Outsourcing Adjudication to Strategic Jurors

Ioannis Caragiannis and **Nikolaj Schwartzbach**
Department of Computer Science, Aarhus University
iannis@cs.au.dk , nikolaj@ignatieff.io

Abstract

We study a scenario where an adjudication task (e.g., the resolution of a binary dispute) is outsourced to a set of agents who are appointed as jurors. This scenario is particularly relevant in a Web3 environment, where no verification of the adjudication outcome is possible, and the appointed agents are, in principle, indifferent to the final verdict. We consider simple adjudication mechanisms that use (1) majority voting to decide the final verdict and (2) a payment function to reward the agents with the majority vote and possibly punish the ones in the minority. Agents interact with such a mechanism strategically: they exert some effort to understand how to properly judge the dispute and cast a yes/no vote that depends on this understanding and on information they have about the rest of the votes. Eventually, they vote so that their utility (i.e., their payment from the mechanism minus the cost due to their effort) is maximized. Under reasonable assumptions about how an agent's effort is related to her understanding of the dispute, we show that appropriate payment functions can be used to recover the correct adjudication outcome with high probability. Our findings follow from a detailed analysis of the induced strategic game and make use of both theoretical arguments and simulation experiments.

1 Introduction

We consider the problem of incentivizing jurors to properly assess case evidence, so that the resulting adjudication is better than random. The problem is motivated by dispute resolution in Web3 systems, where a reliable solution would find numerous applications in, e.g., supply chain management, banking, and commerce [Schwartzbach, 2021].

Web3 typically assumes no trusted authorities and adjudication must therefore be delegated to ordinary users (or agents), who are appointed as jurors and get compensated for this activity. Such agents are anonymous and cannot easily be held accountable for their actions. They are largely indifferent to the outcome of the adjudication case and typically strategize to maximize their utility. As such, paying a fixed reward to the agents for their participation is insufficient; they can then just vote randomly, without putting in any effort to assess the case evidence, producing a useless adjudication outcome. Instead, to produce a non-trivial adjudication, payments to/from the agents should be in some way conditioned on their vote. Hopefully, if the agents are satisfied with their payments, they will make a reasonable effort to assess the case evidence and collectively come up with a correct adjudication. We ask the following natural question.

> *How can payments be structured to motivate strategic jurors to collectively produce a correct adjudication when they are indifferent to the outcome?*

We consider binary (yes/no) adjudication tasks and the following simple mechanism. Each agent submits a vote with her opinion and the adjudication outcome is decided using majority. Agents are rewarded for voting in accordance with the final verdict and less so for voting otherwise. This approach has been deployed in real systems like Kleros [Lesaege *et al.*, 2019; Lesaege *et al.*, 2021]. Kleros is already deployed on Ethereum and, at the time of writing, it has allegedly settled more than one thousand disputes.

1.1 Our Contribution

Our main conceptual contribution is a new model for the behaviour of strategic agents. The model aims to capture the two important components of strategic behaviour while participating in an adjudication task. The first one is to decide the effort the agent needs to exert to get sufficient understanding of the task and form her opinion. The second one is whether she will cast this opinion as vote or she will vote for the opposite alternative. We assume that, when dealing with an adjudication task, agents do not communicate with each other. Instead, each of them has access to the outcome of similar tasks from the past. An agent can compare these outcomes to her own reasoning for them, which allows her to conclude whether her background knowledge is positively correlated, negatively correlated, or uncorrelated to the votes cast by the other agents. Payments can be used to amplify the agent's incentive to take such correlation into account. A strategic agent then acts as follows. If there is positive correlation, her opinion for the new adjudication task will be cast as vote. If the correlation is negative, she will cast the opposite vote. If there is no correlation, the agent will vote randomly.

We assume that each adjudication task has a ground truth alternative that we wish to recover. Agents are distinguished into well-informed and misinformed ones. Well-informed (respectively, misinformed) agents are those whose opinions get closer to (respectively, further away from) the ground truth with increased effort. The ground truth is unobservable and, thus, the agents are not aware of the category to which they belong.

After presenting the strategic agent model, we characterize the strategies of the agents at equilibria of the induced game. We use this characterization to identify a sufficient condition for payments so that equilibria are simple, in the sense that the agents either vote randomly or they are all biased towards the same alternative. Next, we focus on a simple scenario with a population of well-informed and misinformed agents with complementary effort functions and show how to efficiently find payments that result in adjudication that recovers the ground truth with a given probability. Finally, we conduct experiments to justify that strategic play of a population with a majority of well-informed agents results in correct adjudication when payments are set appropriately.

1.2 Related Work

Voting, the main tool we use for adjudication, has received enormous attention in the social choice theory literature — originating with the seminal work of Arrow [1951] — and its recent computational treatment [Brandt *et al.*, 2016]. However, the main assumption there is that agents have preferences about the alternatives and thus an interest for the voting outcome, in contrast to our case where agents' interest for the final outcome depends only on whether this gives them compensation or not. Strategic voter behaviour is well-known to alter the intended outcome of all voting rules besides two-alternative majority voting and dictatorships [Gibbard, 1973; Satterthwaite, 1975]. Positive results are possible with the less popular approach of introducing payments to the voting process; e.g., see Posner and Weyl [2018].

