Sharing Rewards Based on Subjective Opinions

by

Arthur Carvalho

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Arthur Carvalho

Abstract

Fair division is the problem of dividing one or several goods among a set of agents in a way that satisfies a suitable fairness criterion. Traditionally studied in economics, philosophy, and political science, fair division has drawn a lot of attention from the multiagent systems community, since this field is strongly concerned about how a surplus (or a cost) should be divided among a group of agents.

Arguably, the Shapley value is the single most important contribution to the problem of fair division. It assigns to each agent a share of the resource equal to the expected marginal contribution of that agent. Thus, it is implicitly assumed that individual marginal contributions can be objectively computed. In this thesis, we propose a game-theoretic model for sharing a joint reward when the quality of individual contributions is subjective.

In detail, we consider scenarios where a group has been formed and has accomplished a task for which it is granted a reward, which must be shared among the group members. After observing the contribution of the peers in accomplishing the task, each agent is asked to provide evaluations for the others. Mainly to facilitate the sharing process, agents can also be requested to provide predictions about how their peers are evaluated. These subjective opinions are elicited and aggregated by a central, trusted entity, called the mechanism, which is also responsible for sharing the reward based exclusively on the received opinions.

Besides the formal game-theoretic model for sharing rewards based on subjective opinions, we propose three different mechanisms in this thesis. Our first mechanism, the *peer-evaluation mechanism*, divides the reward proportionally to the evaluations received by the agents. We show that this mechanism is fair, budget-balanced, individually rational, and strategy-proof, but that it can be collusion-prone.

Our second mechanism, the *peer-prediction mechanism*, shares the reward by considering two aspects: the evaluations received by the agents and their truth-telling scores. To compute these scores, this mechanism uses a *strictly proper scoring rule*. Under the assumption that agents are Bayesian decision-makers, we show that this mechanism is weakly budget-balanced, individually rational, and incentive-compatible. Further, we present approaches that guarantee the mechanism to be collusion-resistant and fair.

Our last mechanism, the *BTS mechanism*, is the only one to elicit both evaluations and predictions from the agents. It considers the evaluations received by the agents and their truth-telling scores when sharing the reward. For computing the scores, it uses the Bayesian truth serum method, a powerful scoring method based on the surprisingly common criterion. Under the assumptions that agents are Bayesian decision-makers, and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations, we show that this mechanism is incentive-compatible, budget-balanced, individually rational, and fair.

Acknowledgements

First, and foremost, I would like to thank my supervisor, Professor Kate Larson. The freedom that she gave me and the constant support that she provided through this program makes her a perfect model of how a supervisor should behave. I would also like to thank the readers of my thesis, Professor Robin Cohen and Professor Pascal Poupart, for their valuable feedback.

Dedication

This thesis is dedicated to my parents, Ulissiano Batista and Remédios Carvalho, my brother, Jardel Carvalho, and my partner, Carolina da Paz. I just cannot describe how important you are in my life. This work would never have been possible without your support and love.

Contents

List of Tables

List of Figures

List of Algorithms

Glossary

α	A constant that fine-tunes the weight given to the truth-telling scores
β	A constant that fine-tunes the weight given to the prediction score in the BTS method
$C(\mathbf{p} \mid \mathbf{q})$	The expected score for the stated assessment $\mathbf{p}$ at the true opinion $\mathbf{q}$
$\Gamma(\mathbf{s})$	A function that maps each strategy profile to a vector of shares
$\Gamma_i(\mathbf{s})$	The share of V given to agent i when all the reported opinions are $\mathbf{s}$
ϵ	A recalibration coefficient used in the BTS method
Φ_i^j	a M-dimensional unit vector, where $\Phi_i^{j^k} = 1$, if $x_i^j = k$, for $1 \leq k \leq M$, and 0 otherwise.
ζ_i	The truth-telling score of agent i
$h(x_i^j, k)$	A function indicating if the evaluation $x_i^j = k$
$L(\mathbf{p} \mid \mathbf{q})$	The expected score loss for the stated assessment $\mathbf{p}$ at the true opinion $\mathbf{q}$
M	The top possible evaluation that an agent can receive or give
μ_i	The average evaluation received by agent i
N	A set of agents that must share the reward V

n	The size of the set N
$nint$	The nearest integer function
Ω_i	Belief of agent i
$R(\mathbf{p}, e)$	A function that provides a score for the assessment $\mathbf{p}$ upon observing the event e
$\mathbf{r}_i$	Vector with agent i's truthful predictions
$\mathbf{r}_i^j$	Agent i's prediction about the empirical distribution of evaluations received by agent j
r_i^{jk}	Agent i's truthful prediction on the percentage of agents that give the evaluation k to agent j
S	The set with available strategies
S_i	The set with the strategies available to agent i
$\overset{*}{\mathbf{s}}_i$	The truthful strategy of agent i
$\hat{\mathbf{s}}_i$	The false strategy of agent i
$\mathbf{s}$	A strategy profile
$\mathbf{s}_i$	The strategy of agent i
$serum(i, j)$	The score provided by the Bayesian truth serum method to agent i given its evaluation x_i^j and prediction $\mathbf{y}_i^j$
$\mathbf{t}_i$	The vector with agent i's truthful evaluations
t_i^j	Agent i's truthful evaluation for agent j
V	The reward to be shared
ω_j	A common prior over the evaluations for agent j
$\bar{\chi}^i$	The aggregation of the scaled evaluations received by agent i
$\bar{x}_k$	The average frequency of the evaluation k
χ_i^j	The scaled evaluation given by agent i to agent j
χ_i	The scaled evaluations reported by agent i
$\mathbf{x}_i$	The vector with agent i's reported evaluations

x_i^j	Agent i's reported evaluation for agent j
$\bar{y}_k$	The geometric average of the predicted frequencies for the evaluation k
$\mathbf{y}_i$	The vector with agent i's reported predictions
$\mathbf{y}_i^j$	Agent i's reported prediction about the empirical distribution of evaluations received by agent j
$y_i^{j^k}$	Agent i's reported prediction on the percentage of agents that give the evaluation k to agent j

Chapter 1

Introduction

Fair division is the problem of dividing one or several resources among a set of agents in a way that satisfies a suitable fairness criterion. The first step toward a formal definition of fair division was perhaps given by Aristotle, more than 2000 years ago:

> "Equals should be treated equally, and unequals unequally, in proportion to relevant similarities and differences" (Nicomachean Ethics).

Traditionally studied in economics, philosophy, and political science, fair division has drawn a lot of attention from the multiagent systems community, since this field is strongly concerned about how a surplus (or a cost) should be divided among a group of agents [42]. The modern philosophical treatment of fair division involves four basic principles [30, 15, 8]:

- **Exogenous Right:** Certain principles guiding the allocation of the resources are external to the consumption of the same and to the responsibility of the consumers in their production;
- **Compensation:** Agents who most need the resources are privileged with greater shares of them;
- **Reward:** The distribution of the resources is based on individual contributions for producing the same;
- **Fitness:** Resources must go to whomever makes the best use of them.

The idea of *collective welfare*, from the modern microeconomic thinking, yields a systematic interpretation of fitness and compensation by means of Pareto optimality and

collective utility functions. In the last seventy years, microeconomic theory has made significant progress toward understanding the reward principle [30]. Arguably, its single most important discovery is the concept known as *Shapley value* [41], which gives an interpretation of the reward principle in the context of production.

The Shapley value is a systematic formula used to divide a joint cost (or surplus) among a group of agents, where the share assigned to each member is its average marginal cost (or surplus). Thus, the Shapley value is computed directly from the cost or production function, and it implicitly assumes that the marginal contribution of each agent can be objectively obtained. However, there exist settings where the quality of such contributions is subjective, or cannot be objectively obtained. For example, employees of cooperatives may share the profit from successfully completing some project. While employees who contributed most to the success of the project should receive a greater share of the profit, it may be difficult to pin-point what the key contributions were from each employee. Similarly, when a professor wants to share a collective grade among a group of students, individual contributions to the success of the group may be subjective and difficult to separate.

Focusing on these settings where the quality of the contributions of each group member is subjective, we propose in this thesis a game-theoretic model for sharing a joint reward based on the idea of *subjective opinions*. In detail, we consider scenarios where a group has been formed and has accomplished a task for which it is granted a reward, which must be shared among the group members. After observing the contribution of the peers in accomplishing the task, each agent is asked to provide *evaluations* for the others. Mainly to facilitate the sharing process, agents can also be requested to provide *predictions* about how their peers are evaluated. Thus, we consider two kinds of subjective opinions: evaluations and predictions. These opinions are elicited and aggregated by a central, trusted entity, called the *mechanism*, which is also responsible for sharing the reward based exclusively on the received opinions.

Every agent is assumed to want more of the reward. Therefore, we can identify an agent's share with its welfare. In terms of the general fair division principles, the focus of this thesis is almost exclusively on the reward principle, and, more specifically, on the interpretation of "individual contributions".

Since individual contributions are derived from subjective opinions, we propose a fairness criterion more appropriate to our model. It essentially means that if an agent unanimously receives better evaluations than a peer, then this agent should also receive a greater share of the reward than that peer.

A major concern that arises when eliciting and aggregating subjective opinions from rational agents is to guarantee honest reporting. For example, an agent may deliberately lie and give all other agents a low evaluation so that, in comparison, it looks good and receives a greater share of the reward. Further, agents can deliberately lie in their evaluations

looking for side-payments from the beneficiaries. Thus, an important issue in our work is to ensure that each agent is better off telling the truth than lying.

1.1 Contributions

Besides the game-theoretic model for sharing a reward based on subjective opinions, we propose in this thesis three different mechanisms to elicit and aggregate opinions, as well as for determining agents' shares, keeping the issues of truthfulness and fairness in mind.

Our first mechanism, the *peer-evaluation mechanism*, divides the reward proportionally to the evaluations received by the agents. We show that this mechanism is fair, budget-balanced, individually rational, and strategy-proof, but that it can be collusion-prone.

Our second mechanism, the *peer-prediction mechanism*, shares the reward by considering two aspects: the evaluations received by the agents and their truth-telling scores. To compute these scores, this mechanism uses a *strictly proper scoring rule* [47]. Under the assumption that agents are Bayesian decision-makers, we show that this mechanism is weakly budget-balanced, individually rational, and incentive-compatible. Further, we present approaches that guarantee the mechanism to be collusion-resistant and fair.

Finally, our last mechanism, called the *BTS mechanism*, is the only one to elicit both evaluations and predictions from agents. It considers the evaluations received by the agents and their truth-telling scores when sharing the reward. For computing the scores, it uses the Bayesian truth serum method [32], a powerful scoring method based on the surprisingly common criterion. Under the assumptions that agents are Bayesian decision-makers, and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations, we show that our mechanism is incentive-compatible, budget-balanced, individually rational, and fair.

1.2 Thesis Organization

Besides this introductory chapter, the rest of this thesis is organized as follows:

In Chapter 2, we present our basic model, mathematical notation, and concepts used throughout this thesis.

In Chapter 3, we present the peer-evaluation mechanism, our first mechanism for sharing rewards using subjective opinions. It takes the stand-alone evaluations received by the agents and divides the reward proportionally to them. We show that this simple, yet powerful, mechanism is fair, budget-balanced, individually rational, and strategy-proof, but that it can be collusion-prone.

In Chapter 4, we present the peer-prediction mechanism. The main difference in relation to our previous mechanism is that now we explicitly incentivise truth-telling by using scoring rules. In detail, each evaluation submitted by an agent can be seen as a bet on the average of the others' evaluations. Under mild assumptions, we show that the best bet of an agent, in an expected sense, is to truthfully report its evaluations. Also, we show that this mechanism is weakly budget-balanced, individually rational, and we discuss the trade-off that exists between fairness and collusion-resistance.

In Chapter 5, we present our last mechanism, the BTS mechanism. It is based on the theory that a knowledgeable agent is not only informed about its truthful evaluation for a peer, but it may also know the likely distribution of the evaluations received by that peer. This allows us to use the Bayesian truth serum method to incentivise truthfulness. Under the assumptions that agents are Bayesian decision-makers and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations, we show that the BTS mechanism is incentive-compatible and budget-balanced, and we present strategies to guarantee that this mechanism will always be individually rational and fair.

In Chapter 6, we empirically investigate the influence of the mechanisms' parameters on agents' shares. We study the consequences of using different values for the top possible evaluation that an agent can give or receive, and for the parameter that fine-tunes the weight given to the truth-telling scores. Also, we investigate how the proposed mechanisms behave when dealing with populations of different sizes.

In Chapter 7, we review the literature related to our work, pointing out the differences and similarities to our model and mechanisms. In detail, we review similar ideas from economics, multiagent systems, game theory, cake-cutting, and mechanism design literature.

