{21}$, and $P_{11} > P_{21}$. Simple calculations show this: $\lambda_1 > \frac{P_1}{P_1 + P_2} > \lambda_2$, where the first inequality is due to $P_{11} > P_{21}$ and the second inequality is due to $P_{22} > P_{12}$. □

Throughout this section, we use C to denote the following constant.

$$C = \frac{1}{3} \min \{T_{21} - P_{21}, T_{11} - T_{21}, P_{11} - T_{11}\} \tag{8}$$

3.2 The Wisdom of the Crowd Voting Mechanism

Our mechanism is shown in Mechanism 1.

Mechanism 1 The Wisdom of the Crowd Voting Mechanism for $M = N = 2$

1: Each agent t reports to the mechanism the signal (s)he receives (either 1 or 2), his/her threshold θ_t (0.5, 1.5 or 2.5; candidate-friendly, contingent, or candidate-unfriendly), his/her posterior belief of the fraction of agents who will report signal 2. As a result, the report space is $\mathcal{R} = \{1, 2\} \times \{0.5, 1.5, 2.5\} \times [0, 1]$. Let $(\bar{s}_t, \bar{\theta}_t, \bar{\delta}_t)$ be t's report.

2: If agent t reports $\bar{\theta}_t = 0.5$, his reported signal will be automatically treated as $\bar{s}_t = 2$; if agent t reports $\bar{\theta}_t = 2.5$, his reported signal will be automatically treated as $\bar{s}_t = 1$. *The prediction $\bar{\delta}_t$ in the previous step should be made with this treatment being considered, and the mechanism makes this clear to the agents.*

3: Compute the *median* of the reported $\bar{\delta}_t$, denoted by $\bar{\delta}$.

4: If more than half of the agents report $\bar{\theta}_t = 0.5$, announce **A** being the winning alternative; if more than half of the agents reports $\bar{\theta}_t = 2.5$, announce **R** being the winning alternative.

5: If the number of agents reporting $\bar{s}_t = 2$ is more than the median $\bar{\delta}$, announce **A** being the winning alternative; otherwise, announce **R** being the winning alternative.

In our running example, the questionnaire corresponding to our mechanism looks like the followings.

An example for the questionnaire:

1. What was your impression of this candidate during the individual interview between you and this candidate?
   (a) I had a good impression.
   (b) I did not have a good impression.
2. Choose one of the following.
   (a) I definitely want to accept this candidate.
   (b) I definitely want to reject this candidate.
   (c) I am open to considering the inputs of others.

   If your choice is (a), your answer in Q1 will be treated as (a); if your choice is (b), your answer in Q1 will be treated as (b).
3. What percentage of the faculty do you believe had a good impression during the individual interview? Please answer this question assuming the treatment in the previous question. That is, report the fraction of faculty you believe either definitely want to accept this candidate or are open to considering the inputs of others and had a good impression.

3.3 Main Theoretical Results for Two Worlds and Two Signals

We first show that our mechanism indeed achieves (with an exponentially small failure probability) the goal of outputting the alternative favored by the majority, assuming agents are truth-telling.

Given a strategy profile Σ, we define $I(\Sigma) \equiv P_1\lambda_1(\Sigma) + P_2(1-\lambda_2(\Sigma))$. Notice that, when we have $\alpha_{2.5} > 0.5$ and $\alpha_{0.5} > 0.5$, $I(\Sigma)$ is the error rate of a strategy: the probability the mechanism selects the alternative that is not preferred by the majority.

Theorem 3.2. *If all the agents play the truthful strategy Σ^*, then, with probability at least $1-2\exp(-2C^2\alpha_{1.5}T)$ (where C is the constant defined in Eqn. (8)), our mechanism outputs an alternative that is favored by more than half of the agents.*

Proof. Suppose all the agents report truthfully.

If $\alpha_{2.5} > 0.5$ or $\alpha_{0.5} > 0.5$, Step 4 of the mechanism guarantees that the alternative favored by more than half of the agents will be announced with probability 1, which implies the theorem. It remains to consider the case with $\alpha_{2.5} < 0.5$ and $\alpha_{0.5} < 0.5$ (recall that we have assumed T is an odd number, so $\alpha_{2.5}$ and $\alpha_{0.5}$ cannot be 0.5). In this case, **A** is favored by the majority if the actual world is 2, and **R** is favored by the majority if the actual world is 1.

If an agent t receives signal 1, (s)he will believe that a T_{11} fraction of agents receive 1, and (s)he will report $\bar{\delta}_t = \alpha_{1.5}T_{11} + \alpha_{2.5}$. Similarly, an agent receiving signal 2 will report $\bar{\delta}_t = \alpha_{1.5}T_{21} + \alpha_{2.5}$. Therefore, $\bar{\delta} \in [\alpha_{1.5}T_{21} + \alpha_{2.5}, \alpha_{1.5}T_{11} + \alpha_{2.5}]$.

Suppose the actual world is 1. The expected fraction of the agents receiving signal 1 would be P_{11}, and the expected fraction of the agents reporting signal 1 (after the treatment in Step 2) would be $\alpha_{1.5}\cdot P_{11}+\alpha_{2.5}$. By a Chernoff bound, with probability at least $1-2\exp(-2C^2\alpha_{1.5}T)$, the fraction of agents reporting signal 1 is in the interval $[\alpha_{1.5}\cdot(P_{11}-C)+\alpha_{2.5}, \alpha_{1.5}\cdot(P_{11}+C)+\alpha_{2.5}]$, which is greater than $\alpha_{1.5}\cdot T_{11}+\alpha_{2.5} \geq \bar{\delta}$ by Theorem 3.1 and (8). Step 5 of the mechanism indicates that **R** will be announced.

The analysis for the case where 2 is the actual world is similar. □

Next, we show that the truthful strategy profile is an ε-Strong Bayes Nash Equilibrium of our mechanism for some exponentially small ε.

Theorem 3.3. *The truthful strategy profile is an ε-Strong Bayes Nash Equilibrium, where $\varepsilon = (2B^2 + 4B)\exp(-2C^2\alpha_{1.5}T)$.*

3.4 Proof Sketch for Theorem 3.3

We first describe a sketch of the proof. We consider three cases: 1) $\alpha_{0.5} > 0.5$, 2) $\alpha_{2.5} > 0.5$ and 3) $\alpha_{0.5} < 0.5$ and $\alpha_{2.5} < 0.5$.

Case 1) $\alpha_{0.5} > 0.5$**.** For the first case, more than half of the agents are candidate-friendly, and **A** will be announced according to Step 4 of the mechanism if these agents report truthfully. The truthful strategy profile forms a (0-)Strong Bayes Nash Equilibrium, as those candidate-friendly agents receive their maximum utilities by truth-telling and the remaining agents are not able to stop the mechanism from outputting **A** regardless of what they report.

Case 2) $\alpha_{2.5} > 0.5$**.** The analysis for the second case is the analogous to the first.

Case 3) $\alpha_{0.5} < 0.5$ **and** $\alpha_{2.5} < 0.5$**.** Under the third case, **A** is favored by the majority if the actual world is 2, and **R** is favored by the majority if the actual world is 1. By the same analysis as in the proof of Theorem 3.2, supposing agents report truthfully, we know that **A** will be output with probability at least $1 - 2\exp(-2C^2\alpha_{1.5}T)$ if the actual world is 2, and **R** will be output with probability at least $1 - 2\exp(-2C^2\alpha_{1.5}T)$ if the actual world is 1. Therefore, the contingent agents (with $\theta_t = 1.5$) receive very near their maximum utilities, and thus have no incentive to deviate from the truthful strategy.

To conclude that truth-telling is an ε-Strong Bayes Nash Equilibrium, we will study two cases and show that no coalition of deviating agents D exists in either case. Recall that in a deviating coalition all agents must benefit and some agent must benefit by at least ε. Let Σ' be the strategy profile after D's deviation.

In the first case, $I(\Sigma')$ is small, so the mechanism nearly always chooses the majority preferred alternative under Σ'. No agent can be much better off because all the agents have nearly the same utilities as before (Claim 3.6).

In the second case, $I(\Sigma')$ is not small, so the mechanism sometimes fails to choose the majority preferred alternative under Σ'. Here, the contingent agents do not fare better (Claim 3.7) and thus no contingent agent can be in the deviating coalition.

The technical key is then to show that D cannot contain both a candidate-friendly and a candidate-unfriendly agent. Lemma 3.4 shows that a significant increase in a candidate-friendly agent's (*ex-ante*) utility always results a decrease in a candidate-unfriendly agent's (*ex-ante*) utility, and vice versa. This is obvious if we are dealing with *ex-post* utilities, as a candidate-friendly agent and a candidate-unfriendly agent always want the opposite alternatives. However, this becomes much less obvious for *ex-ante* utilities.

Thus, any deviating coalition can only be comprised of either candidate-friendly voters or candidate-unfriendly voters. Finally, we show that a minority coalition comprised of only candidate-friendly voters, cannot change the outcome by misreporting.