The assumption for a ground truth alternative has been also inspired from voting theory [Caragiannis *et al.*, 2016; Conitzer and Sandholm, 2005; Young, 1988]. In a quite popular approach, votes are considered as noisy estimates of an underlying ground truth; typically, agents tend to favor the ground truth more often than the opposite ones. Our assumption for a majority of well-informed agents is in accordance with this. However, an important feature here is that the ground truth is unobservable (also after the fact). This is a typical assumption in the area of peer prediction mechanisms for unverifiable information (see Faltings and Radanovic [2017], Chapter 3), where a set of agents are used to decide about the quality of data. However, that line of work has a mechanism design flavour and assumes compensations to the agents so that their evaluation of the available data is truthful (e.g., see Witkowski *et al.* [2018]). This is significantly different than our modeling assumptions here. In particular, any evaluation of the quality of the agents — a task that is usually part of crowdsourcing systems; e.g., see Shah *et al.* [2015] — is in our case infeasible. Still, our payment optimization is similar in spirit to automated mechanism design [Sandholm, 2003] but, instead of aiming for truthful agent behaviour, we have a particular equilibrium as target.

Numerous works study voting scenarios where agents expend varying effort to obtain better or worse signals [Gersbach, 1995; Persico, 2004; Gerardi and Yariv, 2008; Michelini *et al.*, 2022]. This line of work typically assumes parties have preferences on the outcome and tries to design mechanisms that incentivize truthful behavior. By contrast, we study scenarios where agents are indifferent to the outcome and use payments to obtain a good outcome.

2 Modeling Assumptions and Notation

We assume that adjudication tasks with two alternatives are outsourced to n agents. We use the integers in $[n] = \{1, 2, \ldots, n\}$ to identify the agents. For an adjudication task, each agent casts a vote for one of the alternatives and the majority of votes defines the adjudication outcome. In the case of a tie, an outcome is sampled uniformly at random. To motivate voting, payments are used. A *payment function* $p : [0, 1] \to \mathbb{R}$ indicates that agent i gets a payment of $p(x)$ when the total fraction of agents casting the same vote as i is x. Payments can be positive or negative (corresponding to monetary transfers to and from the agents, respectively). We make no additional assumptions on the payment functions.

The objective of an adjudication task is to recover the underlying *ground truth*. We denote by T the ground truth and by F the other alternative. We use the terms T-vote and F-vote to refer to a vote for alternative T and F, respectively. To decide which vote to cast, agents put an effort to understand the adjudication case and get a *signal* of whether the correct adjudication outcome is T or F. Each agent i is associated with an *effort function* $f_i : \mathbb{R}_{\geq 0} \to [0, 1]$ which relates the quality of the signal received by an agent with the effort she exerts as follows: the signal agent i gets when she exerts an effort $x \geq 0$ is for the ground truth alternative T with probability $f_i(x)$ and for alternative F with probability $1 - f_i(x)$. We partition the agents into two categories, depending on whether their background knowledge is sufficient so that the quality of the signal they receive increases with extra effort (*well-informed* agents) or worsens (*misinformed* agents). We assume that effort functions are continuously differentiable and have $f_i(0) = 1/2$. The effort function for a well-informed agent i is strictly increasing and strictly concave. The effort function for a misinformed agent is strictly decreasing and strictly convex. The functions $f_i(x) = 1 - \frac{e^{-x}}{2}$ and $f_i(x) = \frac{e^{-x}}{2}$ are representative examples of effort functions for a well-informed and a misinformed agent, respectively.

Agents are rational. They are involved in a *strategic game* where they aim to maximize their *utility*, consisting only of the payment they receive minus the effort they exert. In particular, we assume the agents are entirely indifferent to the outcome. This may lead to voting differently than what their signal indicates. We denote by (λ_i, β_i) the *strategy* of agent i, where λ_i is the effort put and β_i is the probability of casting a vote that is identical to the signal received (and, thus, the agent casts a vote for the opposite alternative with probability $1 - \beta_i$). A strategy of $\beta_i = 1$ thus corresponds to an agent voting truthfully. Allowing for $\beta_i < 1$ models a sce-

nario in which the agents participate mainly to make money and may try each possible strategy that can be beneficial to them (which is the typical assumption in blockchain environments). The utility of an agent is *quasilinear*, i.e., equal to the amount of payments received minus the effort exerted. We assume that agents are *risk neutral* and thus aim to maximize the expectation of their utility. Denote by m_i the random variable indicating the number of agents different than i who cast a T-vote. Clearly, m_i depends on the strategies of all agents besides i but, for simplicity, we have removed this dependency from our notation. Now, the expected utility of agent i when using strategy (λ_i, β_i) is

$$\begin{aligned}
&\mathbb{E}[u_i(\lambda_i, \beta_i, m_i)] \\
&= -\lambda_i + f_i(\lambda_i)\beta_i \cdot \mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right] \\
&\quad + f_i(\lambda_i)(1-\beta_i) \cdot \mathbb{E}\left[p\left(\frac{n-m_i}{n}\right)\right] \\
&\quad + (1-f_i(\lambda_i))\beta_i \cdot \mathbb{E}\left[p\left(\frac{n-m_i}{n}\right)\right] \\
&\quad + (1-f_i(\lambda_i))(1-\beta_i) \cdot \mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right] \\
&= -\lambda_i + \mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right] \\
&\quad + (\beta_i(2f_i(\lambda_i)-1) - f_i(\lambda_i)) \cdot Q(m_i). \qquad (1)
\end{aligned}$$

The quantities $p\left(\frac{1+m_i}{n}\right)$ and $p\left(\frac{n-m_i}{n}\right)$ are the payments agent i receives when she votes for alternatives T and F, respectively. The four positive terms in the RHS of the first equality above are the expected payments for the four cases defined depending on the signal received and whether it is cast as a vote or not. In the second equality, we have used the abbreviation

$$Q(m_i) = \mathbb{E}\left[p\left(\frac{1+m_i}{n}\right) - p\left(\frac{n-m_i}{n}\right)\right],$$

which we also use extensively in the following. Intuitively, given the strategies of the other agents, $Q(m_i)$ is the additional expected payment agent i gets when casting a T-vote compared to an F-vote.