In Chapter 8, we conclude our work by highlighting its main contributions. We also suggest future work to expand our proposed model and mechanisms.

Chapter 2

Background

In this chapter, we present our basic model, mathematical notation, and concepts used throughout this thesis.

2.1 Model

A set of agents $N = \{1, \ldots, n\}$, for $n \geq 3$, has accomplished a task for which it is granted a *reward* $V \in \Re^+$. Every agent is assumed to want more of the reward. Therefore, we can identify an agent's share with its welfare. We are interested in settings where the share of V that an agent receives depends, in a meaningful way, on the *subjective opinions* of its peers concerning that agent's contribution to the group.

Each agent privately observes $n-1$ signals, each one related to a peer. These signals are direct assessments of the peers' performance in accomplishing the task. Given a positive integer parameter $1 \leq M \leq V$, the signals observed by an agent $i \in N$ are formally represented by the vector $\mathbf{t}_i = (t_i^1, \ldots, t_i^{i-1}, t_i^{i+1}, \ldots, t_i^n)$, where $t_i^j \in \{1, \ldots, M\}$ represents the signal observed by agent i about agent j's performance. We call $\mathbf{t}_i$ the *truthful evaluations* made by agent i. Agents are requested to report the observed signals. Thus, the parameter M represents the top possible evaluation that an agent can give or receive.

Based on their truthful evaluations, agents can also be requested to predict how their peers are evaluated. The predictions made by an agent $i \in N$ are formally represented by the vector $\mathbf{r}_i = (r_i^1, \ldots, r_i^{i-1}, r_i^{i+1}, \ldots, r_i^n)$, where $\mathbf{r}_i^j = (r_i^{j^1}, \ldots, r_i^{j^M}) \in \Delta^M$, *i.e.*, an element from the unit simplex in $\Re^M$, representing agent i's prediction about the empirical distribution of evaluations received by agent j, *i.e.*, $0 \leq r_i^{j^k} \leq 1$, for $1 \leq k \leq M$, and $\sum_{k=1}^{M} r_i^{j^k} = 1$. The vector $\mathbf{r}_i$ is the result of a private function $\mathbb{F}_i(\mathbf{t}_i)$, indicating that if agent i observes the signal t_i^j, then it will make the prediction $\mathbf{r}_i^j$. Thus, we expect

agents to come up with predictions based on some experience-related reasoning process (*e.g.*, false-consensus [26]).

In this way, we consider two kinds of subjective opinions in this thesis: evaluations and predictions. For avoiding a biased self-judgment, we assume that an agent neither observes a signal about its performance nor predicts its received evaluations. In order to illustrate the notation defined above, consider the following example.

Example 1. *Consider four agents that want to share* \$1000 *using a "five-star" evaluation scheme. Thus, we have* $N = \{1, 2, 3, 4\}$, $V = 1000$, *and* $M = 5$. *Suppose that agent* 1 *observed that agent* 2 *did an excellent job, agent* 3 *was above average, and agent* 4 *did not contribute too much to the group. Thus, agent* 1*'s truthful evaluations can be* $t_1^2 = 5$, $t_1^3 = 4$, $t_1^4 = 1$, *and, consequently,* $\mathbf{t}_1 = (5, 4, 1)$. *Further, assume that agent* 1 *believes that the other agents think that agent* 2 *did not contribute too much to the group. Thus, agent* 1*'s prediction about the truthful evaluations received by agent* 2 *can be* $r_1^{2^1} = r_1^{2^2} = r_1^{2^4} = 0$, $r_1^{2^3} = 0.67$, *and* $r_1^{2^5} = 0.33$, *consequently,* $\mathbf{r}_1^2 = (0, 0, 0.67, 0, 0.33)$.

We make the following assumptions about our model:

Assumption 1 (Independent signals)**.** *The observed signals are independent.*

Assumption 2 (Self-interestedness)**.** *Agents act to maximize their expected shares.*

Assumption 1 means that an agent's truthful evaluation for a peer does not influence its truthful evaluation for another peer. This is a reasonable assumption since agents are expected to come up with subjective opinions based solely on their individual perceptions, and it is intuitively appealing since it takes the autonomy of the individual for granted. Assumption 2 implies that agents are *risk neutral*. This assumption is traditional in both game-theoretic [31] and multiagent systems [42] literature.

A consequence of this last assumption is that agents may deliberately lie when reporting their evaluations and/or predictions. For example, an agent may intentionally give all other agents a low evaluation so that, in comparison, it looks good and receives a greater share of V. Therefore, we distinguish between the truthful evaluations made by every agent $i \in N$, $\mathbf{t}_i$, and the evaluations that it reports, $\mathbf{x}_i = (x_i^1, \ldots, x_i^{i-1}, x_i^{i+1}, \ldots, x_i^n)$. Similarly, we distinguish between the truthful predictions made by every agent $i \in N$, $\mathbf{r}_i$, and the predictions that it reports, $\mathbf{y}_i = (y_i^1, \ldots, y_i^{i-1}, y_i^{i+1}, \ldots, y_i^n)$.

We define the *strategy* of an agent $i \in N$ to be its reported opinions, representing it by $\mathbf{s}_i$. Depending on how opinions are elicited, strategies can be of two kinds: $\mathbf{s}_i = \mathbf{x}_i$, when only evaluations are elicited, and $\mathbf{s}_i = (\mathbf{x}_i, \mathbf{y}_i)$ when both evaluations and predictions are elicited. We overload the notation by always using $\mathbf{s}_i$ to denote the opinions reported by agent i, but we make clear its meaning when necessary. S_i is the set of strategies

available to agent i, and $S = S_1 \times \ldots \times S_n$. Each vector $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n) \in S$ is a *strategy profile*. As customary, let the subscript "$-i$" denote a vector without agent i's component, *e.g.*, $\mathbf{s}_{-i} = (\mathbf{s}_1, \ldots, \mathbf{s}_{i-1}, \mathbf{s}_{i+1}, \ldots, \mathbf{s}_n)$. If the opinions reported by agent i are equal to its truthful opinions, *i.e.*, $\mathbf{x}_i = \mathbf{t}_i$ for evaluations and $\mathbf{y}_i = \mathbf{r}_i$ for predictions, we say that agent i's strategy is *truthful*. We explicitly represent a single truthful opinion or a vector with truthful opinions by using the superscript '$*$', *e.g.*, $\overset{*}{x}_i^j$ and $\overset{*}{\mathbf{s}}_i$. Similarly, we explicitly represent a false opinion or a vector with false opinions by using the superscript '$\hat{}$', *e.g.*, $\hat{x}_i^j$ and $\hat{\mathbf{s}}_i$. We say that a strategy profile is *collectively truthful* if all the reported strategies are truthful.

Opinions are elicited and aggregated by a central, trusted entity, called the *mechanism*, which is also responsible for sharing the reward among agents. This entity relies only on the reported opinions when determining agents' shares, and so it has no additional information. Formally:

Definition 1 (Mechanism)**.** *A mechanism is a pair (S, Γ), where:*

- *$S = S_1 \times \ldots \times S_n$, where S_i is the set of strategies available to agent $i \in N$;*
- *$\Gamma : S \to \Re^n$ is a sharing function that maps each strategy profile to a vector of shares.*

It is important to note that we do not include the parameter M in Definition 1. We assume that its value is common knowledge. Thus, mechanisms differ based on how they share the reward and based on the opinions that they elicit from the agents. We denote the share of V given to agent i, when all the reported opinions are $\mathbf{s}$, by $\Gamma_i(\mathbf{s})$. We use Γ_i when $\mathbf{s}$ is either irrelevant or clear from the context.

Throughout this thesis, we use the solution concepts called *dominant-strategy equilibrium* and *Bayes-Nash equilibrium*.

Definition 2 (Dominant-strategy equilibrium)**.** *We say that $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is a dominant-strategy equilibrium if for each agent $i \in N$, strategy $\mathbf{s}'_i \neq \mathbf{s}_i \in S_i$, and every strategy profile $\sigma_{-\mathbf{i}} \in S_{-i}$, we have that $\Gamma_i(\mathbf{s}_i, \sigma_{-i}) \geq \Gamma_i(\mathbf{s}'_i, \sigma_{-i})$.*

In words, each agent follows a strategy that returns the greatest possible share, no matter how its opponents may play.

Definition 3 (Bayes-Nash equilibrium)**.** *We say that $\sigma = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is a Bayes-Nash equilibrium if for each agent $i \in N$, and strategy $\mathbf{s}'_i \neq \mathbf{s}_i \in S_i$, $\mathbb{E}\left[\Gamma_i(\mathbf{s}_i, \sigma_{-i}) | \overset{*}{\mathbf{s}}_i\right] \geq \mathbb{E}\left[\Gamma_i(\mathbf{s}'_i, \sigma_{-i}) | \overset{*}{\mathbf{s}}_i\right]$.*

In words, for each agent $i \in N$, $\mathbf{s}_i$ is the best response, in an expected sense, that i has to σ_{-i}, given that its truthful strategy is $\overset{*}{\mathbf{s}}_i$. Since the expectation may be taken with respect to different distributions, we elucidate this point when necessary.

When the inequalities in Definition 2 and 3 hold strictly (with "$>$" instead of "$\geq$"), then the strategy profile is classified, respectively, as *strictly dominant-strategy equilibrium* and *strictly Bayes-Nash equilibrium*. Otherwise, if the definitions hold with equality for at least one agent, then the strategy profile is classified, respectively, as *weakly dominant-strategy equilibrium* and *weakly Bayes-Nash equilibrium*.

A simple idea that ensures all possible strategy profiles to be weakly dominant-strategy equilibria is to employ *dictatorial mechanisms*, *i.e.*, mechanisms that throw away agents' opinions and assign shares in a fixed way. This happens because no matter what the agents report, they end up with the same share. An example of a dictatorial mechanism is the *egalitarian mechanism*, *i.e.*, a mechanism that equally shares the reward among the agents. Formally, $\forall i \in N, \forall \mathbf{s} \in S, \Gamma_i(\mathbf{s}) = V/n$. Intuitively, dictatorial mechanisms are unfair because the share assigned to each agent does not reflect the reported opinions about that agent. Formalizing, we define a *fairness* criterion appropriate to our model.

Definition 4 (Fairness)**.** *Consider a strategy profile $\mathbf{s} \in S$ in which the reported evaluation of every agent z for agent i is paired up with agent z's reported evaluation for agent j, for $i \neq j \neq z \in N$, so that $x_z^i > x_z^j$. Further, the evaluations of agent i and agent j for each other are paired up, so that $x_j^i > x_i^j$. Then, we say that a mechanism is fair if $\Gamma_i(\mathbf{s}) > \Gamma_j(\mathbf{s})$.*

In words, if an agent unanimously receives better evaluations than a peer, then this agent should also receive a greater share than that peer.

2.2 Properties

Besides fairness, there are several other key properties we wish the mechanisms to have. In this section, we define such properties.

Definition 5 (Budget Balance)**.** *We say that the mechanism is budget-balanced if $\forall \mathbf{s} \in S, \sum_{i=1}^{n} \Gamma_i(\mathbf{s}) = V$.*

In words, a budget-balanced mechanism allocates the entire reward V back to the agents. As stated, this is a strong definition because we do not put constraints on $\mathbf{s}$, *e.g.*, we do not require $\mathbf{s}$ to be an equilibrium strategy profile. When the mechanism makes a profit, *i.e.*, $\sum_{i=1}^{n} \Gamma_i(\mathbf{s}) < V$, for at least one $\mathbf{s} \in S$, we say that it is *weakly budget-balanced*.

Definition 6 (Individual Rationality)**.** *A mechanism is individually rational if $\forall i \in N, \forall \mathbf{s} \in S, \Gamma_i(\boldsymbol{s}) \geq 0$.*

This condition requires the share received by each agent to be greater than or equal to zero. In other words, all agents are weakly better off participating in the mechanism than not participating at all.

Definition 7 (Incentive Compatibility)**.** *A mechanism is incentive-compatible iff collective truth-telling is an equilibrium strategy profile.*

A mechanism in which a strategy profile $\mathbf{s} \in S$ is both collectively truthful and a dominant-strategy equilibrium is called *strategy-proof*. Intuitively, the best that each agent can do when the mechanism is strategy-proof is to truthfully report its opinions, no matter what the others are reporting. The incentive compatibility concept is weaker when the collectively truthful strategy profile is a Bayes-Nash equilibrium. In this case, it is best, in an expected sense, for each agent to tell the truth provided that the others are also doing so.

Definition 8 (Collusion-Resistance)**.** *A mechanism is collusion-resistant if agents have no incentive to enter into a priori agreements (private contracts) in order to undermine the mechanism.*

In the following section, we extend our discussion on *collusions*. By no means we argue that the properties defined here are exhaustive. However, we believe that they are among the most desirable ones in real applications.