The proof of Lemma 3.4 depends on i) truth-telling nearly always selecting the majority preferred alternative; and ii) the monotonicity of the *ex-post* utilities $v_t(\cdot, \mathbf{A})$ and $v_t(\cdot, \mathbf{R})$ in the first argument. In particular, Lemma 3.4 does not hold if truth-telling is replaced by an arbitrary strategy profile.

3.5 No Win-win Lemma

A key part of the proof is the following lemma which states that it is not possible that predetermined voters of different alternatives both gain from deviating from truthful. As a corollary, any deviating coalition can only contain predetermined voters of one type.

Lemma 3.4. *Suppose* $\alpha_{0.5} < 0.5$ *and* $\alpha_{2.5} < 0.5$*. Let* Σ^* *be the truthful strategy profile and* Σ' *be an arbitrary strategy profile. Let* t_1 *be an arbitrary candidate-friendly agent and* t_2 *be an arbitrary candidate-unfriendly agent. For any* $\Delta \geq 2B\exp(-2C^2\alpha_{1.5}T)$*, we have*

(i) *If* $u_{t_1}(\Sigma') - u_{t_1}(\Sigma^*) > \Delta$*, then* $u_{t_2}(\Sigma') - u_{t_2}(\Sigma^*) < 0$*.*

(ii) *If* $u_{t_2}(\Sigma') - u_{t_2}(\Sigma^*) > \Delta$*, then* $u_{t_1}(\Sigma') - u_{t_1}(\Sigma^*) < 0$*.*

Proof. We will only show 1, as the proof for 2 is similar.

Since $v_t(1, \mathbf{A}) < v_t(2, \mathbf{A})$ and $v_t(1, \mathbf{R}) > v_t(2, \mathbf{R})$ for any agent t, we have

$$0 < v_{t_1}(1, \mathbf{A}) - v_{t_1}(1, \mathbf{R}) < v_{t_1}(2, \mathbf{A}) - v_{t_1}(2, \mathbf{R}) \quad \text{and} \quad v_{t_2}(1, \mathbf{R}) - v_{t_2}(1, \mathbf{A}) > v_{t_2}(2, \mathbf{R}) - v_{t_2}(2, \mathbf{A}) > 0. \tag{9}$$

By referring to (7), this intuitively says that t_1's utility difference $u_{t_1}(\Sigma') - u_{t_1}(\Sigma^*)$ is more sensitive to λ_2 while u_2's utility difference $u_{t_2}(\Sigma') - u_{t_2}(\Sigma^*)$ is more sensitive to λ_1.

This provides the intuition for the proof of the Lemma. Both candidates will gain in utility if errors are decreased, but this is not possible to an non-negligible extent. If errors are increased, then at least one candidate is strictly worse off.

Suppose $u_{t_1}(\Sigma') - u_{t_1}(\Sigma^*) > \Delta$ as it is assumed in (i). By (7), we have

$$P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_{t_1}(1, \mathbf{A}) - v_{t_1}(1, \mathbf{R})) + P_2(\lambda_2(\Sigma') - \lambda_2(\Sigma^*))(v_{t_1}(2, \mathbf{A}) - v_{t_1}(2, \mathbf{R})) > \Delta. \tag{10}$$

We consider two cases: $\lambda_2(\Sigma') \geq \lambda_2(\Sigma^*)$ and $\lambda_2(\Sigma') < \lambda_2(\Sigma^*)$.

Case 1: $\lambda_2(\Sigma') \geq \lambda_2(\Sigma^*)$**.** By Theorem 3.2, we have $\lambda_2(\Sigma^*) \geq 1 - 2\exp(-2C^2\alpha_{1.5}T) > 0$, which implies $\lambda_2(\Sigma') - \lambda_2(\Sigma^*) \leq 2\exp(-2C^2\alpha_{1.5}T)$, which further implies

$$P_2(\lambda_2(\Sigma') - \lambda_2(\Sigma^*))(v_{t_1}(2, \mathbf{A}) - v_{t_1}(2, \mathbf{R})) \leq P_2 \cdot 2\exp(-2C^2\alpha_{1.5}T) \cdot B < \Delta.$$

Putting this into (10), we have $P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_{t_1}(1, \mathbf{A}) - v_{t_1}(1, \mathbf{R}))$, which implies $\lambda_1(\Sigma') > \lambda_1(\Sigma^*)$. We then must have $u_{t_2}(\Sigma') - u_{t_2}(\Sigma^*) < 0$ since we have

$$u_{t_2}(\Sigma') - u_{t_2}(\Sigma^*) = P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_{t_2}(1, \mathbf{A}) - v_{t_2}(1, \mathbf{R})) + P_2(\lambda_2(\Sigma') - \lambda_2(\Sigma^*))(v_{t_2}(2, \mathbf{A}) - v_{t_2}(2, \mathbf{R}))$$

by (7), $\lambda_2(\Sigma') \geq \lambda_2(\Sigma^*)$ (Case 1 assumption), $\lambda_1(\Sigma') > \lambda_1(\Sigma^*)$ (we have just shown), $v_{t_2}(1, \mathbf{A}) - v_{t_2}(1, \mathbf{R}) < 0$ and $v_{t_2}(2, \mathbf{A}) - v_{t_2}(2, \mathbf{R}) < 0$ (since t_2 is candidate-unfriendly).

Case 2: $\lambda_2(\Sigma') < \lambda_2(\Sigma^*)$**.** By (10) and $\Delta > 0$, we have

$$P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_{t_1}(1, \mathbf{A}) - v_{t_1}(1, \mathbf{R})) > P_2(\lambda_2(\Sigma^*) - \lambda_2(\Sigma'))(v_{t_1}(2, \mathbf{A}) - v_{t_1}(2, \mathbf{R})),$$

which, by (9), further implies

$$P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*)) > P_2(\lambda_2(\Sigma^*) - \lambda_2(\Sigma')).$$

By (9) again, this implies

$$P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_{t_2}(1, \mathbf{R}) - v_{t_2}(1, \mathbf{A})) > P_2(\lambda_2(\Sigma^*) - \lambda_2(\Sigma'))(v_{t_2}(2, \mathbf{R}) - v_{t_2}(2, \mathbf{A})),$$

which further implies

$$\begin{aligned} u_{t_2}(\Sigma') - u_{t_2}(\Sigma^*) = P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_{t_2}(1, \mathbf{A}) - v_{t_2}(1, \mathbf{R})) \\ + P_2(\lambda_2(\Sigma') - \lambda_2(\Sigma^*))(v_{t_2}(2, \mathbf{A}) - v_{t_2}(2, \mathbf{R})) < 0. \end{aligned}$$

The lemma concludes. □

Corollary 3.5. *Suppose $\alpha_{0.5} < 0.5$ and $\alpha_{2.5} < 0.5$. The set of deviating agents D cannot contain both candidate friendly and candidate unfriendly agents.*

Proof. By 3 in Definition 2.1, there must be an agent t such that $u_t(\Sigma') - u_t(\Sigma^*) > \varepsilon \geq 2B\exp(-2C^2\alpha_{1.5}T)$. Assume this agent is candidate-friendly. Then by Lemma 3.4, for any candidate-unfriendly agent t', we have $u_{t'}(\Sigma') - u_{t'}(\Sigma^*) < 0$. Thus no candidate-unfriendly agent can be in the deviating coalition.

An analogous argument works if the benefiting agent is candidate-unfriendly. □

3.6 Proof of Theorem 3.3

Now we are ready to prove Theorem 3.3.

Suppose this is not the case. There exists a set of deviating agents D that can deviate from the truthful strategy such that all of them receive utilities that are at least their original utilities and some of them receive utilities that are ε strictly higher than their original utilities. Let Σ' be the strategy profile after agents in D deviate.

We discuss three different cases: 1) $\alpha_{0.5} > 0.5$, 2) $\alpha_{2.5} > 0.5$ and 3) $\alpha_{0.5} < 0.5$ and $\alpha_{2.5} < 0.5$. Notice that n being odd implies neither $\alpha_{0.5}$ nor $\alpha_{2.5}$ can be exactly 0.5.

Case 1: $\alpha_{0.5} > 0.5$. If all agents report truthfully, **A** will be announced with probability 1 according to Step 4 of the mechanism. That is, $\lambda_1(\Sigma^*) = \lambda_2(\Sigma^*) = 1$. By 3 of Definition 2.1, there exists $t \in D$ such that $u_t(\Sigma') > u_t(\Sigma^*) + \varepsilon$. Since $\lambda_1(\Sigma')$ and $\lambda_2(\Sigma')$ completely determine each agent's utility, we must have either $\lambda_1(\Sigma') \neq \lambda_1(\Sigma^*)$ or $\lambda_2(\Sigma') \neq \lambda_2(\Sigma^*)$. This means either $\lambda_1(\Sigma') < 1$ or $\lambda_2(\Sigma') < 1$.