We say that a set of strategies, in which agent $i \in [n]$ uses strategy (λ_i, β_i), is an *equilibrium* in the strategic game induced, if no agent can increase her utility by unilaterally changing her strategy. In other words, the quantity $\mathbb{E}[u_i(x, y, m_i)]$ is maximized with respect to x and y by setting $x = \lambda_i$ and $y = \beta_i$ for $i \in [n]$.

3 Equilibrium Analysis

We are now ready to characterize equilibria. We remark that the cases (a), (b), and (c) of Lemma 1 correspond to the informal terms no correlation, positive correlation, and negative correlation used in the introductory section.

Lemma 1 (equilibrium conditions)**.** *The strategy of agent i at equilibrium is as follows:*

(a) If $|f_i'(0) \cdot Q(m_i)| \leq 1$, then $\lambda_i = 0$ and β_i can have any value in $[0, 1]$.

(b) If $f_i'(0) \cdot Q(m_i) > 1$, then λ_i is positive and such that $f_i'(\lambda_i) \cdot Q(m_i) = 1$ and $\beta_i = 1$.

(c) If $f_i'(0) \cdot Q(m_i) < -1$, then λ_i is positive and such that $f_i'(\lambda_i) \cdot Q(m_i) = -1$ and $\beta_i = 0$.

Proof. First, observe that when agent i selects $\lambda_i = 0$, her expected utility is,

$$\mathbb{E}(u_i(0, \beta_i, m_i)] = \mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right] - \frac{1}{2}Q(m_i),$$

i.e., it is independent of β_i. So, β_i can take any value in $[0, 1]$ when $\lambda_i = 0$.

In case (b), we have $f_i'(0) \cdot Q(m_i) > 0$ which, by the definition of the effort function f_i, implies that $(2f_i(\lambda_i) - 1) \cdot Q(m_i) > 0$ for $\lambda_i > 0$. By inspecting the dependence of expected utility on β_i at the RHS of Eq. (1), we get that if agent i selects $\lambda_i > 0$, she must also select $\beta_i = 1$ to maximize her expected utility. Similarly, in case (c), we have $f_i'(0) \cdot Q(m_i) < 0$ which implies that $(2f_i(\lambda_i) - 1) \cdot Q(m_i) < 0$ for $\lambda_i > 0$. In this case, if agent i selects $\lambda_i > 0$, she also selects $\beta_i = 0$ to maximize her expected utility.

So, in the following, it suffices to reason only about the value of λ_i. Let

$$\begin{aligned}
\Delta_i(\lambda_i) &= \frac{\partial \mathbb{E}[u_i(\lambda_i, \beta_i, m_i)]}{\partial \lambda_i} \\
&= -1 + (2\beta_i - 1) f'(\lambda_i) \cdot Q(m_i) \qquad (2)
\end{aligned}$$

denote the derivative of the expected utility of agent i with respect to λ_i. For (a), by the strict concavity/convexity of f_i, we have $|f_i'(\lambda_i) \cdot Q(m_i)| < 1$ for $\lambda_i > 0$ and,

$$\begin{aligned}
\Delta_i(\lambda_i) &= -1 + (2\beta_i - 1) f_i'(\lambda_i) \cdot Q(m_i) \\
&\leq -1 + |2\beta_i - 1| \cdot |f_i'(\lambda_i) \cdot Q(m_i)| < 0.
\end{aligned}$$

Hence, the expected utility of agent i strictly decreases with $\lambda_i > 0$ and the best strategy for agent i is to set $\lambda_i = 0$.

Otherwise, in cases (b) and (c), the derivative $\Delta_i(\lambda_i)$ has strictly positive values for λ_i arbitrarily close to 0 (this follows by the facts that f is strictly convex/concave and continuously differentiable), while it is clearly negative as λ_i approaches infinity (where the derivative of f approaches 0). Hence, the value of λ_i selected by agent i at equilibrium is one that nullifies the RHS of (2), i.e., such that $f_i'(\lambda_i) \cdot Q(m_i) = 1$ in case (b) and $f_i'(\lambda_i) \cdot Q(m_i) = -1$ in case (c). □

Using Lemma 1, we can now identify some properties about the structure of equilibria.

Lemma 2. *For any payment function, no effort by all agents (i.e., $\lambda_i = 0$ for $i \in [n]$) is an equilibrium.*

The proof of Lemma 2 is omitted. We will use the term *non-trivial* for equilibria having at least one agent putting some effort.

The next lemma reveals the challenge of adjudication in our strategic environment. It essentially states that for every equilibrium that yields probably correct adjudication, there is an equilibrium that yields probably incorrect adjudication with the same probability.