2.3 Collusion

In this thesis, we consider the collusion model in which a single agent can deliberately lie about its evaluation for a specific peer, aiming to increase that peer's share. Because of the self-interestedness assumption, the liar agent looks forward to receiving a side-payment from the beneficiary greater than the expected loss caused by the lie. Formally:

Definition 9 (Collusion)**.** *Given a strategy profile $\mathbf{s} \in S$, a collusion between agents i and j occurs when agent i changes its truthful evaluation for agent j, resulting in the report $\hat{\mathbf{s}}_i \neq \overset{*}{\mathbf{s}}_i$, where $\hat{x}_i^j > \overset{*}{x}_i^j$, and for doing this it receives a side-payment p from agent j so that:*

$$\mathbb{E}\left[\Gamma_i(\hat{\mathbf{s}}_i, \mathbf{s}_{-i}) + p\right] > \mathbb{E}\left[\Gamma_i(\overset{*}{\mathbf{s}}_i, \mathbf{s}_{-i})\right]$$

and

$$\mathbb{E}\left[\Gamma_j(\hat{\mathbf{s}}_i, \mathbf{s}_{-i}) - p\right] > \mathbb{E}\left[\Gamma_j(\overset{*}{\mathbf{s}}_i, \mathbf{s}_{-i})\right]$$

In words, a collusion occurs when, in exchange for misreporting its evaluation, which could lead to a lower share for itself, the liar agent receives a side-payment from the agent who benefits from the misreporting so that both agents end up with a greater expected share than if no collusion had occurred.

While this collusion model may seem narrow, because it is defined considering only two agents, we note that it can be successfully used to model larger collusions provided that they can be decomposed into a union of independent collusions between two agents (a liar and a beneficiary). Thus, this model discards more complex collusions, *e.g.*, when an agent lies about its opinions for a group of peers considering only the joint side-payments. However, we believe that the model is still useful, given the complexity that might arise in coordinating a large group of colluders [19].

We say that a mechanism is *collusion-resistant* when, for all pair of agents i, j, and strategies $\hat{\mathbf{s}}_i \neq \overset{*}{\mathbf{s}}_i \in S_i$, such that $\hat{x}_i^j > \overset{*}{x}_i^j$, the following inequality holds:

$$\mathbb{E}\left[\Gamma_i(\hat{\mathbf{s}}_i, \mathbf{s}_{-i}) + \Gamma_j(\hat{\mathbf{s}}_i, \mathbf{s}_{-i})\right] \leq \mathbb{E}\left[\Gamma_i(\overset{*}{\mathbf{s}}_i, \mathbf{s}_{-i}) + \Gamma_j(\overset{*}{\mathbf{s}}_i, \mathbf{s}_{-i})\right] \tag{2.1}$$

A point useful to discuss is the relationship between our collusion model and the *group strategy-proofness* concept from the mechanism design literature (*e.g.*, [29]). In our setting, a mechanism is group strategy-proof if it is always in the best interest of all subsets of agents to reveal their opinions truthfully. While both concepts deal with a group of colluders, group strategy-proofness is stronger in a sense that it prevents any coalition of agents to gain by lying, but weaker because it does not explicitly capture the idea of side-payments. Further, group strategy-proofness is not well-defined in Bayesian settings, which would severely limit our analysis on the collusion-resistance aspect of the proposed mechanisms.

2.4 Scoring Rules

Later in this thesis, a concept used by a mechanism to incentivise truthfulness is called *scoring rules* [47]. Consider an uncertain quantity with possible outcomes $o_1, \ldots, o_z$ and a probability vector $\mathbf{p} = (p_1, \ldots, p_z)$. A scoring rule $R(\mathbf{p}, e)$ is a function that provides a score for the assessment $\mathbf{p}$ upon observing the event o_e.

A scoring rule is called *strictly proper* when an agent receives its maximal expected score if and only if its stated assessment $\mathbf{p}$ corresponds to its true assessment $\mathbf{q} = (q_1, \ldots, q_z)$ [47, 38]. The *expected score* of $\mathbf{p}$ at $\mathbf{q}$ for a real value scoring rule $R(\mathbf{p}, e)$ is:

$$C(\mathbf{p}|\mathbf{q}) = \sum_{e=1}^{z} q_e \, R(\mathbf{p}, e), \tag{2.2}$$

and the *expected score loss* is defined by the equation:

$$L(\mathbf{p}|\mathbf{q}) = C(\mathbf{q}|\mathbf{q}) - C(\mathbf{p}|\mathbf{q}). \tag{2.3}$$

The literature contains a number of strictly proper scoring rules. The best known and their scoring ranges are [39]:

logarithmic:	$R(\mathbf{p}, i) = \log p_i$	$(-\infty, 0]$
quadratic:	$R(\mathbf{p}, i) = 2p_i - \sum_{e=1}^{z} p_e^2$	$[-1, 1]$
spherical:	$R(\mathbf{p}, i) = \frac{p_i}{(\sum_{e=1}^{z} p_e^2)^{1/2}}$	$[0, 1]$

For proving an equilibrium result later in this thesis, we use the following properties of strictly proper scoring rules.

Lemma 1. *If $R(\phi, e)$ is a strictly proper scoring rule, then a positive affine transformation of R, i.e., $\alpha R(\phi, e) + \beta$, for $\alpha > 0$ and $\beta \in \Re$, is also strictly proper.*

Proof. Given that $\mathbf{p}$ is the assessment that maximizes the expected score in $\alpha R(\phi, e) + \beta$, and $\mathbf{q}$ is the assessment that maximizes the expected score in $R(\phi, e)$, *i.e.*, the true assessment, then we have that:

$$\begin{aligned} \mathbf{p} &= \arg\max_{\phi} \sum_{e=1}^{z} (\alpha q_e \, R(\phi, e) + \beta) \\ &= \arg\max_{\phi} \sum_{e=1}^{z} q_e \, R(\phi, e) \\ &= \mathbf{q} \end{aligned}$$

where the second equality follows from the facts that α and β are constants, and $\alpha > 0$. □

Lemma 2. *Let $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_w)$ be a vector of independent assessments, where $\mathbf{d}_j = (p_1, \ldots, p_z)$, for $1 \leq j \leq w$, is a probability distribution over the outcomes $o_1, \ldots, o_z$. Also, let $\mathbf{E} = (e_1, \ldots, e_w)$ be a vector of independently observed events. Consider the scoring function $H(\mathbf{D}, \mathbf{E}) = \frac{1}{w} \sum_{j=1}^{w} R_j(\mathbf{d}_j, e_j)$, where $R_1, \ldots, R_w$ are strictly proper scoring rules. Then, $H(\mathbf{D}, \mathbf{E})$ is also strictly proper.*

Proof. Consider the vector of assessments that maximizes the expected score in $H(\mathbf{D}, \mathbf{E})$:

$$\begin{aligned}\underset{(\mathbf{d}_1,\ldots,\mathbf{d}_w)}{\arg\max}\,\mathbb{E}\left[\frac{1}{w}\sum_{j=1}^{w} R_j(\mathbf{d}_j, e_j)\right] &= \underset{(\mathbf{d}_1,\ldots,\mathbf{d}_w)}{\arg\max}\,\mathbb{E}\left[\sum_{j=1}^{w} R_j(\mathbf{d}_j, e_j)\right] \\ &= (\underset{\mathbf{d}_1}{\arg\max}\,\mathbb{E}\left[R_1(\mathbf{d}_1, e_1)\right], \ldots, \underset{\mathbf{d}_w}{\arg\max}\,\mathbb{E}\left[R_w(\mathbf{d}_w, e_w)\right]).\end{aligned}$$

This last equality follows from the fact that the scores given for each assessment are independent among themselves. Then, given that $R_1, \ldots, R_w$ are strictly proper scoring rules, we can conclude that the vector $\mathbf{D}$ that maximizes the expected value of $H(\mathbf{D}, \mathbf{E})$ is composed by true assessments. Thus, the scoring function $H(\mathbf{D}, \mathbf{E})$ is strictly proper. □

Throughout this thesis, we use the following strictly proper scoring rule:

$$R(\mathbf{p}, e) = 1 + 2p_e - \sum_j p_j^2. \tag{2.4}$$

This scoring rule is a positive affine transformation of the quadratic scoring rule, and its scoring range is [0, 2]. Selten [39] shows the proof that the quadratic scoring rule is indeed strictly proper, and some of its interesting properties.

2.5 The Bayesian Truth Serum Method

Another method used later in this thesis to incentivise truthfulness is the *Bayesian Truth Serum* method (BTS) [32]. Working on a single multiple-choice question with a finite number of alternatives, this method requires each agent to endorse the answer most likely to be true and to predict the empirical distribution of the endorsed answers.

Agents are evaluated by the accuracy of their predictions (how well they matched the empirical frequencies) as well as how *surprisingly common* are their personal answers in relation to the empirical predictions, *e.g.*, an answer endorsed by 50% of the population against a predicted frequency of 25% is surprisingly common and should receive a high score. Alternatively, this answer would be surprisingly uncommon if predictions averaged 75% and, consequently, it would receive a lower score. This relatively simple scoring criterion removes all the necessity of biasing answers toward the likely consensus.

To have a concrete problem in mind, suppose that a question is asking for evaluations for an agent $j \in N$. Using the notation previously defined, let $h(x_i^j, k)$ be a zero-one indicator function, *i.e.*,

$$h(x_i^j, k) = \begin{cases} 1 & \text{if } x_i^j = k \\ 0 & \text{otherwise} \end{cases}$$

The score provided by the BTS method to agent i, given its reported answer (evaluation) x_i^j and prediction $\mathbf{y}_i^j$, is calculated as follows:

$$\sum_{k=1}^{M} h(x_i^j, k) \log \frac{\bar{x}_k}{\bar{y}_k} + \beta \sum_{k=1}^{M} \bar{x}_k \log \frac{y_i^{j^k}}{\bar{x}_k}, \tag{2.5}$$

where $\bar{x}_k$ is the average frequency of the evaluation k, and $\bar{y}_k$ is the geometric average of the predicted frequencies for the evaluation k,

$$\bar{x}_k = \frac{1}{n-1} \sum_{q \neq j} h(x_q^j, k),$$

$$\bar{y}_k = \exp\left(\frac{1}{n-1} \sum_{q \neq j} \log y_q^{j^k} \right).$$

Since agent j does not evaluate itself, the denominators of the above averages are equal to $n-1$. The BTS method has two major components. The first part, called the *information score*, selects the evaluation endorsed by an agent i, and multiplies it by the log-ratio of the actual-to-predicted frequency of evaluations. Here, an evaluation scores high to the extent that it is more common than collectively predicted. The second part, called the *prediction score*, is a penalty proportional to the relative entropy (or the Kullback-Leibler divergence [9]) between the empirical distribution of evaluations and agent i's prediction about that distribution. The constant $\beta > 0$ fine-tunes the weight given to the prediction score. In this thesis, we are always considering $\beta = 1$. There are four assumptions required for the perfect operation of the BTS method:

Assumption 3 (Common prior)**.** *There exists a common prior over the answers of the members of the population. In our model, this means that for each agent $j \in N$, there exists a common prior $p(\omega_j)$ over the truthful evaluations for agent j.*

Assumption 4 (Rationality)**.** *Every agent forms a posterior over the population distribution of answers. In our model, this means that every agent $i \in N$, with truthful evaluation t_i^j, forms a posterior by applying Bayes' rule to the common prior $p(\omega_j)$, i.e., $p(\omega_j|t_i^j)$.*

Assumption 5 (Stochastic relevance)**.** *Different answers imply different posterior distributions. In our model, this means that $\forall i, q$, $p(\omega_j|t_i^j) = p(\omega_j|t_q^j)$ if and only if $t_i^j = t_q^j$.*

Assumption 6 (Large population)**.** *The population of agents must be sufficiently large so that a single answer cannot significantly affect the empirical distribution of answers.*

Under these four assumptions, and using Equation 2.5 to compute agents' scores, the following theorems hold [32]:

Theorem 1. *Collective truth-telling is a strictly Bayes-Nash equilibrium.*

Theorem 2. *The expected information score in any Bayes-Nash equilibrium is non-negative.*

Theorem 3. *The expected information score is (weakly) greater in the collectively truthful strategy profile than in any other Bayes-Nash equilibrium.*

Theorem 4. *If the constant $\beta = 1$, then the Bayesian truth serum method is zero-sum.*

Theorem 1 says that the best response for an agent, in an expected sense, when everyone else is telling the truth is also to tell the truth. Theorem 2 and 3 mean that the expected value of the left part of Equation 2.5 is greater than or equal to zero in any Bayes-Nash equilibrium, and that it is weakly maximized in the collectively truthful strategy profile than in any other Bayes-Nash equilibrium. Theorem 4 states that if we set $\beta = 1$, then the sum of the scores received by the responders is equal to zero. In these theorems, the expectations are taken with respect to the posterior distribution.