Since a candidate-friendly agent's utility is maximized when both λ_1 and λ_2 are 1, a candidate-friendly agent's utility will decrease if the strategy profile is switched from Σ^* to Σ'. By 2 of Definition 2.1, D does not contain any candidate-friendly agent. However, if this is the case, there are still more than half of the agents that will report $\bar{\theta}_t = 0.5$ (as $\alpha_{0.5} > 0.5$), and **A** will always be announced by Step 4 of the mechanism. We conclude that $\lambda_1(\Sigma') = \lambda_2(\Sigma') = 1$, which contradicts to what we have concluded in the previous paragraph.

Case 2: $\alpha_{2.5} > 0.5$. The analysis is similar to the previous case.

Case 3: $\alpha_{0.5} < 0.5$ and $\alpha_{2.5} < 0.5$. We consider two sub-cases: $I(\Sigma') < (2B+2)\exp(-2C^2\alpha_{1.5}T)$ and $I(\Sigma') \geq (2B+2)\exp(-2C^2\alpha_{1.5}T)$

Firstly, we consider $I(\Sigma') < (2B+2)\exp(-2C^2\alpha_{1.5}T)$.

Claim 3.6. *If $\alpha_{0.5} < 0.5$, $\alpha_{2.5} < 0.5$ and $I(\Sigma') < (2B+2)\exp(-2C^2\alpha_{1.5}T)$, then $u_t(\Sigma') - u_t(\Sigma) < \varepsilon$ for every agent t.*

The ideas behind this proof is that the outcome for Σ' is too close to the outcome of the truthful strategy profile Σ^*, so no agent can get significantly more benefit.

Proof. The proof of this claim shows that none of the three types of agents can benefit by ε because nothing is substantially different from when agents play truthfully.

By rearranging the inequality $I(\Sigma') = P_1\lambda_1(\Sigma') + P_2(1-\lambda_2(\Sigma')) < (2B+2)\exp(-2C^2\alpha_{1.5}T)$ and noticing $\lambda_1(\Sigma'), \lambda_2(\Sigma') \in [0,1]$ and $P_1 + P_2 = 1$, we have

$$1 \geq \lambda_2(\Sigma') \geq 1 - \frac{(2B+2)\exp(-2C^2\alpha_{1.5}T)}{P_2}$$

and

$$0 \leq \lambda_1(\Sigma') \leq \frac{(2B+2)\exp(-2C^2\alpha_{1.5}T)}{P_1}.$$

Therefore, for any candidate-friendly agent, we have

$$\begin{aligned}
u_t(\Sigma') - u_t(\Sigma^*) &= P_1(\lambda_1(\Sigma') - \lambda_1(\Sigma^*))(v_t(1,\mathbf{A}) - v_t(1,\mathbf{R})) + P_2(\lambda_2(\Sigma') - \lambda_2(\Sigma^*))(v_t(2,\mathbf{A}) - v_t(2,\mathbf{R})) \\
&\leq P_1\left(\frac{(2B+2)\exp(-2C^2\alpha_{1.5}T)}{P_1} - 0\right)B + P_2\left(1 - (1 - 2\exp(-2C^2\alpha_{1.5}T))\right)B \\
&< (2B^2 + 4B)\exp(-2C^2\alpha_{1.5}T) = \varepsilon.
\end{aligned}$$

For any contingent agent, we have

$$\begin{aligned}
u_t(\Sigma') - u_t(\Sigma^*) &= P_1(\lambda_1(\Sigma^*) - \lambda_1(\Sigma'))(v_t(1,\mathbf{R}) - v_t(1,\mathbf{A})) + P_2(\lambda_2(\Sigma') - \lambda_2(\Sigma^*))(v_t(2,\mathbf{A}) - v_t(2,\mathbf{R})) \\
&\leq P_1(2\exp(-2C^2\alpha_{1.5}T) - 0)B + P_2\left(1 - (1 - 2\exp(-2C^2\alpha_{1.5}T)\right)B \\
&= 2B\exp(-2C^2\alpha_{1.5}T) < \varepsilon.
\end{aligned}$$

For any candidate-unfriendly agent, we have

$$\begin{aligned}u_t(\Sigma') - u_t(\Sigma^*) &= P_1(\lambda_1(\Sigma^*) - \lambda_1(\Sigma'))(v_t(1,\mathbf{R}) - v_t(1,\mathbf{A})) + P_2(\lambda_2(\Sigma^*) - \lambda_2(\Sigma'))(v_t(2,\mathbf{R}) - v_t(2,\mathbf{A}))\\ &\le P_1(2\exp(-2C^2\alpha_{1.5}T) - 0)B + P_2\left(1 - \left(1 - \frac{(2B+2)\exp(-2C^2\alpha_{1.5}T)}{P_2}\right)\right)B\\ &< (2B^2 + 4B)\exp(-2C^2\alpha_{1.5}T) = \varepsilon.\end{aligned}$$

We conclude that none of the agents has a utility gain of at least ε, which contradict 3 of Definition 2.1. □

Next, we consider the second case $I(\Sigma') \ge (2B+2)\exp(-2C^2\alpha_{1.5}T)$. This case is more complicated. We first show that no contingent agent in Σ' can do as well as in the truthful profile Σ^*.

Claim 3.7. *If $\alpha_{0.5} < 0.5$, $\alpha_{2.5} < 0.5$ and $I(\Sigma') \ge (2B+2)\exp(-2C^2\alpha_{1.5}T)$, then $u_t(\Sigma') - u_t(\Sigma^*) < 0$ for every contingent agent t.*

The ideas behind this proof is that those contingent agents already receive almost optimal utilities in Σ^*; therefore, if the error rate of the strategy Σ' is high enough, the utilities of the contingent agents will decrease.

Proof. By Theorem 3.2, we have

$$\begin{aligned}P_1(1 - \lambda_1(\Sigma^*)) + P_2\lambda_2(\Sigma^*) &\ge P_1(1 - 2\exp(-2C^2\alpha_{1.5}T)) + P_2(1 - 2\exp(-2C^2\alpha_{1.5}T))\\ &= 1 - 2\exp(-2C^2\alpha_{1.5}T). \qquad (11)\end{aligned}$$

Therefore

$$\begin{aligned}P_1(1 - \lambda_1(\Sigma^*)) + P_2\lambda_2(\Sigma^*) &\ge 1 - 2\exp(-2C^2\alpha_{1.5}T)\\ &= 2B\exp(-2C^2\alpha_{1.5}T) + 1 - (2B+2)\exp(-2C^2\alpha_{1.5}T)\\ &\ge 2B\exp(-2C^2\alpha_{1.5}T) + 1 - I(\Sigma')\\ &= P_1(1 - \lambda_1(\Sigma')) + P_2\lambda_2(\Sigma') + 2B\exp(-2C^2\alpha_{1.5}T) \qquad (12)\end{aligned}$$

By (6), we have

$$\begin{aligned}&u_t(\Sigma') - u_t(\Sigma^*)\\ =&P_1(v_t(1,\mathbf{R}) - v_t(1,\mathbf{A}))((1 - \lambda_1(\Sigma')) - (1 - \lambda_1(\Sigma^*))) + P_2(v_t(2,\mathbf{A}) - v_t(2,\mathbf{R}))(\lambda_2(\Sigma') - \lambda_2(\Sigma^*)) \qquad (13)\end{aligned}$$

We will show $u_t(\Sigma') - u_t(\Sigma^*) < 0$ for an arbitrary contingent agent t (with $\theta_t = 1.5$). Recall that $v_t(1,\mathbf{R}) - v_t(1,\mathbf{A}) > 0$ and $v_t(2,\mathbf{A}) - v_t(2,\mathbf{R}) > 0$. We consider three cases:

- If $\lambda_2(\Sigma') \le \lambda_2(\Sigma^*)$ and $(1 - \lambda_1(\Sigma')) \le (1 - \lambda_1(\Sigma^*))$, then one of these two inequalities must be strict by (12). Equation (13) then implies $u_t(\Sigma') - u_t(\Sigma^*) < 0$.
- If $\lambda_2(\Sigma') > \lambda_2(\Sigma^*)$, then we have $P_1(1-\lambda_1(\Sigma^*)) - P_1(1-\lambda_1(\Sigma')) > 2B\exp(-2C^2\alpha_{1.5}T)$ by (12). Since $\lambda_2(\Sigma') \le 1$ and $\lambda_2(\Sigma^*) \ge 1 - 2\exp(-2C^2\alpha_{1.5}T)$, we have $\lambda_2(\Sigma') - \lambda_2(\Sigma^*) \le 2\exp(-2C^2\alpha_{1.5}T)$. We also have $v_t(2,\mathbf{A}) - v_t(2,\mathbf{R}) \le B$ and $v_t(1,\mathbf{R}) - v_t(1,\mathbf{A}) \ge 1$ (recall that $v_t(1,\mathbf{A})$, $v_t(1,\mathbf{R})$, $v_t(2,\mathbf{R})$ and $v_t(2,\mathbf{A})$ are integers bounded by B). Putting those into (13), we have
$$u_t(\Sigma') - u_t(\Sigma^*) < -1 \cdot 2B\exp(-2C^2\alpha_{1.5}T) + P_2 \cdot B \cdot 2\exp(-2C^2\alpha_{1.5}T) < 0.$$
- If $1 - \lambda_1(\Sigma') > 1 - \lambda_1(\Sigma^*)$, then we have $P_2\lambda_2(\Sigma^*) - P_2\lambda_2(\Sigma') > 2B\exp(-2C^2\alpha_{1.5}T)$ by (12). Similar to the second case, we have $(1 - \lambda_1(\Sigma')) - (1 - \lambda_1(\Sigma^*)) \le 2\exp(-2C^2\alpha_{1.5}T)$, $v_t(2,\mathbf{A}) - v_t(2,\mathbf{R}) \ge 1$ and $v_t(1,\mathbf{R}) - v_t(1,\mathbf{A}) \le B$. Putting those into (13), we have
$$u_t(\Sigma') - u_t(\Sigma^*) < P_1 \cdot B \cdot 2\exp(-2C^2\alpha_{1.5}T) + 1 \cdot (-2B\exp(-2C^2\alpha_{1.5}T)) < 0.$$