Lemma 3. *For any payment function, if the set of strategies $(\lambda_i, \beta_i)_{i\in[n]}$ is an equilibrium, so is the set of strategies $(\lambda_i, 1-\beta_i)_{i\in[n]}$.*

We say that an equilibrium is *simple* if there exists an alternative $a \in \{T, F\}$ such that all agents cast a vote for alternative a with probability $\geq 1/2$. Intuitively, this makes prediction of the agents' behaviour at equilibrium easy. Together with Lemma 1, this implies that, in a simple equilibrium, an agent putting no effort (i.e., $\lambda_i = 0$) can use any strategy β_i. For agents putting some effort, a well-informed agent uses $\beta_i = 1$ if $a = T$ and $\beta_i = 0$ if $a = F$ and a misinformed agent uses $\beta_i = 0$ if $a = T$ and $\beta_i = 1$ if $a = T$.

Lemma 4 (simple equilibrium condition)**.** *When the payment function p satisfies*

$$p\left(\frac{2+m}{n}\right) - p\left(\frac{1+m}{n}\right) + p\left(\frac{n-m}{n}\right) - p\left(\frac{n-m-1}{n}\right) \geq 0, \qquad (3)$$

for every $m \in \{0, 1, \ldots, n-2\}$, all equilibria are simple.

It can be verified that the payment function

$$p(x) = \begin{cases} \frac{\omega}{xn}, & x \geq 1/2 \\ -\frac{\ell}{xn}, & x < 1/2 \end{cases}$$

with $\omega \leq \ell$ satisfies the condition of Lemma 4. We refer to this function as the award/loss sharing payment function. Essentially, the agents with the majority vote share an award of ω while the ones in minority share a loss of ℓ. Note that for $\omega = \ell$, the payment function is strictly budget balanced unless all votes are unanimous. This is similar to the payment function used in Kleros. A sufficient condition for simple equilibria which is quite broad but does not include Kleros' payments is the following.

Corollary 5. *When the payment functions are monotone non-decreasing, all equilibria are simple.*

4 Selecting Payments for Correct Adjudication

We now focus on the very simple scenario in which some of the n agents are well-informed and have the same effort function f and the rest are misinformed and have the effort function $1-f$. Can we motivate an expected x-fraction of them to vote for the ground truth?

Of course, we are interested in values of x that are higher than $1/2$. This goal is directly related to asking for a high probability of correct adjudication. Indeed, as the agents cast their votes independently, the realized number of T-votes is sharply concentrated around their expectation and thus the probability of incorrect adjudication is exponentially small in terms of the number of agents n and the quantity $(x-1/2)^2$. This can be proved formally by a simple application of well-known concentration bounds, e.g., Hoeffding's inequality [Hoeffding, 1963].

So, our aim here is to define appropriate payment functions so that a set of strategies leading to an expected x-fraction of T-votes is an equilibrium. We will restrict our attention to payments satisfying the condition of Lemma 4; then, we know that all equilibria are simple. We will furthermore show that all equilibria are *symmetric*, in the sense that all agents cast a T-vote with the same probability. This means that there are $\lambda > 0$ and $\beta \in \{0, 1\}$ so that all well-informed agents use strategy (λ, β) and all misinformed agents use the strategy $(\lambda, 1-\beta)$.

Lemma 6. *Consider the scenario with n agents, among which the well-informed agents use the same effort function f and the misinformed agents use the effort function $1-f$. If the payment function p satisfies the condition of Lemma 4, then all equilibria are symmetric.*

Lemma 6 implies that, for $x > 1/2$, an equilibrium with an expected x-fraction of T-votes has each agent casting a T-vote with probability $f(\lambda) = x$; the well-informed agents use the strategy $(\lambda, 1)$ and the misinformed agents use $(\lambda, 0)$. As agents vote independently, the random variables m_i follow the same binomial distribution $\text{Bin}(n-1, x)$ with $n-1$ trials, each having success probability x. Also, notice that the fact that the effort function is strictly monotone implies that λ is uniquely defined from x as $\lambda = f^{-1}(x)$.

We now aim to solve the optimization task of selecting a payment function p which satisfies the conditions of Lemma 4, induces as equilibrium the strategy $(\lambda, 1)$ for well-informed agents and the strategy $(\lambda, 0)$ for misinformed agents, ensures non-negative expected utility for all agents (individual rationality), and minimizes the expected amount given to the agents as payment. As all agents cast a T-vote with the same probability and the quantities m_i are identically distributed for different is, it suffices to minimize the expected payment

$$x \cdot \mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right] + (1-x) \cdot \mathbb{E}\left[p\left(\frac{n-m_i}{n}\right)\right] \qquad (4)$$

of a single agent. By the definition of expected utility in equation (1), restricting this quantity to values at least as high as $f^{-1}(x)$ gives the individual rationality constraints for all agents. Furthermore, by Lemma 1, the equation,

$$f'(f^{-1}(x)) \cdot Q(m_i) = 1, \qquad (5)$$

gives the equilibrium condition for both well-informed and misinformed agents.