For proving an equilibrium result later in this thesis, we use the following lemmas related to the BTS method. In what follows, consider $n-1$ independent questions, where each question asks for an evaluation for a specific agent different than agent i. Let $\sigma_i^j = (x_i^j, \mathbf{y}_i^j)$, *i.e.*, a vector with both the answer given by agent i to the question about agent j's evaluation and the prediction made by agent i about the empirical distribution of answers to the same question. Also, let $\sigma^j_{-\{ij\}}$ be a vector with all agents' answers and predictions to that question, except those from agent i and j. Finally, consider that a score for agent i is computed by using the following scoring scheme:

$$g_i = \sum_{j \neq i} \mathbb{G}\left(\sigma_i^j, \sigma^j_{-\{i,j\}}\right), \tag{2.6}$$

where $\mathbb{G}(\cdot, \cdot)$ is computed using the BTS method (Equation 2.5).

Lemma 3. *The scoring scheme g_i is incentive-compatible.*

Proof. Suppose that every peer of agent i always reports its answers and predictions truthfully. Consider the vector $\sigma_i = (\sigma_i^1, \ldots, \sigma_i^{i-1}, \sigma_i^{i+1}, \ldots, \sigma_i^n)$, with the answers and predictions that maximize the expected score of agent i, *i.e.*,

$$
\begin{aligned}
\sigma_i &= \underset{(\sigma_i^1,\ldots,\sigma_i^{i-1},\sigma_i^{i+1},\ldots,\sigma_i^n)}{\arg\max} \left(\mathbb{E}\left[\sum_{j \neq i} \mathbb{G}\left(\sigma_i^j, \sigma^j_{-\{i,j\}} \right) \right] \right) \\
&= \left(\arg\max_{\sigma_i^1} \left(\mathbb{E}\left[\mathbb{G}\left(\sigma_i^1, \sigma^1_{-\{i,1\}} \right) \right] \right), \ldots, \arg\max_{\sigma_i^{i-1}} \left(\mathbb{E}\left[\mathbb{G}\left(\sigma_i^{i-1}, \sigma^{i-1}_{-\{i,i-1\}} \right) \right] \right), \right. \\
&\quad \left. \arg\max_{\sigma_i^{i+1}} \left(\mathbb{E}\left[\mathbb{G}\left(\sigma_i^{i+1}, \sigma^{i+1}_{-\{i,i+1\}} \right) \right] \right), \ldots, \arg\max_{\sigma_i^{n}} \left(\mathbb{E}\left[\mathbb{G}\left(\sigma_i^{n}, \sigma^{n}_{-\{i,n\}} \right) \right] \right) \right).
\end{aligned}
$$

The second equality follows from the fact that the $n-1$ questions are independent. According to Theorem 1, each expectation inside this last vector is strictly maximized when agent i tells the truth, because, by assumption, everyone else is telling the truth. Thus, collective truth-telling is a strictly Bayes-Nash equilibrium, and the scoring scheme g_i is incentive-compatible. □

Lemma 4. *A positive affine transformation of the scoring scheme g_i is incentive-compatible.*

Proof. Suppose that every peer of agent i always reports its answers and predictions truthfully. Consider the vector $\sigma_i = (\sigma_i^1, \ldots, \sigma_i^{i-1}, \sigma_i^{i+1}, \ldots, \sigma_i^n)$, with the answers and predictions that maximize the expected score received by agent i from the scoring scheme $g_i' = \kappa\, g_i + \lambda$, for $\kappa > 0$ and $\lambda \in \Re$, *i.e.*,:

$$
\begin{aligned}
\sigma_i &= \underset{(\sigma_i^1,\ldots,\sigma_i^{i-1},\sigma_i^{i+1},\ldots,\sigma_i^n)}{\arg\max} \left(\mathbb{E}\left[\kappa \left(\sum_{j \neq i} \mathbb{G}\left(\sigma_i^j, \sigma^j_{-\{i,j\}} \right) \right) + \lambda \right] \right) \\
&= \underset{(\sigma_i^1,\ldots,\sigma_i^{i-1},\sigma_i^{i+1},\ldots,\sigma_i^n)}{\arg\max} \left(\kappa \mathbb{E}\left[\sum_{j \neq i} \mathbb{G}\left(\sigma_i^j, \sigma^j_{-\{i,j\}} \right) \right] + \lambda \right) \\
&= \underset{(\sigma_i^1,\ldots,\sigma_i^{i-1},\sigma_i^{i+1},\ldots,\sigma_i^n)}{\arg\max} \left(\mathbb{E}\left[\sum_{j \neq i} \mathbb{G}\left(\sigma_i^j, \sigma^j_{-\{i,j\}} \right) \right] \right).
\end{aligned}
$$

The third equality follows from the facts that κ and λ are constant, and that $\kappa > 0$. Thus, we reduce this lemma to Lemma 3, completing the proof. □

Chapter 3

The Peer-Evaluation Mechanism

In this chapter, we present the *peer-evaluation mechanism*, our first mechanism for sharing rewards using subjective opinions. It works by taking the stand-alone evaluations received by the agents and dividing the reward proportionally to them. We show that this mechanism is fair, budget-balanced, individually rational, and strategy-proof, but that it can be collusion-prone.

3.1 The Mechanism

The mechanism starts by requesting evaluations from the agents, *i.e.*, $\forall i \in N, \mathbf{s}_i = \mathbf{x}_i$. For each vector with evaluations, $\mathbf{x}_i$, the mechanism creates a second vector, $\chi_i = (\chi_i^1, \ldots, \chi_i^{i-1}, \chi_i^{i+1}, \ldots, \chi_i^n)$, by scaling the elements of the first one to sum up to V. Mathematically,

$$\forall i, j,\ \chi_i^j = x_i^j \left(\frac{V}{\sum_{q \neq i} x_i^q} \right).$$

This simple adjustment in the agents' evaluations ensures that the sum of the final shares is not orders of magnitude lower than the reward V. The mechanism aggregates the evaluations received by each agent $i \in N$ by summing the scaled evaluations received by it. The share of each agent $i \in N$ is then equal to this aggregated value divided by n, *i.e.*,

$$\Gamma_i = \frac{\sum_{j \neq i} \chi_j^i}{n}. \tag{3.1}$$

The intuition behind the peer-evaluation mechanism is that the share received by each agent is directly proportional to its received evaluations. It is important to note that the peer-evaluation mechanism computes an agent's share by summing the scaled evaluations received by that agent and dividing the result by n, differently from the arithmetic mean that would use $n-1$ in the denominator. This is useful to ensure important properties for the mechanism. Algorithm 1 presents the sharing function of the peer-evaluation mechanism from an algorithmic perspective.

Algorithm 1 The Peer-Evaluation Mechanism

1: **for** $i = 1$ to n **do**
2: **for** $j \neq i$ **do**
3: $\chi_j^i = x_j^i \left(\frac{V}{\sum_{q \neq j} x_j^q}\right)$
4: **end for**
5: $\Gamma_i = \frac{\sum_{j \neq i} \chi_j^i}{n}$
6: **end for**

To illustrate the peer-evaluation mechanism, consider the following example:

Example 2. *Suppose that four agents A, B, C, and D, want to share the reward $V = 1000$ using the peer-evaluation mechanism. Let $M = 10$, and the reported evaluations shown in Table 3.1. Each numeric cell beneath the label "Evaluation" can be interpreted as the evaluation given by the agent in the row to the agent in the column, e.g, $x_B^A = 7$. Each cell beneath the label "Share" represents the resulting share of the agent in the row.*

Table 3.1: Example of the peer-evaluation mechanism.

	Evaluation				Share
	A	B	C	D	
A	-	5	3	3	231.15
B	7	-	6	5	345.78
C	2	3	-	2	214.02
D	1	2	1	-	209.05

For illustration, the share received by agent A is:

$$\begin{aligned}
\Gamma_A &= \frac{x_B^A \left(\frac{V}{x_B^A + x_B^C + x_B^D}\right) + x_C^A \left(\frac{V}{x_C^A + x_C^B + x_C^D}\right) + x_D^A \left(\frac{V}{x_D^A + x_D^B + x_D^C}\right)}{n} \\
&= \frac{7\left(\frac{1000}{7+6+5}\right) + 2\left(\frac{1000}{2+3+2}\right) + 1\left(\frac{1000}{1+2+1}\right)}{4} \\
&\approx 231.15.
\end{aligned}$$

3.2 Properties

This simple, yet powerful, mechanism has very interesting properties. First, because the reported evaluations are always greater than zero, an agent cannot receive a negative share. Consequently, the peer-evaluation mechanism is individually rational. Since the share received by an agent does not depend on its strategy, the mechanism is also non-consensual, *i.e.*, an agent will not increase its share by following a group consensus. Further, the mechanism is budget-balanced, as proved in the following proposition.

Proposition 1. *The peer-evaluation mechanism is budget-balanced.*

Proof. The sum of the shares received by the agents is:

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{\sum_{j\neq i}\chi_j^i}{n}\right) &= \sum_{j=1}^{n}\left(\frac{\sum_{i\neq j}\chi_j^i}{n}\right)\\ &= n\left(\frac{V}{n}\right)\\ &= V.\end{aligned}$$

where the second equality follows from the fact that the scaled evaluations sum up to V. □

The following theorems state our main results concerning the properties of the peer-evaluation mechanism.

Theorem 5. *The peer-evaluation mechanism is strategy-proof.*

Proof. The share received by an agent does not depend on its reported evaluations. Consequently, an agent $i \in N$ cannot improve its own share by reporting a vector of strategies $\hat{\mathbf{s}}_i \neq \overset{*}{\mathbf{s}}_i$. Then, the collectively truthful strategy profile is trivially a weakly dominant-strategy equilibrium. □

Theorem 6. *If $M < \sqrt{n-1}$, then the peer-evaluation mechanism is fair.*

Proof. Consider a pair of different agents $i, j \in N$ and a strategy profile $\mathbf{s} \in S$ where $x_j^i > x_i^j$ and, for every other agent $z \neq i, j$, $x_z^i > x_z^j$. The mechanism must satisfy the following inequality to be considered fair:

$$\begin{aligned} \Gamma_i &> \Gamma_j \Rightarrow \\ \frac{\sum_{z\neq i,j} x_z^i \left(\frac{V}{\sum_{q\neq z} x_z^q}\right) + x_j^i \left(\frac{V}{\sum_{q\neq j} x_j^q}\right)}{n} &> \frac{\sum_{z\neq i,j} x_z^j \left(\frac{V}{\sum_{q\neq z} x_z^q}\right) + x_i^j \left(\frac{V}{\sum_{q\neq i} x_i^q}\right)}{n}. \end{aligned}$$

After doing some algebraic manipulations we have:

$$\sum_{z\neq i,j} \left(\frac{x_z^i - x_z^j}{\sum_{q\neq z} x_z^q}\right) + \frac{x_j^i}{\sum_{q\neq j} x_j^q} - \frac{x_i^j}{\sum_{q\neq i} x_i^q} > 0.$$

In what follows, we restrict ourselves to the worst-case scenario. Since $\forall i, j$, $x_i^j \in \{1, \ldots, M\}$, and $\forall z \neq i, j$, $x_z^i > x_z^j$, we have:

$$\frac{n-2}{(n-1)M} + \frac{1}{(n-1)M} - \frac{M}{(n-1)} > 0.$$

The result follows after simple algebraic manipulations. □

Related to the collusion-resistance property, the peer-evaluation mechanism is not collusion-resistant. To illustrate this point, consider the following example:

Example 3. *Consider the same scenario of Example 3, and assume that the evaluations in Table 3.1 are truthful. As can be seen from Table 3.1, agents C and D receive the smallest shares among the four agents when they report their truthful evaluations. Now, instead of telling the truth, suppose that agents C and D make an a priori agreement. In detail, assume that agent C agrees to increase its evaluation for agent D from $x_C^D = 2$ to $x_C^D = 10$ in exchange for a side-payment of* \$50.00*. The resulting evaluations and shares, after discounting the side-payment, can be seen in Table 3.2. After colluding, agents C and D are able to increase their shares by, respectively,* \$50.00 *and* \$45.24*.*

Table 3.2: Shares after a collusion between agents C and D.

Evaluation					Share
	A	B	C	D	
A	-	5	3	3	193.06
B	7	-	6	5	288.64
C	2	3	-	10	264.02
D	1	2	1	-	254.29

In the following proposition, we provide a bound on the maximum value that an agent can receive due exclusively to a collusive behavior.

Proposition 2. *The maximum value that an agent can receive due to a collusive behavior is* $\frac{V(M-1)}{(n-1)n}$.