Putting these three cases together, we have $u_t(\Sigma') - u_t(\Sigma^*) < 0$ for an arbitrary contingent agent t. □

Therefore, D cannot contain any contingent agents by 2 of Definition 2.1. Corollary 3.5 says that D cannot simultaneously contain an candidate-friendly agent and a candidate-unfriendly agent. Thus any D must contain either only candidate-friendly agents or only candidate-unfriendly agents.

The following claim states that neither type of predetermined agents alone are powerful enough to change the outcome to their favor. This concludes the proof as we have shown there is no deviating coalition.

Claim 3.8. *Suppose $\alpha_{0.5} < 0.5$ and $\alpha_{2.5} < 0.5$. If D contains only candidate-friendly agents, then $\lambda_1(\Sigma') \leq \lambda_1(\Sigma^*)$ and $\lambda_2(\Sigma') \leq \lambda_2(\Sigma^*)$. If D contains only candidate-unfriendly agents, then $\lambda_1(\Sigma') \geq \lambda_1(\Sigma^*)$ and $\lambda_2(\Sigma') \geq \lambda_2(\Sigma^*)$.*

Proof. We focus on the case that D contains only candidate-friendly agents. The candidate unfriendly case is analogous.

No matter how these candidates deviate, fewer than $1/2$ the agents report $\alpha_{0.5}$. If more than $1/2$ the agents report $\alpha_{2.5}$, then the candidate will surely be rejected. If some deviate to $\alpha_{1.5}$ it also cannot help. Their reported signals $\bar{\theta}_t$ are already being treated as 2 and their threshold $\bar{\delta}_t$ as 0. Any change can only make the candidate less likely to be accepted. □

4 Unknown/Partially Known Distribution of Agent Types

Before generalizing our result to the setting with general M and N, we take a detour and discuss the necessity of the assumption that the distribution of agent types is common knowledge. As we mentioned in Sect. 2, we assume the distribution of agent types, $\alpha_{0.5}, \alpha_{1.5}, \ldots, \alpha_{N+0.5}$, is a common knowledge among the agents. This assumption is natural by its own in many scenarios including our running example (if a theory candidate is applying at a computer science department, those theory faculty members are more inclined to accept the candidate than the faculty members in AI, software, hardware; moreover, the numbers of the theory, AI, software, hardware faculty members are public information). In this section, we will see that this assumption is also necessary for the existence of a mechanism that satisfies Theorem 3.2 and Theorem 3.3.

Before describing our impossibility result, we first formally define the model with unknown agent types. Let $\Delta_{N+1} = \{(x_1, \ldots, x_{N+1}) \mid \forall i : x_i \in [0,1]; \sum_{i=1}^{N+1} x_i = 1\}$. The distribution of types, $(\alpha_{0.5}, \alpha_{1.5}, \ldots, \alpha_{N+0.5})$, is then an element of Δ_{N+1}. To model an unknown/partially known distribution of agent types, let $\mathcal{D}_{\Delta_{N+1}}$ be a distribution over Δ_{N+1} where each agent believes the distribution of the agent types, $(\alpha_{0.5}, \alpha_{1.5}, \ldots, \alpha_{N+0.5})$, is drawn from $\mathcal{D}_{\Delta_{N+1}}$.

Next, we describe a natural property that is shared by most social choice mechanism, including the one in this paper.

Definition 4.1. A mechanism is *anonymous* if it always outputs the same alternative for any two collections of reports $\mathbf{r}^{(1)} = (r_1^{(1)}, \ldots, r_T^{(1)}) \in \mathcal{R}^T, \mathbf{r}^{(2)} = (r_1^{(2)}, \ldots, r_T^{(2)}) \in \mathcal{R}^T$ such that $\mathbf{r}^{(1)}$ is a permutation of $\mathbf{r}^{(2)}$.

In other words, an anonymous mechanism cannot decide the output alternative based on agents' identities.

We have the following strong impossibility result, whose proof is available in Appendix B.

Theorem 4.2. *Under the setting with unknown distribution of agent types, there exists a constant $\tau > 0$ such that no anonymous mechanism always outputs the alternative favored by more than half of the agents with probability more than $1 - \tau$ in any τ-strong symmetric Bayes Nash equilibrium. This is true even with $M = N = 2$.*

Since the truthful strategy profile is symmetric, we have the following corollary about the impossibility of a truthful mechanism.

Corollary 4.3. *Under the setting with unknown distribution of agent types defined above, there exists a constant $\tau > 0$ such that no anonymous mechanism satisfies both of the followings:*

- *the mechanism outputs the alternative favored by more than half of the agents with probability more than $1 - \tau$;*
- *under the mechanism, the truthful strategy profile is a τ-strong Bayes Nash equilibrium.*

This is true even with $M = N = 2$.

5 Non-binary Worlds

In this section, we consider the generalization to the setting with more than two worlds $N > 2$, while keeping the binary signal assumption $M = 2$. We will see in the next section that the generalization to non-binary signals is simple.

Throughout this section, we define $\vartheta_\top \in \{0.5, 1.5, \ldots, N + 0.5\}$ to be the *median threshold* satisfying $\sum_{\vartheta=1}^{\vartheta_\top} \alpha_\vartheta > 0.5$ and $\sum_{\vartheta=\vartheta_\top}^{N+0.5} \alpha_\vartheta > 0.5$. It is easy to see that $\vartheta_\top$ is the median over all the agents' thresholds.

In the case $M = N = 2$, we have asked each agent his/her received signal, threshold, and posterior belief on the fraction of agents who will report signal 1. In the case $N > 2$ here, while it is still natural to ask an agent for his/her signal, asking for posterior prediction and keeping the mechanism as before will not work here. In particular, this will make Theorem 3.2 fail. To reason this intuitively, it is easy to see that, if the mechanism is required to output the alternative favored by the majority, the mechanism will output **A** if $W > \vartheta_\top$ and output **R** if $W < \vartheta_\top$. To ensure this, we need to make sure the median of the posterior prediction is between $P_{\vartheta_\top+0.5,1}$ and $P_{\vartheta_\top-0.5,1}$ (recall that P_{n1} is the probability that signal 1 is received if the actual world is n). However, while this is true for $N = 2$ as Theorem 3.1 suggests (in fact, all the possible posterior predictions, T_{21} and T_{22}, are between P_{21} and P_{22}), this is not necessarily true for $N > 2$.

As a solution to this issue, for each agent t with threshold θ_t, we ask him/her for a value between $P_{\theta_t-0.5,1}$ and $P_{\theta_t+0.5,1}$ (agents with threshold 0.5 report a value between P_{11} and 1 and agents with threshold $N+0.5$ report a value between 0 and P_{N1}), and the median of these values will be between $P_{\vartheta_\top+0.5,1}$ and $P_{\vartheta_\top-0.5,1}$. A natural way to ask an agent for this value can be, *please give a percentage value q such that you would like alternative **A** if the fraction of agents reporting signal* 1 *is less than q percent, and you would like alternative **R** if otherwise.*

A challenge in the non-binary world setting is how to ask agents to report thresholds. The thresholds $0.5, 1.5, \ldots, N + 0.5$, as well as the world index $1, \ldots, N$, may be artificial in some practical settings. To be specific, while, for the case $N = 2$, the thresholds $0.5, 1.5, 2.5$ can represent "candidate-friendly", "contingent" and "candidate-unfriendly" in our running example, or "republican-aligned", "swing" and "democratic-aligned" in the US presidential election case, it is unclear what $0.5, 1.5, \ldots, N + 0.5$ can represent in the $N > 2$ case. This will make designing questionnaire difficult. To overcome this, we will not ask each agent's threshold θ_t, and we will ask each agent to choose one from the three options: I) $\theta_t > \vartheta_\top$, II) $\theta_t = \vartheta_\top$ or III) $\theta_t < \vartheta_\top$.