We can solve the optimization task above using linear programming. Our LP has the payment parameters $p(1/n)$, $p(2/n), \ldots, p(1)$ as variables. The linear inequalities (3) for $m \in \{0, 1, \ldots, n-2\}$ form the first set of constraints, restricting the search to payment functions satisfying the conditions of Lemma 4. Crucially, observe that the quantities $\mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right]$ and $\mathbb{E}\left[p\left(\frac{n-m_i}{n}\right)\right]$ and, subsequently, $Q(m_i)$, can be expressed as linear functions of the payment parameters. Indeed, for $t = 0, 1, \ldots, n-1$, let $z(t) = \Pr[m_i = t]$ be the known probabilities of the binomial distribution $\text{Bin}(n-1, x)$. Clearly,

$$\mathbb{E}\left[p\left(\frac{1+m_i}{n}\right)\right] = \sum_{t=0}^{n-1} z(t) \cdot p\left(\frac{1+t}{n}\right),$$

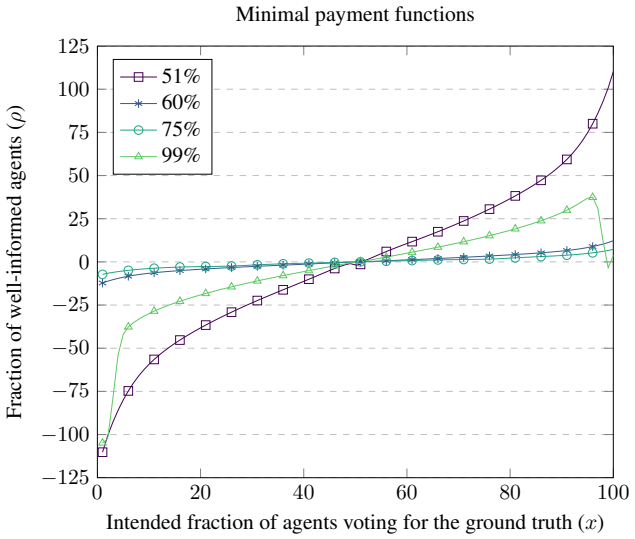

Figure 1: Minimal payment functions that ensure the existence of a simple equilibrium with an x-fraction of the agents casting a T-vote on average, so that all agents have non-negative expected utility. The scenario uses $n = 100$ and the effort function $f(x) = 1 - \frac{e^{-x}}{2}$. The payment functions obtained by solving the linear program from Theorem 7 for $x \in \{0.51, 0.6, 0.75, 0.99\}$ are shown. The sharp decline of the curve for $x = 99\%$ is due to the LP solver wanting to keep the expected fraction of voters casting a T-vote at exactly 99%. We suspect that this is not a unique solution and a smoother payment function would also be close to the minimal payments.

and,

$$\mathbb{E}\left[p\left(\frac{n-m_i}{n}\right)\right] = \sum_{t=0}^{n-1} z(t) \cdot p\left(\frac{n-t}{n}\right).$$

Thus, the objective function (4), the individual rationality constraint, and the equilibrium condition constraint can be expressed as linear functions of the LP variables. Overall, the LP has n variables and $n+1$ constraints (n inequalities and one equality). We may summarize this discussion as follows.

Theorem 7. *Consider the scenario with n agents, where the well-informed ones have the same effort function f and the misinformed ones have the same effort function $1-f$. Given $x \in (1/2, 1)$, selecting the payment function that satisfies the conditions of Lemma 4, induces an equilibrium in which all agents have non-negative expected utility and an expected x-fraction of agents casts a T-vote so that the expected amount given to the agents as payment is minimized, can be done in time polynomial in n using linear programming.*

Our approach can be extended to include additional constraints (e.g., non-negativity or monotonicity of payments), provided they can be expressed as linear constraints of the payment parameters. Fig. 1 depicts four payment solutions obtained by solving the above LP for $n = 100$ and the effort function $f(x) = 1 - \frac{e^{-x}}{2}$, and values of $x \in \{0.51, 0.60, 0.75, 00.99\}$.

5 Computational Experiments

Our goal in this section is to justify that appropriate selection of the payment parameters can lead to correct adjudication in practice, even though Lemma 3 shows the co-existence of both good and bad equilibria. The key property that favours good equilibria more often is that, in practice, jurors are on average closer to being well-informed than misinformed. Formally, this means that $\frac{1}{n} \cdot \sum_{i \in [n]} f_i(x) > 1/2$ for every $x > 0$.

Due to the lack of initial feedback, it is natural to assume that agents start their interaction by putting some small effort and convert their signal to a vote. We claim that this, together with their tendency to being well-informed, is enough to lead to correct adjudication despite strategic behaviour. We provide evidence for this claim through the following experiment implementing the scenario we considered in Section 4.

We have n agents, a ρ-fraction of whom are well-informed and the rest are misinformed. Agent i's effort function is $f_i(x) = 1 - \frac{e^{-x}}{2}$ if she is well-informed and $f_i(x) = \frac{e^{-x}}{2}$ if she is misinformed. We consider the minimal payment functions, defined as the solution of the linear program detailed in the last section, parameterized by the fraction x of agents intended to vote for the ground truth. A small subset of these payment functions can be seen in Fig. 1. In addition, we consider two different payment functions, both defined using a parameter $\omega > 0$:

- $p(x) = \omega$ if $x \geq 1/2$ and $p(x) = 0$, otherwise.
- $p(x) = \frac{\omega}{xn}$ if $x \geq 1/2$ and $p(x) = -\frac{\omega}{xn}$, otherwise.

With the first payment function, each agent gets a payment of ω if her vote is in the majority, while she gets no payment otherwise. With the second payment, the agents in the majority share an award of ω, while the agents in the minority share a loss of ω. Notice that both payment functions satisfy the conditions of Lemma 4. We will refer to them as *threshold* and *award/loss sharing* payment functions, respectively.

In our experiments, we simulate the following dynamics of strategic play. Initially, all agents put an effort of $\epsilon > 0$ and cast the signal they receive as vote. In subsequent rounds, each agent best-responds. In particular, the structure of the dynamics is as follows:

Round 0**:** Agent i puts an effort of ϵ and casts her signal as vote.