Proof. Consider that agent i is lying to increase the share received by agent j by reporting $\hat{x}_i^j > \overset{*}{x}_i^j$. The maximum value that agent j can receive due exclusively to this collusive behavior is:

$$\begin{aligned}
\frac{\hat{x}_i^j\left(\frac{V}{(\sum_{z\neq i,j} x_i^z)+\hat{x}_i^j}\right) - \overset{*}{x}_i^j\left(\frac{V}{(\sum_{z\neq i,j} x_i^z)+\overset{*}{x}_i^j}\right)}{n} &< \frac{\hat{x}_i^j\left(\frac{V}{(\sum_{z\neq i,j} x_i^z)+\overset{*}{x}_i^j}\right) - \overset{*}{x}_i^j\left(\frac{V}{(\sum_{z\neq i,j} x_i^z)+\overset{*}{x}_i^j}\right)}{n} \\
&= \frac{V\left(\frac{\hat{x}_i^j-\overset{*}{x}_i^j}{(\sum_{z\neq i,j} x_i^z)+\overset{*}{x}_i^j}\right)}{n} \\
&\leq \frac{V(M-1)}{(n-1)n}
\end{aligned}$$

The first inequality follows from the fact that $\hat{x}_i^j > \overset{*}{x}_i^j$. The second inequality follows from the fact that $\forall i, j,\ x_i^j \in \{1, \ldots, M\}$.

□

3.3 Concluding Remarks

In this chapter, we presented our first mechanism for sharing rewards using subjective opinions. The share received by each agent from the peer-evaluation mechanism is directly proportional to its received evaluations. Arguably, the most salient property of this mechanism is its simplicity. In actual applications, the mathematics behind the sharing function can be easily taught to the agents. We showed that the peer-evaluation mechanism is budget-balanced, individually rational, and strategy-proof. Further, if $M < \sqrt{n-1}$, then it is also guaranteed to be fair.

The main drawback of the peer-evaluation mechanism is that, although agents do not have direct incentives for misreporting evaluations, they also do not have strong incentives for telling the truth. This makes the mechanism extremely susceptible to collusions. Example 3 illustrated that, although agents are unable to unilaterally improve their shares, by colluding with other agents in the group, they can undermine the peer-evaluation mechanism and significantly improve their shares.

Chapter 4

The Peer-Prediction Mechanism

In this chapter, we propose our second mechanism for sharing rewards based on subjective opinions, the *peer-prediction mechanism*. It encourages truthfulness by using scoring rules. The share received by each agent from this mechanism has two major components. The first one is the average of the received evaluations, and the second one is a truth-telling score. These scores are computed by considering each evaluation submitted by an agent as a bet on the average of the others' evaluations. Under mild assumptions, we show that the peer-prediction mechanism is weakly budget-balanced, individually rational, and incentive-compatible, and we discuss a trade-off that exists between fairness and truthfulness.

4.1 The Mechanism

The mechanism starts by requesting evaluations from agents, *i.e.*, $\forall i \in N, \mathbf{s}_i = \mathbf{x}_i$. Besides the basic assumptions of the model, *i.e.*, independent signals and self-interestedness (see Section 2.1), we make the following extra assumptions:

Assumption 7. *For each agent $j \in N$, there exists a common, Dirichlet prior $p(\omega_j)$ over the truthful evaluations for agent j, and this prior is common knowledge.*

Assumption 8. *Every agent $i \in N$, with truthful evaluation t_i^j, forms a posterior by applying Bayes' rule to the Dirichlet prior $p(\omega_j)$, i.e., $p(\omega_j|t_i^j)$.*

A Dirichlet distribution $\mathbb{D}(\theta, \Theta) \propto \prod_k \theta_k^{\Theta_k - 1}$ over a simplex (multinomial) θ is parameterized by positive numbers Θ_k such that $\Theta_k - 1$ can be interpreted as the number of times that the θ_k-probability event has been observed [18, 14]. Here, we assume that $\Theta_k = 1$, for $1 \leq k \leq M$. Intuitively, this means that the initial belief of every agent about the truthful evaluations for a peer is uninformative, *i.e.*, that the expected distribution, $\mathbb{E}[\omega_j]$,

is uniform over the set $\{1, \ldots, M\}$. Assumption 8 means that, after observing the signal from a peer, the belief of each agent $i \in N$ is updated so that the *posterior predictive distribution* will be:

$$\mathbb{E}\left[\omega_j = k | t_i^j\right] = \begin{cases} \frac{2}{M+1} & \text{if } t_i^j = k \\ \frac{1}{M+1} & \text{otherwise.} \end{cases}$$

Implicitly, this assumption means that each agent's relevant information consists exclusively of its observed signals. Consequently, the updated belief of each agent indicates that the observed signal from a peer is the evaluation most likely to be deserved by that peer. We highlight the reasons for these assumptions throughout this chapter.

The share received by each agent $i \in N$ from the peer-prediction mechanism has two major components. The first one, μ_i, is the average of the evaluations received by agent i:

$$\mu_i = \frac{\sum_{j \neq i} x_j^i}{n-1}. \tag{4.1}$$

The second major component of agent i's share is a truth-telling score. The key idea is that such scores are maximized, in an expected sense, when the agents truthfully report their evaluations. To calculate the scores, let $\Phi_i^j = (\Phi_i^{j^1}, \ldots, \Phi_i^{j^M})$ be the element from the unit simplex in $\Re^M$ where:

$$\Phi_i^{j^k} = \begin{cases} \frac{2}{M+1} & \text{if } x_i^j = k \\ \frac{1}{M+1} & \text{otherwise} \end{cases}$$

for $1 \leq k \leq M$. When computing Φ_i^j, the peer-prediction mechanism is essentially estimating agent i's posterior predictive distribution about the truthful evaluations for agent j from agent i's reported evaluation, x_i^j. The score of agent i is then:

$$\zeta_i = \frac{\sum_{j \neq i} R\left(\Phi_i^j, \, nint\left(\frac{\sum_{z \neq i,j} x_z^j}{n-2}\right)\right)}{n-1}, \tag{4.2}$$

where R is the strictly proper scoring rule in Equation 2.4, and *nint* is the nearest integer function. Thus, agent i's score is the arithmetic mean of the results provided by that scoring rule, where these results are obtained by using agent i's estimated posterior predictive distributions as assessments. Scoring rules require an outcome, or a "reality", to score an assessment. If the mechanism knew *a priori* each agent's truthful opinions, it could compare them to the reported ones and reward agreement. However, due to the subjective nature of the opinions, we are facing a situation where the objective truth is unknowable.

Our solution to this issue is to score Φ_i^j against the average evaluation received by agent j, disregarding agent i's opinion from this last value. Since $\forall i, j, x_i^j \in \{1, \ldots, M\}$, the function *nint* in Equation 4.2 rounds the average evaluation to an integer number inside the set $\{1, \ldots, M\}$. Intuitively, we are treating the evaluations submitted by an agent as bets on the average of the others' evaluations.

Finally, we linearly combine the average evaluation received by agent i, μ_i, and its truth-telling score, ζ_i:

$$\mu_i + \alpha\,\zeta_i,$$

where the constant $\alpha > 0$ fine-tunes the weight given to the truth-telling score ζ_i. We observe that $\forall i \in N$, μ_i and ζ_i are, respectively, in the ranges [1, M] and [0, 2], because $\forall i, j$, $x_i^j \in \{1, \ldots, M\}$, and a result of Equation 2.4 is in the range [0, 2]. Consequently, the above value can be orders of magnitude lower than V. To overcome this problem, we multiply the above value by the constant $\frac{V}{(M+2\alpha)n}$. The reason for using this value will be clear in the next section. Thus, the share of agent i returned by the peer-prediction mechanism is:

$$\Gamma_i = (\mu_i + \alpha\,\zeta_i)\frac{V}{(M+2\alpha)n} \tag{4.3}$$

The intuition behind this sharing function is that agents receive greater shares when they are well-evaluated and when they tell the truth. Algorithm 2 presents this sharing function from an algorithmic perspective.

Algorithm 2 The Peer-Prediction Mechanism

1: **for** $i = 1$ to n **do**
2: $\quad \mu_i = \frac{\sum_{j \neq i} x_j^i}{n-1}$
3: $\quad \zeta_i = \frac{\sum_{j \neq i} R\left(\Phi_i^j, nint\left(\frac{\sum_{z \neq i,j} x_z^j}{n-2}\right)\right)}{n-1}$
4: $\quad \Gamma_i = (\mu_i + \alpha\,\zeta_i)\frac{V}{(M+2\alpha)n}$
5: **end for**

To illustrate the peer-prediction mechanism, consider the following example:

Example 4. *Suppose that four agents A, B, C, and D, want to share the reward* $V = 1000$ *using the peer-prediction mechanism. Let* $M = 5$, $\alpha = 0.5$, *and the reported evaluations shown in Table 4.1. Each numeric cell beneath the label "Evaluation" can be interpreted as the evaluation given by the agent in the row to the agent in the column, e.g,* $x_B^A = 5$.

Table 4.1: Example of the peer-prediction mechanism: reported evaluations.

Evaluation				
	A	B	C	D
A	-	4	2	1
B	5	-	5	5
C	3	5	-	1
D	3	5	2	-

Using these evaluations, the mechanism returns the shares presented in the last column of Table 4.2. For illustration, consider the share received by agent D. The first component of Γ_D *is the arithmetic mean of the evaluations received by agent D, which is:*

$$\begin{aligned} \mu_D &= \frac{1+5+1}{3} \\ &\approx 2.33. \end{aligned}$$

The second component of Γ_D *is the arithmetic mean of the results provided by the scoring rule in Equation 2.4, where each result is related to an evaluation submitted by agent D:*

$$\begin{aligned} \zeta_D &= \frac{R\left(\Phi_D^A, nint\left(\frac{x_B^A+x_C^A}{2}\right)\right) + R\left(\Phi_D^B, nint\left(\frac{x_A^B+x_C^B}{2}\right)\right) + R\left(\Phi_D^C, nint\left(\frac{x_A^C+x_B^C}{2}\right)\right)}{3} \\ &\approx \frac{1.11+1.44+1.11}{3} \\ &\approx 1.22 \end{aligned}$$

Finally, the share of agent D is a linear combination of μ_D *and* ζ_D*, times the constant* $\frac{V}{(M+2\alpha)n}$*,*

$$\begin{aligned} \Gamma_D &= (\mu_D + \alpha\zeta_D)\frac{V}{(M+2\alpha)n} \\ &\approx (2.33 + 0.5 \times 1.22)\frac{1000}{(5+2\times 0.5)4} \\ &\approx 122.50. \end{aligned}$$

Table 4.2: Example of the peer-prediction mechanism: resulting shares.

	μ_i	ζ_i	Γ_i
A	3.67	1.11	175.93
B	4.67	1.11	217.59
C	3.00	1.22	150.46
D	2.33	1.22	122.50

4.2 Properties

In this section, we show the properties of the peer-prediction mechanism. We start by observing that $\forall i \in N$, μ_i and ζ_i are, respectively, in the ranges [1, M] and [0, 2], because $\forall i, j$, $x_i^j \in \{1, \ldots, M\}$, and a result of Equation 2.4 is always in the range [0, 2]. Since $\alpha > 0$, the agents' shares are greater than zero, and, consequently, the peer-prediction mechanism is individually rational.

Proposition 3. *The peer-prediction mechanism is weakly budget-balanced.*

Proof. Mathematically, this proposition says:

$$\begin{aligned} V &\leq \sum_{i=1}^{n} \Gamma_i \\ &= \sum_{i=1}^{n} (\mu_i + \alpha\, \zeta_i)\, \frac{V}{(M+2\alpha)n} \end{aligned}$$

Since $\forall i \in N$, μ_i and ζ_i are, respectively, in the ranges [1, M] and [0, 2], we have:

$$\frac{V}{(M+2\alpha)} \leq \sum_{i=1}^{n} \Gamma_i \leq V, \tag{4.4}$$

completing the proof. □

We note that Equation 4.4 provides a bound on the profit that the mechanism can make, *i.e.*, a value at most $V - \frac{V}{(M+2\alpha)}$. By using the constant $\frac{V}{(M+2\alpha)n}$ to scale agents' shares, we guarantee that the mechanism never takes a loss. Now, we are ready to show our main results concerning the properties of the peer-prediction mechanism.