Our mechanism is presented in Mechanism 2.

Mechanism 2 Mechanism for $N > 2$ and $M = 2$

1: Each agent t reports to the mechanism the signal (s)he receives (either 1 or 2), one of the three options I) $\theta_t > \vartheta_\top$, II) $\theta_t = \vartheta_\top$ or III) $\theta_t < \vartheta_\top$, and a value $q_t \in [0, 1]$ such that (s)he would like **A** *if and only if* the fraction of agents who report 1 is less than q_t. As a result, the report space is $\mathcal{R} = \{1, 2\} \times \{\text{I}, \text{II}, \text{III}\} \times [0, 1]$. Let $(\bar{s}_t, \bar{o}_t, \bar{q}_t)$ be t's report.

2: If agent t reports $\bar{o}_t = \text{I}$, his reported signal will be automatically treated as $\bar{s}_t = 1$; if agent t reports $\bar{o}_t = \text{III}$, his reported signal will be automatically treated as $\bar{s}_t = 2$. *The value $\bar{q}_t$ in the previous step should be given with this treatment being considered, and the mechanism makes this clear to the agents.*

3: Compute the *median* of those reporting $\bar{q}_t$, denoted by $\bar{q}$.

4: If the fraction of the agents reporting $\bar{s}_t = 1$ is more than the median $\bar{q}$, announce **R** being the winning alternative; otherwise, announce **A** being the winning alternative.

In our running example, the questionnaire corresponding to the mechanism looks like the following:

An example for the questionnaire:

1. What is your impression of this candidate during the individual interview between you and this candidate?

   (a) I had a good impression.

   (b) I do not have a good impression.

2. Choose one of the following.

   I I may still want **R** even if more than half of the faculty members prefer **A**;

   II I am happy with **A** if more than half of the faculty members prefer **A**, and I am happy with **R** if more than half of the faculty members prefer **R**;

   III I may still want **A** even if more than half of the faculty members prefer **R**.

   If your choice is I, your answer in Q1 will be treated as (b); if your choice is III, your answer in Q1 will be treated as (a).

3. Please give a percentage value q such that you would like alternative **A** if the fraction of faculty members reporting signal 1 is less than q percent, and you would like alternative **R** if otherwise. Please answer this question assuming the treatment in the previous question.

Next, we will prove that the mechanism outputs the alternative favored by the majority with high probability, and the truthful strategy profile is an ε-Strong Bayes Nash Equilibrium for $\varepsilon = o(1)$. For simplicity and concisely describing the ideas behind the proofs, we will not perform the Chernoff bound analyses as in Sect. 3, and we will assume $T \to \infty$ instead. As a result, for any world n, the fractions of agents receiving signal 1 and 2 are, almost surely, P_{n1} and P_{n2} respectively.

Theorem 5.1. *Suppose $T \to \infty$. If all the agents play the truthful strategy, then our mechanism outputs an alternative favored by more than half of the agents.*

Proof. Suppose all the agents report truthfully. We aim to show that, the mechanism will output **A** if and only if $W > \vartheta_\top$. Let n be the actual world. We assume $n > \vartheta_\top$ without loss of generality, as the analysis for $n < \vartheta_\top$ is similar. Let $\alpha_{\mathrm{I}} = \sum_{\vartheta > \vartheta_\top} \alpha_\vartheta$ be the fraction of agents who report Option I, and let $\alpha_{\mathrm{III}} = \sum_{\vartheta < \vartheta_\top} \alpha_\vartheta$ be the fraction of agents who report Option III. The fraction of agents reporting signal 1 is $\alpha_{\mathrm{I}} + \alpha_{\vartheta_\top} \cdot P_{n1}$.

On the other hand, for an agent t with threshold θ_t, (s)he prefers **A** if $W > \theta_t$, and (s)he prefers **R** if otherwise. If (s)he was asked to give a value such that (s)he would like **A** if and only if the fraction of agents who *receive* signal 1 is less than this value, she would have report a value between $P_{\theta_t+0.5,1}$ and $P_{\theta_t-0.5,1}$. Now, considering that, as instructed by the mechanism, those $\alpha_{\mathrm{I}} \cdot T$ (resp. $\alpha_{\mathrm{III}} \cdot T$) agents will always *report* signal 1 (resp. signal 2) regardless of what they receive, (s)he will report $\bar{q}_t \in (\alpha_{\mathrm{I}} + \alpha_{\vartheta_\top} \cdot P_{\theta_t+0.5,1}, \alpha_{\mathrm{I}} + \alpha_{\vartheta_\top} \cdot P_{\theta_t-0.5,1})$.

It is then easy to see that the median $\bar{q}$ is between $\alpha_{\mathrm{I}} + \alpha_{\vartheta_\top} \cdot P_{\vartheta_\top+0.5,1}$ and $\alpha_{\mathrm{I}} + \alpha_{\vartheta_\top} \cdot P_{\vartheta_\top-0.5,1}$, which is more than the fraction of agents reporting signal 1 (which is $\alpha_{\mathrm{I}} + \alpha_{\vartheta_\top} \cdot P_{n1}$ as computed earlier). The last step of our mechanism will make sure **A** is output. □

As a remark, if we do not assume $T \to \infty$, to show that the statement in Theorem 5.1 fails with an exponentially low probability, we need to make an extra assumption that the median $\bar{q}$ is not exponentially close to the two endpoints $\alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\vartheta_\top+0.5,1}$ and $\alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\vartheta_\top-0.5,1}$. This is a natural assumption, as an agent's reported value should not depend on T. In addition, in practice, it is natural to expect that most agents with threshold θ_t will report values that are around the midpoint of the interval $(\alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\theta_t+0.5,1}, \alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\theta_t-0.5,1})$.

Next, we show that our mechanism satisfies the truthful property. Again, we consider $T \to \infty$.

Theorem 5.2. *Suppose $T \to \infty$. The truthful strategy profile form a strong Bayes Nash Equilibrium.*

Similar as in Sect. 3, given a strategy profile Σ and a value $n \in \{1, \ldots, N\}$, let $\lambda_n(\Sigma)$ be the probability that alternative **A** is announced if the actual world is $W = n$.

The ideas behind the proof of Theorem 5.2 is similar as before. Firstly, those agents with threshold $\vartheta_\top$ will not deviate from the truthful strategy, as their utilities have already been maximized. Secondly, we can prove a lemma similar to Lemma 3.4 showing that there is a conflict of interest between an arbitrary agent with threshold below $\vartheta_\top$ and an arbitrary agent with threshold above $\vartheta_\top$. This shows that the set of deviating agents D can only contain either agents with thresholds below $\vartheta_\top$ or agents with thresholds above $\vartheta_\top$. This further implies that more than half of the agents are truth-telling. Finally, the use of median in our mechanism ensures that less than half of the agents' deviating cannot change the output alternative.

Lemma 5.3. *Let Σ^* be the truthful strategy profile and Σ' be an arbitrary strategy profile. Let t_1 be an arbitrary agent with threshold below $\vartheta_\top$ and t_2 be an arbitrary agent with threshold above $\vartheta_\top$. Suppose $T \to \infty$. We have*

1. *If $u_{t_1}(\Sigma') > u_{t_1}(\Sigma^*)$, then $u_{t_2}(\Sigma') < u_{t_2}(\Sigma^*)$.*
2. *If $u_{t_1}(\Sigma') < u_{t_1}(\Sigma^*)$, then $u_{t_2}(\Sigma') > u_{t_2}(\Sigma^*)$*

Proof. We will only prove (1), as the proof for (2) is similar. For the ease of notation, in this proof, we write $u_1, u_2, v_1, v_2, \theta_1, \theta_2$ to represent $u_{t_1}, u_{t_2}, v_{t_1}, v_{t_2}, \theta_{t_1}, \theta_{t_2}$ respectively. Suppose $u_1(\Sigma') > u_1(\Sigma^*)$, and we aim to show $u_2(\Sigma') < u_2(\Sigma^*)$. Theorem 5.1 implies that $\lambda_n(\Sigma^*) = 0$ for all $n < \vartheta_\top$ and $\lambda_n(\Sigma^*) = 1$ for all $n > \vartheta_\top$. Firstly, we show that

$$\sum_{n:\theta_1<n<\vartheta_\top} P_n\left(\lambda_n(\Sigma') - \lambda_n(\Sigma^*)\right) > \sum_{n:\vartheta_\top<n<\theta_2} P_n\left(\lambda_n(\Sigma^*) - \lambda_n(\Sigma')\right). \tag{14}$$