Round j**, for** $j = 1, 2, \ldots, R$**:** Agent i gets m_i as feedback. She decides her strategy $\beta_i \in \{0, 1\}$ and effort level $\lambda_i \geq 0$. She draws her signal, which is alternative T with probability $f_i(\lambda_i)$ and alternative F with probability $1 - f_i(\lambda_i)$. If $\beta_i = 1$, she casts her signal as vote; otherwise, she casts the opposite of her signal as vote.

In each round after round 0, agents get the exact value of m_i as feedback (as opposed to its distribution)[1] but maximize their expected utility with respect to the components λ_i and β_i of their strategy. Hence, the only difference with what we have seen in earlier sections is that the calculation of expected

[1] An alternative implementation would assume that m_i takes the number of T-votes in a randomly chosen previous round. The results obtained in this way are qualitatively similar to those we present here.

utility considers the actual value of payments and not their expectation, i.e.,

$$\mathbb{E}[u_i(\lambda_i, \beta_i, m_i)] = -\lambda_i + p\left(\frac{1+m_i}{n}\right) + (\beta_i(2f_i(\lambda_i) - 1) - f_i(\lambda_i)) \cdot Q(m_i),$$

where

$$Q(m_i) = p\left(\frac{1+m_i}{n}\right) - p\left(\frac{n-m_i}{n}\right).$$

By applying Lemma 1, we get the following characterization of the best-response of agent i in round $j > 0$.

Corollary 8. *The best response of agent i receiving feedback m_i is as follows:*

(a) If $|Q(m_i)| \leq 2$, then $\lambda_i = 0$ and β_i can take any value in $[0, 1]$.

(b) Otherwise, $\lambda_i = \ln \frac{|Q(m_i)|}{2}$.

(b.1) If agent i is well-informed and $Q(m_i) > 2$ or agent i is misinformed and $Q(m_i) < -2$, then $\beta_i = 1$.

(b.2) If agent i is misinformed and $Q(m_i) > 2$ or agent i is well-informed and $Q(m_i) < -2$, then $\beta_i = 0$.

In our experiments, we consider an agent population of fixed size $n = 100$, with the fraction of well-informed agents ranging from 0 to 1. We simulate the dynamics described above for $R = 50$ rounds and repeat each simulation 20 times. For each experiment, we measure the frequency with which the majority of votes after the R^{th} round is for the ground truth alternative T. We do so for both the threshold and award/loss sharing payment functions, with parameter $\omega \in [0, 5]$ for the threshold payment functions and $\omega \in [0, 100]$ for the award/loss sharing one. We also consider the payment functions that arise as solutions to the linear programs considered in the previous section. In each experiment, we play with the values of two parameters simultaneously. We consider 100 values on each axis and plot the resulting data using a heatmap, with each data point corresponding to the average correctness observed during the experiment.

In the first experiment (Fig. 2.a), we consider the threshold payment function and vary the size of the reward ω and the fraction ρ of well-informed agents. We consider a reasonably high starting effort of $\epsilon = 1$, corresponding to a probability of 0.816 of receiving the ground truth as signal. We observe two distinct regions as we vary the size of the payment. Initially, when the payment is too small (i.e. $\omega \leq 2.5$), the outcome of the adjudication is mostly random. When the payment increases above the threshold, we observe a *sharp phase transition* independent of ρ, where the correctness is extremified by the payment in the following sense: when ρ is sufficiently large (respectively, small), the mechanism recovers the ground truth with high (respectively, low) probability.

In the second experiment (Fig. 2.b), we consider the award/loss sharing payment function. The range of ω is changed from $[0, 5]$ to $[0, 100]$. All other parameters are kept the same. We obtain similar results as for the threshold payment function, i.e. the outcome is mostly random below a threshold above which we observe a sharp phase transition where the outcome of the mechanism is extremified.

In the third experiment (Fig. 2.c), we observe the effect on the correctness by the initial effort. We fix the threshold payment function with $\omega = 3$ such that mechanism has a chance to recover the ground truth, and let ϵ range from 0 to 5. We observe that, when ϵ is small, the outcome of the mechanism is mostly random, while the outcome quickly extremifies as ϵ increases. This means the mechanism only works if the agents initially put in sufficient effort. The results are similar for both the award/loss sharing payment function and the minimal payment functions.

In the fourth experiment (Fig. 2.d), we consider the payment functions obtained from Theorem 7. A subset of the payment functions we use are depicted in Fig. 1. Here, instead of varying the size of the reward, we vary the parameter x used as input to the linear program. This parameter represents the intended fraction of agents voting for the ground truth at equilibrium. We let x range from 0.51 to 1 in increments of 0.01. Here, we observe that for x close to 0.5 and for x close to 1, the mechanism is extremified, while for x close to 0.75 and ρ close to 0.5 the outcome of the mechanism is mostly random. This is rather unexpected since if a 0.75-fraction of the agents vote for the ground truth, the majority vote will be for the ground truth almost certainly. Indeed, we observe that in these games when $\rho \approx 0.5$, the agents exert effort close to zero, hence producing the random outcome. We claim that despite this behavior, the ground truth is still an equilibrium, it is just not a stable equilibrium and the parties converge to the trivial equilibrium.