Theorem 7. *The peer-prediction mechanism is incentive-compatible.*

Proof. Suppose that every peer of an agent $i \in N$ truthfully reports its evaluations. We prove that the strict best response for agent i, in an expected sense, is also to tell the truth. We start by observing that the share received by agent i (Equation 4.3) can be written as:

$$\begin{aligned}\Gamma_i &= \mu_i \frac{V}{(M+2\alpha)n} + \alpha \frac{V}{(M+2\alpha)n}\zeta_i \\ &= C_1 + C_2\zeta_i\end{aligned}$$

where C_1 and C_2 are positive constants, from agent i's point of view, because they do not depend on the evaluations reported by agent i. Focusing on agent i's score, ζ_i, from Lemma 2 (Section 2.4) we know that the scoring function,

$$\zeta_i = \frac{\sum_{j \neq i} R\left(\Phi_i^j,\, nint\left(\frac{\sum_{z \neq i,j} x_z^j}{n-2}\right)\right)}{n-1}$$

is strictly proper, where the expectation is taken with respect to agent i's posterior predictive distributions. Consequently, it is maximized when Φ_i^j is equal to $\mathbb{E}\left[p(\omega_j|t_i^j)\right]$, *i.e.*, when $x_i^j = t_i^j$. Finally, we conclude by observing that agent i's share, Γ_i, can be seen as a positive affine transformation of a strictly proper scoring rule. Then, according to Lemma 1, the expected share of agent i is strictly maximized when it tells the truth. Consequently, the collectively truthful strategy profile is a strictly Bayes-Nash equilibrium, and the peer-prediction mechanism is incentive-compatible. □

Another way to interpret the above result is to imagine that each agent is betting on the average evaluation received by each peer. Since the only information available to an agent are its observed signals, then its best strategy is to bet on these observed signals, *i.e.*, to bet on its truthful evaluations. In the following proposition, we illustrate an approach to avoid collusions by fine-tuning the weight given to the truth-telling scores.

Proposition 4. *If* $\alpha \geq \frac{(M-1)(M+1)^2}{2}$*, then the peer-prediction mechanism is collusion-resistant.*

Proof. For this proof, consider the following notation. Let $\overset{*}{\zeta}_i$ and $\overset{*}{\mu}_j$ be, respectively, agent i's score and the average evaluation for agent j when agent i reports its truthful evaluations, *i.e.*, when $\mathbf{s}_i = \overset{*}{\mathbf{s}}_i$. Also, let $\hat{\zeta}_i$ and $\hat{\mu}_j$ be, respectively, agent i's score and the average evaluation for agent j when agent i lies in its evaluation for increasing agent j's share, *i.e.*, when $\mathbf{s}_i = \hat{\mathbf{s}}_i$, such that $\hat{x}_i^j > \overset{*}{x}_i^j$.

We start the proof by noting that if agent i lies in its evaluation for agent j, neither μ_i nor ζ_j changes, since they do not depend on x_i^j. We prove that Equation 2.1 holds when the agents' shares are computed using the peer-prediction mechanism with specific values for α. Mathematically,

$$\mathbb{E}\left[\mu_i + \alpha\,\hat{\zeta}_i + \hat{\mu}_j + \alpha\,\zeta_j\right]\frac{V}{(M+2\alpha)n} \leq \mathbb{E}\left[\mu_i + \alpha\,\overset{*}{\zeta}_i + \overset{*}{\mu}_j + \alpha\,\zeta_j\right]\frac{V}{(M+2\alpha)n}.$$

After doing some algebraic manipulations, we have:

$$\alpha \geq \frac{\mathbb{E}\left[\hat{\mu}_j - \overset{*}{\mu}_j\right]}{\mathbb{E}\left[\overset{*}{\zeta}_i - \hat{\zeta}_i\right]}.$$

The above inequality implies that we can set a value for α so that the peer-prediction mechanism will be collusion-resistant. While it is not possible to set α precisely, because the mechanism does not know agent i's truthful evaluations, it is possible to set up an upper-bound on:

$$\frac{\mathbb{E}\left[\hat{\mu}_j - \overset{*}{\mu}_j\right]}{\mathbb{E}\left[\overset{*}{\zeta}_i - \hat{\zeta}_i\right]}, \tag{4.5}$$

and, then, we can set up the parameter α to a value greater than or equal to that upper-bound. Starting by the numerator of (4.5), we have:

$$\begin{aligned}\mathbb{E}\left[\hat{\mu}_j - \overset{*}{\mu}_j\right] &= \mathbb{E}\left[\frac{\hat{x}_i^j - \overset{*}{x}_i^j}{n-1}\right] \\ &\leq \frac{M-1}{n-1}\end{aligned} \tag{4.6}$$

The equality follows from our collusion model, where agent i is the only one misreporting its evaluation for agent j. The inequality follows from the fact that $x_i^j \in \{1,\ldots,M\}$. Now, proceeding with the denominator of (4.5), we observe that:

$$\mathbb{E}\left[\overset{*}{\zeta}_i - \hat{\zeta}_i\right] = \frac{L\left((\omega_j|x_i^j)\middle|(\omega_j|t_i^j)\right)}{n-1},$$

where $L(\cdot\,|\,\cdot)$ is the expected scoring loss when agent i misreports its evaluation to benefit agent j (Equation 2.3). The expectation is taken with respect to agent i's posterior predictive distributions. From Equation 2.4, we have:

$$L\left((\omega_j|x_i^j)\middle|(\omega_j|t_i^j)\right) = \sum_{k=1}^{M}\left(p(\omega_j^k|t_i^j) - p(\omega_j^k|x_i^j)\right)^2.$$

Since the distribution ω_j is uniform, we have that $\forall k, p(\omega_j^k|x_i^j)$ and $p(\omega_j^k|t_i^j) \in \{\frac{1}{M+1}, \frac{2}{M+1}\}$. Thus, for $t_i^j \neq x_i^j$, we have:

$$L\left((\omega_j|x_i^j)\middle|(\omega_j|t_i^j)\right) = \frac{2}{(M+1)^2}.$$

In this way,

$$\mathbb{E}\left[\overset{*}{\zeta}_i - \hat{\zeta}_i\right] = \frac{\frac{2}{(M+1)^2}}{n-1}. \tag{4.7}$$

Combining (4.6) and (4.7), we have that if $\alpha \geq \frac{(M-1)(M+1)^2}{2}$, then the peer-prediction mechanism is collusion-resistant. □

Intuitively speaking, this proposition means that if the truth-telling scores have a high weight in the agents' shares, then these agents will not have strong incentives to collude. In the following theorem, we show a quite similar approach that guarantees the peer-prediction mechanism to be always fair.

Theorem 8. *If $\alpha < \frac{1}{2}$, then the peer-prediction mechanism is fair.*

Proof. Consider a pair of different agents $i, j \in N$ and a strategy profile $\mathbf{s} \in S$ where $x_j^i > x_i^j$, and, for every other agent $z \neq i, j$, $x_z^i > x_z^j$. In this proof, we need to show that the peer-prediction mechanism satisfies the following inequality:

$$\begin{aligned} \Gamma_i(\mathbf{s}) &> \Gamma_j(\mathbf{s}) \Rightarrow \\ (\mu_i + \alpha\,\zeta_i)\frac{V}{(M+2\alpha)n} &> (\mu_j + \alpha\,\zeta_j)\frac{V}{(M+2\alpha)n} \end{aligned}$$

After doing some algebraic manipulations, we have:

$$\alpha < \frac{\mu_i - \mu_j}{\zeta_j - \zeta_i}. \tag{4.8}$$

In what follows, we provide a lower-bound for the fraction in (4.8). Thereafter, we set α to be less than that lower-bound. Starting by the numerator, we have:

$$\begin{aligned} \mu_i - \mu_j &= \frac{\sum_{q \neq i} x_q^i - \sum_{q \neq j} x_q^j}{n-1} \\ &\geq \frac{\sum_{q \neq i,j} (x_q^i + 1 - x_q^j) + x_j^i - x_i^j}{n-1} \\ &\geq 1 \end{aligned} \tag{4.9}$$

The first inequality follows from the facts that $\forall q \neq i, j$, $x_q^i > x_q^j$, and $\forall i, j$, $x_i^j \in \{1, \ldots, M\}$. Thus, we have a lower-bound for the numerator of (4.8). Turning now to its denominator, we note that $\zeta_j - \zeta_i \leq 2$, because truth-telling scores are in the range [0, 2] (see Equation 2.4). Consequently, if we set $\alpha < \frac{1}{2}$, then we guarantee that the peer-prediction mechanism is always fair.

□

Intuitively speaking, this theorem means that we can guarantee that the peer-prediction mechanism will always be fair by putting a low weight on the truth-telling scores, so that the resulting shares will depend almost entirely on the received evaluations.

4.3 Concluding Remarks

In this chapter, we presented our second mechanism for sharing rewards using subjective opinions. The peer-prediction mechanism explicitly incentivises truthfulness by considering each evaluation reported by an agent as a bet on the average of the others' evaluations. This idea shares similarities with the work done by Miller *et al* [28]. They propose a mechanism to induce honest rating feedback. Their scheme uses one agent's report to update a probability distribution for the report of someone else (called a reference rater). The first agent is then scored not on the agreement between the ratings, but on a comparison between the probabilities assigned to the reference rater's possible ratings and the reference rater's actual rating. Differently from our mechanism, the key idea in that work is that the correlation in agents' private information can be used to induce truthful reporting.

Under mild assumptions, we showed that the peer-prediction mechanism is individually rational, weakly budget-balanced and incentive-compatible. Further, we presented

approaches that guarantee the mechanism to be collusion-resistant and fair. There are some drawbacks associated with the peer-prediction mechanism. We start by noting that the greatest possible share that each agent can receive is V/n (see Equation 4.3). This implies that the mechanism might make a very high profit. We return to this issue in Chapter 6. The second drawback is that the peer-prediction mechanism is not guaranteed to be fair and collusion-resistant at the same time. There exists a fairly intuitive trade-off here: If the mechanism overly incentivises truthfulness, it can avoid collusions, but, consequently, it makes the received evaluations be less representative inside the shares. Alternatively, if the mechanism does not adequately incentivise truthfulness, agents may have incentives to collude, but the received evaluations will be more representative inside the shares. We argue that the underlying application may help the mechanism' administrator to choose the most desirable property between fairness and collusion-resistance, and thus to set the parameter α accordingly.

Chapter 5

The BTS Mechanism

In this chapter, we propose our last mechanism for sharing rewards using subjective opinions, the *BTS mechanism*. It is based on the principle of *metaknowledge*, *i.e.*, knowledge about knowledge. In detail, we use the theory that a knowledgeable agent (or expert) is not only informed about the answer to a specific problem, but it may also know how the other agents respond to that problem. In our model, this means that an agent is not only informed about its truthful evaluation for a peer, but it may also know the likely distribution of the evaluations received by that peer.

In our previous mechanism, we used scoring rules to promote truthfulness. An evident problem with our approach is that agents may bias their evaluations toward the average evaluations, instead of telling the truth, since scoring rules does not account for variations in the quality of opinions. Hence, trying to take into account the information quality in the reported opinions, we use the BTS method (Section 2.5) to incentivise truthfulness in our new mechanism. This method is particularly suited to scenarios where minority opinions possess a greater likelihood of truth than indicated by their relative popularity.

Under the assumptions that agents are Bayesian decision-makers and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations, we show that the BTS mechanism is incentive-compatible and budget-balanced, and we present strategies to guarantee that the mechanism will always be individually rational and fair.

5.1 The Mechanism

The mechanism starts by requesting both evaluations and predictions from agents, *i.e.*, $\forall i \in N$, $\mathbf{s}_i = (\mathbf{x}_i, \mathbf{y}_i)$. Besides the basic assumptions required by the model, *i.e.*, independent

signals and self-interestedness (Section 2.1), we hold the four extra assumptions stated in Section 2.5 to be possible to use the BTS method. These assumptions essentially mean that agents are Bayesian decision-makers and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations. We discuss the implications of these extra assumptions throughout this chapter. We note that even though the BTS mechanism requires a common prior over the truthful evaluations for each peer, these priors do not need to be uniformly distributed, differently from the peer-prediction mechanism.