This is because

$$\begin{aligned}
0 &< u_1(\Sigma') - u_1(\Sigma^*) && \text{(by our assumption)}\\
&= \sum_{n=1}^{N} P_n\left(\lambda_n(\Sigma') - \lambda_n(\Sigma^*)\right)\left(v_1(n,\mathbf{A}) - v_1(n,\mathbf{R})\right) && \text{(by (4))}\\
&\le \sum_{n:\theta_1<n<\theta_2} P_n\left(\lambda_n(\Sigma') - \lambda_n(\Sigma^*)\right)\left(v_1(n,\mathbf{A}) - v_1(n,\mathbf{R})\right) && (\dagger)\\
&= \sum_{n:\theta_1<n<\vartheta_\top} P_n\left(\lambda_n(\Sigma') - \lambda_n(\Sigma^*)\right)\left(v_1(n,\mathbf{A}) - v_1(n,\mathbf{R})\right)\\
&\quad - \sum_{n:\vartheta_\top<n<\theta_2} P_n\left(\lambda_n(\Sigma^*) - \lambda_n(\Sigma')\right)\left(v_1(n,\mathbf{A}) - v_1(n,\mathbf{R})\right)\\
&\le \sum_{n:\theta_1<n<\vartheta_\top} P_n\left(\lambda_n(\Sigma') - \lambda_n(\Sigma^*)\right)\left(v_1(\vartheta_\top - 0.5,\mathbf{A}) - v_1(\vartheta_\top - 0.5,\mathbf{R})\right)\\
&\quad - \sum_{n:\vartheta_\top<n<\theta_2} P_n\left(\lambda_n(\Sigma^*) - \lambda_n(\Sigma')\right)\left(v_1(\vartheta_\top - 0.5,\mathbf{A}) - v_1(\vartheta_\top - 0.5,\mathbf{R})\right), && (\ddagger)
\end{aligned}$$

which implies (14), where both Step (†) and (‡) are based on the following facts. In particular, (†) is based on the first two facts, and (‡) is based on the first and the third facts.

- for $n < \vartheta_\top$, $\lambda_n(\Sigma') - \lambda_n(\Sigma^*) = \lambda_n(\Sigma') - 0 \ge 0$; for $n > \vartheta_\top$, $\lambda_n(\Sigma^*) - \lambda_n(\Sigma') = 1 - \lambda_n(\Sigma') \ge 0$.
- $v_1(n,\mathbf{A}) - v_1(n,\mathbf{R})$ is negative for $n < \theta_1$ and is positive for $n > \theta_2$. Notice that this is also true for v_2.
- the expression $v_1(n,\mathbf{A}) - v_1(n,\mathbf{R})$ is increasing in n. This is true for any agent t. In particular, for each agent t, $v_t(n,\mathbf{A})$ is increasing in n and $v_t(n,\mathbf{R})$ is decreasing in n.

Next, we show that (14) implies $u_2(\Sigma') < u_2(\Sigma^*)$. By the same calculations and analyses above, we have

$$\begin{aligned}
&u_2(\Sigma') - u_2(\Sigma^*)\\
\le &\sum_{n:\theta_1<n<\vartheta_\top} P_n\left(\lambda_n(\Sigma') - \lambda_n(\Sigma^*)\right)\left(v_2(\vartheta_\top - 0.5,\mathbf{A}) - v_2(\vartheta_\top - 0.5,\mathbf{R})\right) - \sum_{n:\vartheta_\top<n<\theta_2} P_n\left(\lambda_n(\Sigma^*) - \lambda_n(\Sigma')\right)\\
&\left(v_2(\vartheta_\top - 0.5,\mathbf{A}) - v_2(\vartheta_\top - 0.5,\mathbf{R})\right) && \text{(same calculations and analyses above)}\\
<&0, && \text{(by } v_2(\vartheta_\top - 0.5,\mathbf{A}) - v_2(\vartheta_\top - 0.5,\mathbf{R}) < 0 \text{ and (14))}
\end{aligned}$$

which implies the lemma. □

Now we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. Suppose otherwise and there is a set of deviating agents D. Let Σ' be the profile after the deviation of agents in D. Firstly, we show that D cannot contain an agent t with threshold exactly $\vartheta_\top$. Notice that such an agent's utility has already been maximized by the truthful profile Σ^*. Suppose $\lambda_n(\Sigma') \neq \lambda_n(\Sigma^*)$ for certain n. It must be that $\lambda_n(\Sigma') > \lambda_n(\Sigma^*) = 0$ if $n < \vartheta_\top$, and $\lambda_n(\Sigma') < \lambda_n(\Sigma^*) = 1$ if $n > \vartheta_\top$. It is then easy to see that this agent's utility will decrease, which contradicts to 2 of Definition 2.1. Suppose $\lambda_n(\Sigma') = \lambda_n(\Sigma^*)$ for all n. We have $u_t(\Sigma') = u_t(\Sigma^*)$ for every agent t. This already contradicts to 3 of Definition 2.1.

Next, Lemma 5.3 ensures that D cannot contain both an agent with threshold less than $\vartheta_\top$ and an agent with threshold greater than $\vartheta_\top$. Assume without loss of generality that D only contains agents with threshold more than $\vartheta_\top$. In this case, all the agents with threshold less than $\vartheta_\top$ and agents with threshold exactly $\vartheta_\top$ are truth-telling, and notice that there are more than $\frac{1}{2}T$ such agents.

To derive a contradiction, we show that $u_t(\Sigma') \leq u_t(\Sigma^*)$ for each $t \in D$ (this will contradict to 3 of Definition 2.1). Consider an arbitrary agent $\hat{t} \in D$, and notice that $\theta_{\hat{t}} > \vartheta_\top$ by our assumption. For each world $n < \vartheta_\top$, agent $\hat{t}$'s utility is maximized if $\lambda_n(\Sigma') = 0$, and $\hat{t}$'s utility is already maximized by the truthful profile: $\lambda_n(\Sigma^*) = 0$. Therefore, to prove $u_{\hat{t}}(\Sigma') \leq u_{\hat{t}}(\Sigma^*)$, it suffices to show that $\lambda_n(\Sigma') = \lambda_n(\Sigma^*) = 1$ for each $n > \vartheta_\top$ (so that $\hat{t}$'s utility is unchanged for those worlds greater than $\vartheta_\top$, and $\hat{t}$'s utility has already been maximized by the truthful profile making no further utility gain possible for those worlds smaller than $\vartheta_\top$).

Consider an arbitrary world $n > \vartheta_\top$. Let $\alpha_{\mathrm{I}} = \sum_{\vartheta > \vartheta_\top} \alpha_\vartheta$ and $\alpha_{\mathrm{III}} = \sum_{\vartheta < \vartheta_\top} \alpha_\vartheta$. If n is the actual world, P_{n1} fraction of the agents will receive signal 1. The fraction of agents reporting signal 1, after the treatment in Step 2 of the mechanism, is $\alpha_{\mathrm{I}} + \alpha_\top \cdot P_{n1}$ under the truthful strategy profile Σ^*. For Σ', the fraction of agents reporting signal 1 is at most $\alpha_{\mathrm{I}} + \alpha_\top \cdot P_{n1}$, as agents with thresholds $\vartheta_\top$ or below are truth-telling. On the other hand, by a similar analysis in the proof of Theorem 5.1, a truthful agent t will report the value $\bar{q}_t$ that is in the interval $(\alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\theta_t + 0.5,1}, \alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\theta_t - 0.5,1})$. Since agents with thresholds less than or equal to $\vartheta_\top$ are truthful and there are more than $\frac{1}{2}T$ of them, the median of those reported $\bar{q}_t$ is greater than $\alpha_{\mathrm{I}} + \alpha_{\theta_\top} \cdot P_{\vartheta_\top + 0.5,1}$, regardless of the reports of the agents with thresholds more than $\vartheta_\top$. Since $n > \vartheta_\top$ which implies $n \geq \vartheta_\top + 0.5$, we have $P_{n,1} \leq P_{\vartheta_\top + 0.5,1}$, which further implies that the fraction of agents who report 1 is less than the median of those reported $\bar{q}_t$. By our mechanism, $\mathbf{A}$ will always be announced, and $\lambda_n(\Sigma') = 1$. □

6 Non-binary Signals

There is a simple reduction from the non-binary signal setting to the binary-signal setting. Suppose the signal space is $\{1, \ldots, M\}$. To reduce it to a binary signal space $\{1, 2\}$, we set an arbitrary non-integer number $s_\top$ between 1 and M. All the signals less than $s_\top$ are reduced to the "bad" signal 1, and all the signals greater than $s_\top$ are reduced to the "good" signal 2. The mechanisms in the previous sections can be adapted to the setting here. The mechanisms are the same as before, except for the following change: whenever the mechanism asks an agent for a binary signal in the previous setting, the mechanism asks the agent whether the signal (s)he received is less than or more than $s_\top$, which corresponds to signal 1 and 2 respectively.