In a fifth experiment (Fig. 2.e), we consider a different set of minimal payment functions, obtained by relaxing the equality constraint Eq. (5) to a lower bound inequality. This has the effect of no longer requiring an exact x-fraction of the agents vote for the ground truth, but instead gives a lower bound on their number. This slightly changes the payment functions, though they are qualitatively similar to those shown in Fig. 1. Here, we again vary the fraction ρ of well-informed agents on the y-axis, and the intended fraction x of agents voting for the ground truth, ranging from $x = 0.51$ to $x = 1$ in increments of 0.01. However, we obtain different and considerably better results than those in Fig. 2.d. In particular, we obtain a good adjudication outcome for any x when $\rho > 0.75$.

In our sixth and final experiment (Fig. 2.f), we aim to explain the enigmatic behaviour of the LP-computed payments for $x \approx 0.75$. We fix the payment function to be the minimal payment function with $x = 0.75$ and vary the number of rounds as $R = 1 \ldots 100$. We do not take into account round 0 where all parties exert $\epsilon > 0$ effort in the estimation of the correctness of the outcome. We observe that the outcome is extremified when the number of rounds is small and decays as we increase the number of rounds. We can explain this result by considering the payment function for $x = 0.75$ in Fig. 1 whose distribution is mostly flat when the outcome is close to being a tie. Here, the value of $Q(m_i)$ is small so the agent will lower the effort they exert, making it more likely that the outcome will be disputed. This creates a pull towards the trivial equilibrium. By contrast, the curves for $x \in \{0.51, 0.99\}$ have a higher slope close to 0.5, which makes this effect less pronounced. This explains why the ad-