For each vector with evaluations, $\mathbf{x}_i$, the BTS mechanism creates a second vector, $\chi_i = (\chi_i^1, \ldots, \chi_i^{i-1}, \chi_i^{i+1}, \ldots, \chi_i^n)$, by scaling the elements of the first one to sum up to V. Mathematically,

$$\forall i, j,\ \chi_i^j = x_i^j \left(\frac{V}{\sum_{q \neq i} x_i^q} \right).$$

This simple adjustment in agents' evaluations assures that the sum of the final shares is not orders of magnitude lower than the reward V. Similarly to the peer-prediction mechanism, the share received by each agent $i \in N$ has two major components. The first one, $\bar{\chi}^i$, reflects the aggregation of the evaluations received by agent i. It is calculated by summing the scaled evaluations received by agent i, and dividing the result by n, *i.e.*,

$$\bar{\chi}^i = \frac{\sum_{j \neq i} \chi_j^i}{n}.$$

The fact that we are dividing by n instead of taking the arithmetic mean, *i.e.*, dividing by $n-1$, assures important properties for the mechanism. The second component of agent i's share is a score that explicitly incentivises truthfulness. The main idea here is that agents have their scores maximized, in an expected sense, when they truthfully report their evaluations and predictions. The score of agent i is calculated as follows:

$$\zeta_i = \frac{\sum_{j \neq i} serum(i, j)}{n - 1},$$

where $serum(i, j)$ is a value received by agent i from the BTS method when agent i reports x_i^j and $\mathbf{y}_i^j$. Thus, the score of agent i is the arithmetic mean of the results provided by the Bayesian truth serum method, where each result is directly related to an evaluation and a prediction submitted by agent i. It is important to note that we use a slightly different version of the BTS method than the original one presented in Section 2.5 (Equation 2.5). In detail, we use a recalibration coefficient, $0 < \epsilon < 1$, to adjust predictions and averages away from 0/1 extremes. Formally,

$$serum(i,j) = \sum_{k=1}^{M} h(x_i^j, k) \log \frac{\bar{x}_k}{\bar{y}_k} + \sum_{k=1}^{M} \bar{x}_k \log \left(\frac{(1-\epsilon) y_i^{j^k} + \frac{\epsilon}{M}}{\bar{x}_k} \right), \tag{5.1}$$

where $\bar{x}_k$ and $\bar{y}_k$ are now:

$$\bar{x}_k = (1-\epsilon) \left(\frac{1}{n-1} \sum_{q \neq j} h(x_q^j, k) \right) + \frac{\epsilon}{M},$$

$$\bar{y}_k = \exp \left(\frac{1}{n-1} \sum_{q \neq j} \log \left((1-\epsilon) y_q^{j^k} + \frac{\epsilon}{M} \right) \right).$$

where $h(x_i^j, k)$ is a zero-one indicator function. Such a change prevents problems related to values for $\log(0)$ and $\log(0/0)$. We argue that it does not have influence on the original properties of the BTS method, *i.e.*, Theorem 1 to 4 still hold, since we are essentially reducing the range of predictions and averages. Now, we have that:

$$0 < \frac{\epsilon}{M} < \bar{x}_k, \bar{y}_k < 1 - \epsilon + \frac{\epsilon}{M} < 1 \tag{5.2}$$

We can see the computation of each agent's score as if it is answering multiple independent questions and receiving multiple independent scores for its answers, where each question asks for both an evaluation for a peer and a prediction for the empirical distribution of evaluations received by that peer. At the end, a final score is calculated as the arithmetic mean of those multiple independent scores. In this way, the function $serum(i,j)$ represents the score given to agent i when it answers the question about the evaluation and prediction for agent j.

Finally, the share of agent i is a linear combination of the scaled evaluations received by it, $\bar{\chi}^i$, and agent i's score, ζ_i, *i.e.*,

$$\Gamma_i = \bar{\chi}^i + \alpha \, \zeta_i, \tag{5.3}$$

where the constant $\alpha > 0$ fine-tunes the weight given to the score ζ_i. It is useful to note that despite the assumptions of the prior and posterior distributions required to use the BTS method, they are neither known nor requested by the mechanism, only evaluations and predictions are elicited from agents. Algorithm 3 presents the sharing function of the BTS mechanism from an algorithmic perspective. To illustrate the BTS mechanism, consider the following example:

Algorithm 3 The BTS Mechanism

for $i = 1$ to n **do**
  for $j \neq i$ **do**
    $\chi_i^j = x_i^j \left(\frac{V}{\sum_{q \neq i} x_i^q} \right)$
  end for
end for
for $i = 1$ to n **do**
  $\bar{\chi}^i = \frac{\sum_{j \neq i} \chi_j^i}{n}$
  $\zeta_i = \frac{\sum_{j \neq i} serum(i,j)}{n-1}$
  $\Gamma_i = \bar{\chi}^i + \alpha\, \zeta_i$
end for

Example 5. *Suppose that six agents A, B, C, D, E and F, want to share a reward $V = 1000$ using the BTS mechanism and a "positive/negative" scheme, i.e., $M = 2$. The reported evaluations and predictions can be seen, respectively, in the left part of Table 5.1 and in Table 5.2.*

In Table 5.1, each numeric cell beneath the label "Evaluation" can be interpreted as the evaluation given by the agent in the row to the agent in the column. Each numeric cell beneath the label "Scaled Evaluation" has the same meaning, but this time the evaluations reported by each agent are scaled to sum up to V.

Table 5.1: Example of the BTS mechanism: reported evaluations.

	Evaluation						**Scaled Evaluation**					
	A	B	C	D	E	F	A	B	C	D	E	F
A	-	2	2	1	1	1	-	285.71	285.71	142.86	142.86	142.86
B	1	-	2	2	1	2	125.00	-	250.00	250.00	125.00	250.00
C	1	2	-	1	1	2	142.86	285.71	-	142.86	142.86	285.71
D	1	2	2	-	1	2	125.00	250.00	250.00	-	125.00	250.00
E	2	2	1	2	-	2	222.22	222.22	111.11	222.22	-	222.22
F	2	2	1	2	1	-	250.00	250.00	125.00	250.00	125.00	-

*In Table 5.2, each numeric cell can be interpreted as the prediction made by the agent in the row about the percentage of agents that submit the evaluation in the second row (**"1"** or **"2"**) to the agent in the column. For example, the emphasized value* 0.8*, below the label "A" and on the right of the label "B", means that agent B predicts that 80% of the population give the evaluation "1" to agent A. Using these evaluations and predictions, and setting the mechanism's parameters $\alpha = 100$ and $\epsilon = 0.01$, the mechanism returns the shares presented in the last column of Table 5.3.*

Table 5.2: Example of the BTS mechanism: reported predictions.

	A		**B**		**C**		**D**		**E**		**F**	
	"1"	**"2"**	**"1"**	**"2"**	**"1"**	**"2"**	**"1"**	**"2"**	**"1"**	**"2"**	**"1"**	**"2"**
A	-	-	0	1	0.4	0.6	0.2	0.8	1	0	0.2	0.8
B	*0.8*	0.2	-	-	0.2	0.8	0.2	0.8	1	0	0.4	0.6
C	0.8	0.2	0	1	-	-	0.4	0.6	1	0	0.4	0.6
D	0.8	0.2	0.2	0.8	0.6	0.4	-	-	0.8	0.2	0.4	0.6
E	0.8	0.2	0	1	0.6	0.4	0.4	0.6	-	-	0.4	0.6
F	0.8	0.2	0.8	0.2	0.6	0.4	0.4	0.6	0.8	0.2	-	-

Table 5.3: Example of the BTS mechanism: resulting shares.

	$\bar{\chi}^i$	ζ_i	Γ_i
A	144.18	0.05	149.39
B	215.61	-0.06	209.73
C	170.30	0.09	179.80
D	167.99	-0.02	166.03
E	110.12	0.16	125.76
F	191.80	-0.21	170.80

For illustration's sake, consider the share received by agent F. The first component of Γ_F is the aggregation of the scaled evaluations received by agent F, that is:

$$\begin{aligned} \bar{\chi}^F &= \frac{142.86 + 250.00 + 285.71 + 250.00 + 222.22}{6} \\ &\approx 191.80. \end{aligned}$$

The second component of Γ_F is the arithmetic mean of the results provided by the BTS method (Equation 5.1), each one related to an evaluation and a prediction reported by agent F:

$$\begin{aligned} \zeta_F &= \frac{serum(F,A) + serum(F,B) + serum(F,C) + serum(F,D) + serum(F,E)}{5} \\ &\approx \frac{0.58 - 1.19 - 0.18 - 0.11 - 0.11}{5} \\ &\approx -0.21. \end{aligned}$$

Finally, the share of agent F is a linear combination of $\bar{\chi}^F$ and ζ_F.

$$
\begin{aligned}
\Gamma_F &= \bar{\chi}^F + \alpha\,\zeta_F \\
&\approx 191.80 + 100 \times (-0.21) \\
&\approx 170.80.
\end{aligned}
$$

5.2 Properties

In this section, we show the properties concerning the BTS mechanism.

Proposition 5. *The BTS mechanism is budget-balanced.*

Proof. The sum of the shares received by the agents is equal to:

$$
\begin{aligned}
\sum_{i=1}^{n} \left(\bar{\chi}^i + \alpha\,\zeta_i\right) &= \sum_{i=1}^{n} \bar{\chi}^i + \alpha \sum_{i=1}^{n} \zeta_i \\
&= \sum_{i=1}^{n} \left(\frac{\sum_{j \neq i} \chi^i_j}{n} \right) + \alpha \sum_{i=1}^{n} \left(\frac{\sum_{j \neq i} serum(i,j)}{n-1} \right) \\
&= \sum_{j=1}^{n} \left(\frac{\sum_{i \neq j} \chi^i_j}{n} \right) + \alpha \sum_{j=1}^{n} \left(\frac{\sum_{i \neq j} serum(i,j)}{n-1} \right) \\
&= n \left(\frac{V}{n} \right) + \frac{\alpha}{n-1} \left(\sum_{j=1}^{n} \sum_{i \neq j} serum(i,j) \right).
\end{aligned}
$$

The left part of this last equation follows from the fact that the scaled evaluations sum up to V. From Theorem 4, we know that if the BTS's parameter $\beta = 1$, then $\sum_{i \neq j} serum(i,j) = 0$. Then, the right part of this last equation is equal to zero, completing the proof. □

Proposition 6. *If* $\alpha \leq \frac{V}{2Mn \log\left(\frac{M}{\epsilon}\right)}$, *then the BTS mechanism is individually rational.*

Proof. We start the proof by observing that $\forall i \in N$, $\bar{\chi}^i \geq 0$. Thus, whenever the scores of the agents are positive, their shares will also be positive because $\alpha > 0$. So, we focus on the case where the scores are negative. In what follows, we set a value for α so that we guarantee that the share of every agent $i \in N$ is greater than or equal to zero, even when $\zeta_i \leq 0$. Formally, we assure that the following inequality is still valid when $\zeta_i \leq 0$:

$$\frac{\bar{\chi}^i}{-\zeta_i} \geq \alpha.$$

For doing this, we compute a lower-bound for the above fraction. Thereafter, we set α to be less than or equal to that lower-bound. Starting by the numerator, we have:

$$\begin{aligned} \bar{\chi}^i &= \frac{\sum_{j \neq i} x_j^i \left(\frac{V}{\sum_{q \neq j} x_j^q} \right)}{n} \\ &\geq \frac{\sum_{j \neq i} x_j^i \left(\frac{V}{M\,(n-1)} \right)}{n} \\ &\geq \frac{V(n-1)}{M\,n(n-1)} \\ &= \frac{V}{M\,n} \end{aligned} \tag{5.4}$$

The inequalities follow from the fact $\forall i, j,\, x_i^j \in \{1, \ldots, M\}$. Now, we compute the smallest negative value that ζ_i can have. Since ζ_i is the average of $n-1$ results from the BTS method, we can restrict ourselves to find the smallest negative value that can be returned by the BTS method (Equation 5.1). For simplicity's sake, consider the value returned by $serum(i, j)$. Focusing on the left part of Equation 5.1, we have:

$$\begin{aligned} \sum_{k=1}^{M} h(x_i^j, k) \log \frac{\bar{x}_k}{\bar{y}_k} &\geq \sum_{k=1}^{M} h(x_i^j, k) \log \bar{x}_k \\ &\geq \log\left(\frac{\epsilon}{M}\right) \end{aligned} \tag{5.5}$$

The first inequality follows from $0 < \bar{y}_k < 1$. The second inequality follows from $\frac{\epsilon}{M} < \bar{x}_k < 1 - \epsilon + \frac{\epsilon}{M}$. Moving to the right part of Equation 5.1, we have:

$$\begin{aligned} \sum_{k=1}^{M} \bar{x}_k \log\left(\frac{(1-\epsilon) y_i^{j^k} + \frac{\epsilon}{M}}{\bar{x}_k}\right) &\geq \sum_{k=1}^{M} \bar{x}_k \log\left(\frac{\epsilon}{M}\right) \\ &= \left(1 - \epsilon + \frac{\epsilon}{M}\right) \log\left(\frac{\epsilon}{M}\right) \\ &\geq \log\left(\frac{\epsilon}{M}\right) \end{aligned} \tag{5.6}$$

The first inequality follows from the facts that $0 < \bar{x}_k < 1$ and $(1-\epsilon)y_i^{j^k} \geq 0$. The second inequality follows from the facts that $\log(\epsilon/M) < 0$, and $\left(1 - \epsilon + \frac{\epsilon}{M}\right) < 1$. Joining 5.5 and 5.6, we have:

$$serum(i,j) \geq 2\log\left(\frac{\epsilon}{M}\right) \tag{5.7}$$

Finally, joining 5.4 and 5.7, we conclude that:

$$\begin{aligned} \alpha &\leq \frac{V}{M\,n\,\left(-2\log\left(\frac{\epsilon}{M}\right)\right)} \Rightarrow \\ \alpha &\leq \frac{V}{2Mn\log\left(\frac{M}{\epsilon}\right)} \end{aligned}$$

□

Since agents' scores can be negative, the above proposition says that we can guarantee that the shares will always be positive by putting a low weight on scores. For illustration's sake, if $\alpha \leq 7.87$, then the shares returned by the BTS mechanism, in the scenario of Example 5, are greater than zero, regardless the evaluations and predictions reported by the agents.