For the mechanism in Sect. 5, all the properties, including that the mechanism outputs the alternative favored by more than half of the agents and that the truth-telling strategy profile form a strong Bayes Nash Equilibrium, continue to hold in the non-binary signal setting with exactly the same proofs. For the mechanism in Sect. 3, it is easy to see that these properties also continue to hold here if we prove the following inequality that is similar to the one in Theorem 3.1:

$$\forall m \in \{1, \ldots, M\} : P_{21} < T_{m1} < P_{11}, \tag{15}$$

where P_{11} and P_{21} are the probabilities that a signal below $s_\top$ is received if the actual world is 1 and 2 respectively (notice that we are dealing with binary-world setting), and T_{m1} is the probability that an agent who receives signal m believes that another agent will receive a signal that is less than $s_\top$. Intuitively, if

(15) holds, all the agents' posterior predictions are still between P_{21} and P_{11}, and the "correct" signal will still be "surprisingly popular". The proof of (15) is by straightforward Bayesian analysis, and is left to the readers.

7 Conclusion

We presented a mechanism that elicits and aggregates the information and preferences of voters over two alternatives. In particular, voters' truthfully reporting their signals forms a strong Bayes Nash equilibrium, and in this case the mechanism outputs the alternative that is favored by the majority with overwhelming probability.

We mention two possible future directions. The first is deployment. It would be interesting to test this mechanism in the real world, and then test to see if participants are, in aggregate, more happy when this mechanism is used as compared with a majority vote mechanism. For example, groups could choose a movie to watch where different participants have different information about the potential movies. In particular, not everyone has seen both movies. Participants could be surveyed afterwards about how enjoyable the movie was.

The second direction is to understand under what more general models the mechanism works. For example, what if the agents are not true Bayesians? When should we still expect this mechanism to work if the agents are performing heuristic calculations? Additionally, are there modifications that should be made to the mechanism to make it increasingly robust for use with non-Bayesian agents?

References

[1] David Austen-Smith and Jeffrey S. Banks. Information aggregation, rationality, and the condorcet jury theorem. *American political science review*, 90(1):34–45, 1996.

[2] Duncan Black. On the Rationale of Group Decision-making. *Journal of political economy*, 56(1), 1948.

[3] Kay-Yut Chen, Leslie R Fine, and Bernardo A Huberman. Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50(7):983–994, 2004.

[4] Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 9–26, 2018.

[5] Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330. International World Wide Web Conferences Steering Committee, 2013.

[6] N. De Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix.* Cambridge University Press, 2014.

[7] Ofer Dekel, Felix Fischer, and Ariel D Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

[8] Boi Faltings and Goran Radanovic. Game theory for data science: Eliciting truthful information. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 11(2):1–151, 2017.

[9] Timothy Feddersen and Wolfgang Pesendorfer. Voting behavior and information aggregation in elections with private information. *Econometrica*, 65(5):1029–1058, 1997. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/2171878`.

[10] A. Gao, J. R. Wright, and K. Leyton-Brown. Incentivizing Evaluation via Limited Access to Ground Truth: Peer-Prediction Makes Things Worse. *ArXiv e-prints*, June 2016.

[11] Ralph Hertwig. Tapping into the wisdom of the crowd—with confidence. *Science*, 336(6079):303–304, 2012.

[12] Hadi Hosseini, Debmalya Mandal, Nisarg Shah, and Kevin Shi. Surprisingly popular voting recovers rankings, surprisingly! In *Proceedings of the International Joint Conference on Artificial Intelligence (forthcoming)*, 2021.

[13] Harold Hotelling. Stability in Competition. *The Economic Journal*, 39(153), 1929.

[14] Leslie K John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

[15] Yuqing Kong. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2020.

[16] Yuqing Kong and Grant Schoenebeck. Equilibrium selection in information elicitation without verification via information monotonicity. In *Proceedings of the 9th Innovations in Theoretical Computer Science (ITCS 2018)*, January 2018.

[17] Yuqing Kong and Grant Schoenebeck. Eliciting expertise without verification. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pages 195–212, 2018.

[18] Yuqing Kong and Grant Schoenebeck. A Framework For Designing Information Elicitation Mechanisms That Reward Truth-telling. *ACM Transactions on Economics and Computation (TEAC)*, 7(1), 2019.

[19] Yuqing Kong, Grant Schoenebeck, Biaoshuai Tao, and Fang-Yi Yu. Information elicitation mechanisms for statistical estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(2), pages 2095–2102, 2020.

[20] Krishna K. Ladha. The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science*, 36(3), 1992.

[21] N. Miller. Information, Electorates, and Democracy: some Extensions and Interpretations of the Condorcet Jury Theorem. *Information Pooling and Group Decision Making*, 1986.

[22] N. Miller, P. Resnick, and R. Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, pages 1359–1373, 2005.

[23] Javier Perote and Juan Perote-Pena. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47(2):153–176, 2004.

[24] D. Prelec. A Bayesian Truth Serum for subjective data. *Science*, 306(5695):462–466, 2004.

[25] Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

[26] Grant Schoenebeck and Fang-Yi Yu. Two strongly truthful mechanisms for three heterogeneous agents answering one question. In *Web and Internet Economics*, pages 119–132, Cham, 2020. Springer International Publishing. ISBN 978-3-030-64946-3.

[27] Grant Schoenebeck and Fang-Yi Yu. Learning and strongly truthful multi-task setting peer prediction: A variational approach. In *Proceedings of the 12th Innovations in Theoretical Computer Science (ITCS 2021)*, January 2021.

[28] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196, 2016.

[29] Joseph P Simmons, Leif D Nelson, Jeff Galak, and Shane Frederick. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15, 2011.

[30] Peyton Young. Condorcet's Theory of Voting. *American Political Science*, 82(4), 1988.

[31] Peter Zhang and Yiling Chen. Elicitability and knowledge-free elicitation with peer prediction. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 245–252, 2014.

A Comparison with Feddersen and Pesendorfer's Work

Feddersen and Pesendorfer [9] consider a two-alternative setting similar to our model. As mentioned before, Feddersen and Pesendorfer [9] consider the standard majority voting where each agent votes for an alternative, while assuming agents play a (Bayes) Nash equilibrium strategy profile. We, on the other hand, design a more sophisticated mechanism to incentive truth-telling.

Other than this difference, the state space, the signal space and the space of agents' types in Feddersen and Pesendorfer [9] are all continuous. For Feddersen and Pesendorfer's continuous setting, in the Nash equilibrium, agents' strategies have three types: always vote for one alternative, always vote for the other, and vote the alternative based on the signal. These are similar to our three types: candidate-friendly, contingent, candidate-unfriendly. However, due to continuity, each agent needs to compute his/her type by solving an equation with a complicated Riemann integral (while agents' know their types directly from their thresholds in our setting). A phenomenon in their setting due to continuity is that the fraction of contingent voters in the Nash equilibrium approaches to zero when the number of the voters goes to infinity.

Although agents can be classified by three types in both settings, we would like to clarify a fundamental difference in the motivation behind this classification. In our setting, each agent's type is determined by his/her threshold and reflects his/her preference over the two alternatives. In Feddersen and Pesendorfer's setting, each agent "chooses" a type in a specific way so that the majority voting scheme outputs the correct alternative with high probability. Therefore, in their setting, agents' types are chosen for collaboratively aggregating information, and should not be viewed as reflections of their preferences. Although an agent's preference does affect his/her choice, the purpose for choosing a type is for information aggregation, not for reflecting the preference.

At a high level, our mechanism includes some novel techniques, including the surprising popular technique and the median trick, to ensure the output of the correct alternative in the setting with strategic agents. In Feddersen and Pesendorfer's setting, it may be surprising that the simple majority voting scheme is already enough for output the correct alternative. The reason behind this is that certain implicit techniques for guaranteeing the correct output are "embedded" into agents' strategic behaviors. In other words, the agents are the ones who work out those techniques, not the mechanism. That is why we mentioned in the introduction that the agents in Feddersen and Pesendorfer's setting need to have much more sophistication compared with the agents in our setting.

Another difference is that they are considering a Nash equilibrium strategy profile, while our mechanism satisfies the much stronger criterion that truth-telling strategies form a *strong* Bayes Nash equilibrium.

B Proof of Theorem 4.2

Consider an anonymous mechanism and an arbitrary symmetric strategy profile Σ. For $\vartheta \in \{0.5, 1.5, \ldots, N+0.5\}$, let α_ϑ denote the fraction of agents having threshold ϑ. Those $\alpha_\vartheta \cdot T$ agents will report based on the signal they received, and there can only be M different reports. The mechanism can only see how many different reports there are, and how many agents reporting each of them; in particular, the mechanism cannot see who reports which. The following proposition follows immediately from the above remarks.

Proposition B.1. *Fix a symmetric strategy profile Σ. Let $\beta_{\vartheta,m}$ be the fraction of agents with threshold ϑ that receive signal m. If a mechanism is anonymous, then the values $\{\alpha_\vartheta : \vartheta = 0.5, 1.5, \ldots, N+0.5\} \cup \{\beta_{\vartheta,m} : \vartheta = 0.5, 1.5, \ldots, N+0.5; m = 1, \ldots, M\}$ completely determines the output of the mechanism.*

We also need the following technical lemma.