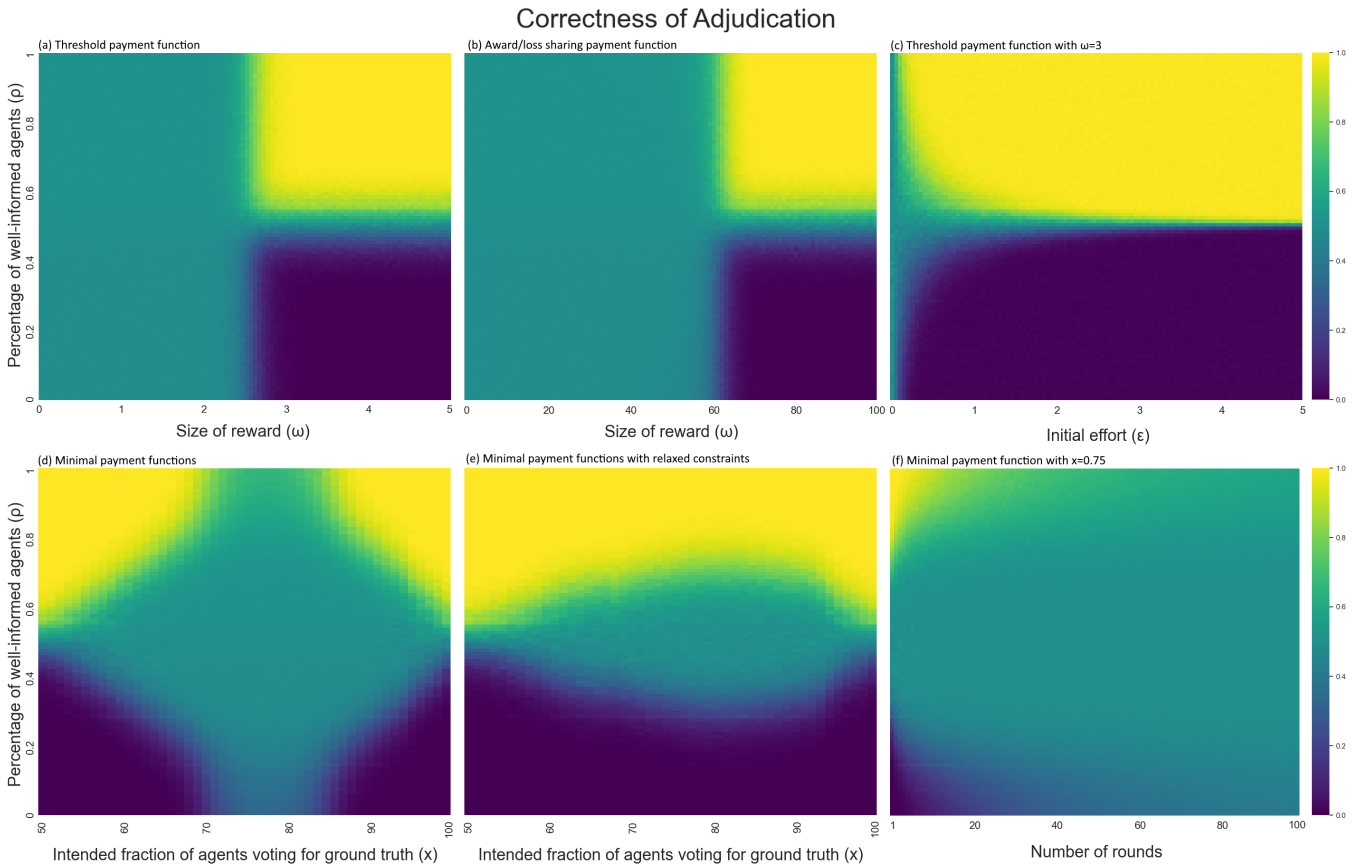

Figure 2: Heatmap of the correctness of the adjudication, plotted with the fraction of well-informed agents on the y-axis, with six varying x-axes. In each plot, we run $R = 50$ rounds with a jury of size $n = 100$, using 1000 samples for each data point. The color of a data point indicates the average measured correctness with the given parameters, using the viridis color scale displayed in the legend on the right. Yellow corresponds to good recovery, while dark blue corresponds to poor recovery of the ground truth, while random outcomes are represented by turquoise. The six x-axes are as follows: (a) Size of the reward for the threshold payment function, ranging from $\omega = 0$ to $\omega = 5$, with $\epsilon = 1$. (b) Size of the reward for the award/loss sharing payment function, ranging from $\omega = 0$ to $\omega = 100$, with $\epsilon = 1$. (c) The initial effort ϵ, ranging from $\epsilon = 0$ to $\epsilon = 5$, with the payment function being the threshold payment function with $\omega = 3$. (d) The intended fraction x of agents voting for the ground truth, ranging from $x = 0.51$ to $x = 1$, with the payment function defined by Theorem 7. (e) The intended fraction x of agents voting for the ground truth, ranging from $x = 0.51$ to $x = 1$, with the payment functions obtained from Theorem 7 by relaxing Eq. (5) to an inequality. (f) The number of rounds, ranging from $R = 1$ to $R = 100$, with the payment function being the minimal payment function with $x = 0.75$ from Fig. 1.

judication outcome is mostly random for $x \approx 0.75$. By design, the linear program finds minimal payments that ensure there is an equilibrium where an x-fraction of the agents vote in favor of the ground truth. However, it does not constrain the solution to have the property that the good equilibrium is *stable*. In some sense, the fact that the non-trivial equilbrium is stable when x is far from 0.5 is happenstance and begs the deeper question why the solutions to the linear program are of the form we observe. Intuitively, it makes sense that attaining a high accuracy requires large payments. A similar phenomenon seemingly holds for accuracies close to 0.51 which can be explained informally as follows. Combinatorially, there are only a few ways to attain an accuracy of 0.51 which necessitates the use of large punishment and rewards when the vote is close to being a tie. By contrast, for larger ρ, there are more ways to attain an accuracy of 0.75 in the majority, hence loosening the requirements on the payments. This suggests that the case $x = 0.75$ does not provide positive results in practice because of *instability of equilibria*. It would be interesting to explore whether it is possible to extend our approach with additional natural constraints that ensure the non-trivial equilibrium is also stable.

Our experiments suggest that several classes of payment functions can be used to recover the ground truth with high probability, provided the agents are well-informed on average. Clearly, there is much work yet to be done in designing payment functions with desirable properties: while the threshold function and the award/loss sharing function seem to recover the ground truth reliably, it might be difficult in practice to pinpoint the location of the phase transition, as this requires estimating the effort functions used by actual jurors. The same holds true for the minimal payment functions.

Acknowledgments

We would like to thank Luca Nizzardo, Irene Giacomelli, Matteo Campanelli, and William George for interesting discussions in several stages of this work. IC was partially supported by a research advisorship grant from Protocol Labs.

References

[Arrow, 1951] Kenneth J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, 1951.

[Brandt *et al.*, 2016] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[Caragiannis *et al.*, 2016] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation*, 4(3):15:1–15:30, 2016.

[Conitzer and Sandholm, 2005] Vincent Conitzer and Tuomas Sandholm. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 145–152, 2005.

[Faltings and Radanovic, 2017] Boi Faltings and Goran Radanovic. *Game Theory for Data Science: Eliciting Truthful Information*. Morgan & Claypool Publishers, 2017.

[Gerardi and Yariv, 2008] Dino Gerardi and Leeat Yariv. Information acquisition in committees. *Games and Economic Behavior*, 62(2):436–459, 2008.

[Gersbach, 1995] Hans Gersbach. Information efficiency and majority decisions. *Social Choice and Welfare*, 12(4):363–370, 1995.

[Gibbard, 1973] Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

[Hoeffding, 1963] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[Lesaege *et al.*, 2019] Clément Lesaege, Federico Ast, and William George. Kleros Short Paper v1.0.7. Technical report, Kleros, 09 2019.

[Lesaege *et al.*, 2021] Clément Lesaege, William George, and Federico Ast. Kleros Long Paper v2.0.2. Technical report, Kleros, 07 2021.

[Michelini *et al.*, 2022] Matteo Michelini, Adrian Haret, and Davide Grossi. Group wisdom at a price: Jury theorems with costly information. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 419–425, 2022.

[Persico, 2004] Nicola Persico. Committee Design with Endogenous Information. *The Review of Economic Studies*, 71(1):165–191, 01 2004.

[Posner and Weyl, 2018] Eric A. Posner and E. Glen Weyl. *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*. Princeton University Press, 2018.

[Sandholm, 2003] Tuomas Sandholm. Automated mechanism design: A new application area for search algorithms. In *Proceedings of the 9th International Conference on Principles and Practice of Constraint Programming (CP)*, pages 19–36, 2003.

[Satterthwaite, 1975] Mark. A. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

[Schwartzbach, 2021] Nikolaj I. Schwartzbach. An incentive-compatible smart contract for decentralized commerce. In *Proceedings of the 2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–3, 2021.

[Shah *et al.*, 2015] Nihar B. Shah, Dengyong Zhou, and Yuval Peres. Approval voting and incentives in crowdsourcing. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, page 10–19, 2015.

[Witkowski *et al.*, 2018] Jens Witkowski, Rupert Freeman, Jennifer Vaughan, David Pennock, and Andreas Krause. Incentive-compatible forecasting competitions. In *Proceedings of 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[Young, 1988] H. Peyton Young. Condorcet's theory of voting. *American Political Science Review*, 82(4):1231–1244, 1988.