Lemma B.2. *The total variation distance between the two binomial distributions* $\mathrm{Bin}(T, 1/6)$ *and* $\mathrm{Bin}(T/3, 1/2)$ *is less than* 0.123 *for sufficiently large T.*

Proof. By Central Limit Theorem, the total variation distance between $\mathrm{Bin}(T, 1/6)$ and $\mathrm{Bin}(T/3, 1/2)$ is at most the total variation distance between $\mathcal{N}(T/6, 5T/36)$ and $\mathcal{N}(T/6, T/12)$ plus $o(1)$, which, by shifting the mean of the Gaussian distribution, is the total variation distance between $\mathcal{N}(0, 5T/36)$ and $\mathcal{N}(0, T/12)$ plus $o(1)$.

Let $f(x)$ and $g(x)$ be the probability density function for $\mathcal{N}(0, 5T/36)$ and $\mathcal{N}(0, T/12)$ respectively. To calculate the total variation distance, firstly, straightforward calculations reveal that $f(x) < g(x)$ on

$\left(-\sqrt{\frac{5}{24}\ln\frac{5}{3}T}, \sqrt{\frac{5}{24}\ln\frac{5}{3}T}\right)$ and $f(x) > g(x)$ on $\left(-\infty, -\sqrt{\frac{5}{24}\ln\frac{5}{3}T}\right) \cup \left(\sqrt{\frac{5}{24}\ln\frac{5}{3}T}, \infty\right)$. Therefore, the total variation distance between $\mathcal{N}(0, 5T/36)$ and $\mathcal{N}(0, T/12)$ is

$$\begin{aligned}
\int_{-\sqrt{\frac{5}{24}\ln\frac{5}{3}T}}^{\sqrt{\frac{5}{24}\ln\frac{5}{3}T}} g(x) - f(x)dx &= \int_{-\sqrt{\frac{5}{24}\ln\frac{5}{3}T}}^{\sqrt{\frac{5}{24}\ln\frac{5}{3}T}} \frac{1}{\sqrt{2\pi\frac{T}{12}}} e^{-\frac{1}{2}\frac{x^2}{T/12}} - \frac{1}{\sqrt{2\pi\frac{5T}{36}}} e^{-\frac{1}{2}\frac{x^2}{5T/36}} dx \\
&= \int_{-\sqrt{\frac{5}{24}\ln\frac{5}{3}}}^{\sqrt{\frac{5}{24}\ln\frac{5}{3}}} \frac{1}{\sqrt{2\pi\frac{1}{12}}} e^{-\frac{1}{2}\frac{y^2}{1/12}} - \frac{1}{\sqrt{2\pi\frac{5}{36}}} e^{-\frac{1}{2}\frac{y^2}{5/36}} dy \qquad \text{(where } y = x/\sqrt{T}\text{)} \\
&< 0.12295. \qquad \text{(Calculated by computer)}
\end{aligned}$$

Thus, the total variation distance between $\mathrm{Bin}(T, 1/6)$ and $\mathcal{N}(T/6, 5T/36)$ is at most $0.12295 + o(1)$, which implies the lemma. □

Now we are ready to present the proof of Theorem 4.2 (restated below).

Theorem 4.2. *Under the setting with unknown distribution of agent types, there exists a constant $\tau > 0$ such that no anonymous mechanism always outputs the alternative favored by more than half of the agents with probability more than $1 - \tau$ in any τ-strong symmetric Bayes Nash equilibrium. This is true even with $M = N = 2$.*

Proof. We consider the following instance with two worlds and two signals ($M = N = 2$).

The prior distribution of the two worlds (world 1 and world 2) is given by $P_1 = 0.98$ and $P_2 = 0.02$. The probability distribution of the two signals under each of the two worlds is given by $P_{11} = 1$, $P_{12} = 0$, $P_{21} = 5/6$, $P_{22} = 1/6$. For each agent t with threshold θ_t, we have $v_t(n, \mathbf{A}) = 1$ and $v_t(n, \mathbf{R}) = 0$ if $n > \theta_t$, and we have $v_t(n, \mathbf{A}) = 0$ and $v_t(n, \mathbf{R}) = 1$ if $n < \theta_t$. Lastly, $\mathcal{D}_{\Delta_3}$ is defined as follows: with probability $1/2$ we are in setting F (for friendly) and the fractions of agents having thresholds $0.5, 1.5, 2.5$ are $\alpha^{(1)}_{0.5} = 1/3, \alpha^{(1)}_{1.5} = 2/3, \alpha^{(1)}_{2.5} = 0$ respectively; with probability $1/2$ we are in setting N (for no friendly), the fractions of agents having threshold $0.5, 1.5, 2.5$ are $\alpha^{(2)}_{0.5} = 0, \alpha^{(2)}_{1.5} = 1, \alpha^{(2)}_{2.5} = 0$ respectively. This finishes the description of the instance. Note there 2 worlds and 2 settings yielding 4 possible environments which we label $1F$, $1N$, $2F$, and $2N$.

Let $\tau = 0.001$. Suppose there exists a mechanism $\mathcal{M}$ that outputs the correct answer with probability at least $1 - \tau$ in a symmetric strategy profile Σ. We will show that Σ cannot be a τ-strong Bayes Nash equilibrium.

Notice that the contingent agents are the majority in all the four settings. In both environments $2F$ and $2N$, which of each happens with probability 1%, the preference of the majority is always to accept. Thus, the mechanism must accept with probability at least 99% in each environment for otherwise it will be far from being achieving $1 - \tau$ accuracy. Similarly, in both environments $1N$ and $1F$, which of each occurs with probability 49%, the mechanism must accept with probability at most 1% for otherwise it will be far from being achieving $1 - \tau$ accuracy.

To show that Σ cannot be a τ-strong Bayes Nash equilibrium, consider the set of deviating agents are all the candidate-friendly agents. Those candidate-friendly agents pretend they are contingent agents such that signal 1 is received with probability $1/2$ and signal 2 with probability $1/2$. These candidate-friendly agents will follow the strategy of the real contingent agents according to Σ. Let Σ' be the resultant strategy profile. We aim to show that those deviating candidate-friendly agents can increase their utilities significantly in Σ'.

For an intuitively arguments, in $1F$, the mechanism sees that all the agents are contingent, and the fraction of agents receiving signal 2 follows distribution $\mathrm{Bin}(T/3, 1/2)$; in $2N$, the mechanism also sees that all the agents are contingent, and the fraction of agents receiving signal 2 follows distribution $\mathrm{Bin}(T, 1/6)$. Lemma B.2 implies that the mechanism cannot distinguish between environments $1F$ and $2N$ with probability more than 87.7%. Before deviating, the mechanism will output $\mathbf{A}$ with probability at most 1% in $1F$; after deviating, the mechanism will output $\mathbf{A}$ with probability at least $99\% \cdot 87.7\% > 1\%$ in $1F$ by confusing $1F$ with $2N$. The candidate-friendly agents will benefit from deviating.

To make the arguments in the previous paragraph more rigorous, let $\mathcal{M}(X) = 1$ if the mechanism outputs $\mathbf{A}$ when all the agents are contingent, agents play according to Σ, and the number of agents receiving signal

2 is X. Let $\mathcal{M}(X) = 0$ if the output is $\mathbf{R}$ under the same circumstance. This definition is well-defined due to Proposition B.1.

In $2N$, the mechanism outputs $\mathbf{A}$ with probability $\mathbb{E}_{X\sim\mathrm{Bin}(T,1/6)}[\mathcal{M}(X)]$. In $1F$, when the candidate-friendly agents deviate to Σ', the mechanism outputs $\mathbf{A}$ with probability $\mathbb{E}_{X\sim\mathrm{Bin}(T/3,1/2)}[\mathcal{M}(X)]$. Lemma B.2 implies

$$\left|\mathop{\mathbb{E}}_{X\sim\mathrm{Bin}(T,1/6)}[\mathcal{M}(X)] - \mathop{\mathbb{E}}_{X\sim\mathrm{Bin}(T/3,1/2)}[\mathcal{M}(X)]\right| < 0.123.$$

Since we have shown that the mechanism outputs $\mathbf{A}$ with probability at least 99% in $2N$, the mechanism outputs $\mathbf{A}$ with probability at least 86.7% in $1F$ when the candidate-friendly agents deviate. Since environment $1F$ happens with probability 0.49, the expected utility for each candidate-friendly agent is at least $0.49 \cdot 86.7\% = 0.42483$. However, without deviating, $\mathbf{A}$ will be output with probability at most $0.98 \cdot 1\% + 0.02 = 0.0298$. We have seen that the candidate-friendly agents receive a utility gain of at least $0.39503 > \tau$. □

As a remark, our impossibility result Theorem 4.2 holds even for randomized mechanism. If mechanism can be randomized, Proposition B.1 becomes that $\{\alpha_\vartheta\} \cup \{\beta_{\vartheta,m}\}$ completely determines the *probability* that the mechanism output $\mathbf{A}$ (or $\mathbf{R}$). In the proof of Theorem 4.2, $\mathcal{M}(X)$ becomes the *probability* that the mechanism output $\mathbf{A}$, rather than either 0 or 1. The remaining part of the proof is exactly the same.