Theorem 9. *The BTS mechanism is incentive-compatible.*

Proof. Suppose that every peer of an agent $i \in N$ truthfully reports its strategy. We prove that the best response of agent i, in an expected sense, is also to tell the truth. Consider the share received by agent i from the mechanism:

$$\begin{aligned} \Gamma_i &= \bar{\chi}^i + \alpha\,\zeta_i \\ &= \frac{\sum_{j\neq i}\chi_j^i}{n} + \alpha\left(\frac{\sum_{j\neq i} serum(i,j)}{n-1}\right) \\ &= C_1 + C_2\sum_{j\neq i} serum(i,j), \end{aligned}$$

where C_1 and C_2 are positive constants, from agent i's point of view, because they do not depend on the strategy reported by agent i. We note that $\sum_{j\neq i} serum(i,j)$ is similar to the function $g_i(\mathbf{Z})$, defined in Equation 2.6. Consequently, agent i's share can be seen as a positive affine transformation of the scoring function g, and, according to Lemma 4, it is maximized, in an expected sense, when agent i tells the truth, where the expectation in taken with respect to agent i posterior beliefs. So, the BTS mechanism is incentive-compatible and the collectively truthful strategy profile is a Bayes-Nash equilibrium. □

Theorem 10. *If $M \leq \sqrt{n-2}$ and $\alpha \leq \frac{V}{3Mn^2 \log\left(\frac{M}{\epsilon}\right)}$, then the BTS mechanism is fair.*

Proof. Consider a pair of different agents $i, j \in N$ and a strategy profile $\mathbf{s} \in S$ where $x_j^i > x_i^j$, and, for every other agent $z \neq i, j$, $x_z^i > x_z^j$. In this proof, we show that, with appropriated values for α, the BTS mechanism always satisfies the following inequality:

$$\begin{aligned} \Gamma_i(\mathbf{s}) &> \Gamma_j(\mathbf{s}) \Rightarrow \\ \bar{\chi}^i + \alpha\,\zeta_i &> \bar{\chi}^j + \alpha\,\zeta_j \end{aligned}$$

After doing some algebraic manipulations, we have:

$$\alpha < \frac{\bar{\chi}^i - \bar{\chi}^j}{\zeta_j - \zeta_i}. \tag{5.8}$$

In what follows, we provide a lower-bound for the above fraction. Thereafter, we set α to be less than that lower-bound. Starting by its numerator, we have:

$$\begin{aligned} \bar{\chi}^i - \bar{\chi}^j &= \frac{\sum_{z \neq i,j} \left(x_z^i - x_z^j\right) \left(\frac{V}{\sum_{q \neq z} x_z^q}\right) + x_j^i \left(\frac{V}{\sum_{q \neq j} x_j^q}\right) - x_i^j \left(\frac{V}{\sum_{q \neq i} x_i^q}\right)}{n} \\ &\geq \frac{V}{n}\left(\frac{n-2}{(n-1)M} + \frac{1}{(n-1)M} - \frac{M}{(n-1)}\right) \\ &= \frac{V}{n}\left(\frac{n-2+1-M^2}{(n-1)M}\right) \\ &\geq \frac{V}{n(n-1)M} \\ &\geq \frac{V}{n^2 M}. \end{aligned} \tag{5.9}$$

The first inequality follows from the facts that for every agent $z \neq i, j$, $x_z^i > x_z^j$, and that $\forall i, j$, $x_i^j \in \{1, \ldots, M\}$. The second inequality follows from the assumption that $M \leq \sqrt{n-2}$. Moving to the denominator of 5.8, since each score is the arithmetic mean of the results from the BTS method (Equation 5.1), and that scores can be negative, we can restrict ourselves to compute the largest positive and the smallest negative value that the function $serum(i, j)$ can return. In the proof of Proposition 6, we found that $serum(i, j) \geq 2 \log \frac{\epsilon}{M}$. To compute the largest positive value, we start by focusing on the left part of Equation 5.1:

$$\begin{aligned}\sum_{k=1}^{M} h(x_i^j,k)\log\frac{\bar{x}_k}{\bar{y}_k} &\leq \sum_{k=1}^{M} h(x_i^j,k)\log\frac{1}{\bar{y}_k} \\ &\leq \log\left(\frac{1}{\frac{\epsilon}{M}}\right) \\ &= \log\left(\frac{M}{\epsilon}\right)\end{aligned} \tag{5.10}$$

The first inequality follows from $0 < \bar{x}_k < 1$. The second inequality follows from $\frac{\epsilon}{M} < \bar{y}_k < 1 - \epsilon + \frac{\epsilon}{M}$. Moving to the right part of Equation 5.1, we note that it is always less than or equal to zero, because it can be seen as the negative of the Kullback-Leibler divergence, which is always greater than or equal to zero [9]. Thus, we have:

$$serum(i,j) \leq \log\left(\frac{M}{\epsilon}\right) \tag{5.11}$$

Joining 5.7 and 5.11, we have:

$$\zeta_j - \zeta_i \leq \log\left(\frac{M}{\epsilon}\right) - 2\log\left(\frac{\epsilon}{M}\right) \tag{5.12}$$

Finally, joining 5.9 and 5.12, we conclude that:

$$\alpha \leq \frac{V}{Mn^2\left(\log\left(\frac{M}{\epsilon}\right) - 2\log\left(\frac{\epsilon}{M}\right)\right)} = \frac{V}{3Mn^2\log\left(\frac{M}{\epsilon}\right)}$$

□

Since the bound for α in Theorem 10 is less than the bound in Proposition 6, we conclude that whenever the BTS mechanism is guaranteed to be fair, if $M \leq \sqrt{n-2}$, then the mechanism will also be guaranteed to be individually rational. Related to the collusion-resistance property, we briefly note that for making the BTS mechanism collusion-resistant, the following inequality must hold:

$$\alpha \geq \frac{\mathbb{E}\left[\hat{\overset{*}{\chi}}^j - \overset{*}{\hat{\chi}}^j\right]}{\mathbb{E}\left[\overset{*}{\zeta}_i - \hat{\zeta}_i\right]}, \tag{5.13}$$

where $\overset{*}{\zeta}_i$ and $\overset{*}{\bar{\chi}}^j$ are, respectively, agent i's score and the aggregated scaled evaluation for agent j when agent i reports its truthful evaluations, *i.e.*, when $\mathbf{s}_i = \overset{*}{\mathbf{s}}_i$. Oppositely, $\hat{\zeta}_i$ and $\hat{\bar{\chi}}^j$ are, respectively, agent i's score and the aggregated scaled evaluation for agent j when agent i lies in its evaluation for increasing agent j's share, *i.e.*, when $\mathbf{s}_i = \hat{\mathbf{s}}_i$, such that $\hat{x}_i^j > \overset{*}{x}_i^j$.

There exist two problems with the denominator of the above fraction. First, we observe that this expectation can be derived from the proofs of Theorem 1 and 4 [32], but the resulting value depends on probability distributions that directly use the common prior distribution (see Assumption 3, in Section 2.5). Further, there is no guarantee that this expected value is either positive or negative, or even zero. Another way to visualize this fact is by noting that, according to Theorem 2, the expected information score (left part of Equation 5.1) is non-negative, and that the expected prediction score (right part of Equation 5.1) is non-positive, since it is a penalty proportional to the relative entropy. Consequently, we cannot make the BTS mechanism collusion-resistant by fine-tuning the weight given to scores. Further investigation on schemes for guaranteeing collusion-resistance is left as future research work.

Since a beneficiary does not have its score increased when an agent lies to increase its share, the maximum value that an agent can receive from the BTS mechanism due exclusively to a collusive behavior is equal to the similar value from the peer-evaluation mechanism. Thus, Proposition 2 also holds here, *i.e.*, the maximum value that an agent can receive due to a collusive behavior is less than $\frac{V(M-1)}{(n-1)n}$.

5.3 Concluding Remarks

In this chapter, we presented our last mechanism for sharing rewards using subjective opinions. The BTS mechanism is based on the principle of metaknowledge, where a knowledgeable agent knows both its evaluation for a peer and the likely distribution of the reported evaluations for that peer. The mechanism incentivises truthfulness by using the BTS method on the reported evaluations and predictions.

Under the assumptions that agents are Bayesian decision-makers and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations, we showed that the BTS mechanism is incentive-compatible and budget-balanced, and we presented strategies to guarantee that the mechanism will always be individually rational and fair.

Another property of our mechanism worthwhile to discuss is the non-consensuality. An agent does not necessarily increase its share by biasing its evaluations toward the likely group consensus. First, its aggregated scaled evaluations do not depend on its

reported evaluations. Second, its score does not necessarily increase due to the surprisingly common criterion used by the Bayesian truth serum method. For illustration, suppose that a particular evaluation is highly predicted for a peer and endorsed by the majority, but less than the predicted number of agents. Such evaluation is a surprisingly uncommon answer, hence it receives a low score. Then, agents who believe that their evaluations represent a minority view do better in not biasing their evaluations toward the consensus, because, even as a minority, their evaluations can still be surprisingly common and, consequently, receive a higher score. This is, arguably, the main difference between the BTS mechanism and the peer-prediction mechanism, *i.e.*, the former takes into account the information quality in the reported opinions when computing scores.

Finally, despite the fact that we were unable to show a strategy to avoid collusions, we argue that in practical applications it may not be easy for the agents to undermine the BTS mechanism through collusions. In some way, the colluders must develop a sophisticated theory on how the other agents are evaluating and predicting, since a liar agent may have its score substantially decreased if its untruthful evaluation turns out to be surprisingly uncommon.

Related to the four extra assumptions made (see Section 2.5), we note that the only reason for holding them is to make the mathematics behind the BTS method work. The first three mean that agents are Bayesian decision-makers. Despite the fact that there is little evidence that they are realistic, these assumptions are traditional in Bayesian game-theoretic work [31, 21]. By far, the most stringent of our assumptions is the requirement of a large population. However, we note that some works have successfully used the BTS method with small populations. For example, Weiss used the BTS method to asses chess expertise in two different populations: the experts, which have 15 members, and the novices, with 13 members [46]. Prelec and Seung successfully used the BTS method with two distinct populations having sizes, respectively, 51 and 32 [33]. Finally, Prelec and Weaver empirically showed that the BTS method encourages and rewards truthfulness [34]. They used 4 populations, all with size 33. We observe that more important than having a very large population of agents is a guarantee that each empirical distribution of received evaluations is balanced, *i.e.*, that there exists a significant number of endorsements for each possible evaluation. In this way, the influence of a single agent on the empirical distributions of evaluations is reduced. A good rule of thumb is to use a value for the parameter M less than or equal to $\sqrt{n-2}$ because it only allows a small number of possible evaluations in relation to the number of agents, and, with appropriate values for α, this value also assures fairness.

Chapter 6

Numerical Experiments

In this chapter, we empirically investigate the influence of the mechanisms' parameters on agents' shares. We start by studying the consequences of using different values for the parameter M, the top possible evaluation that an agent can give or receive, when the truthful evaluations are either uniformly distributed or normally distributed.

Thereafter, we analyze the behavior of the mechanisms with different values for α, the parameter that fine-tunes the weight given to the truth-telling scores. In detail, we empirically investigate when fairness, individual rationality and collusion-resistance hold, considering that the truthful evaluations are either uniformly distributed or normally distributed. Finally, we analyze how the mechanisms behave when dealing with populations of different sizes.

6.1 Parameter M

The parameter M defines the range of possible evaluations that an agent can give or receive. A natural question to ask is which value of M should be used. We start by noting that M has a great influence on two aspects of the mechanisms. First, as M increases, then the evaluations can be more fine-grained, in that small differences between agents can be recognized and specified by their peers. However, this increased expressivity can also make the evaluation process more challenging, since agents will have more possibilities to evaluate their peers. On the other hand, lower values of M make the evaluation step of the agents simpler, since they do not need to differentiate between their peers as much. However, this may increase the sense of unfairness among the agents, since they cannot be properly differentiated and, thus, they may end up with similar shares.

To better understand the influence of different values of M on agents' shares, we propose the following experiment. For each proposed mechanism, we share the reward $V = 10000$

among 100 agents using the following values for M: $2, 5, 7, 10, 25, 50, 75$ and 100. We deliberately choose a fairly large population to make the results statistically significant. For the peer-prediction mechanism, we use the fixed parameter $\alpha = 0.5$, and, for the BTS mechanism, we use the parameters $\alpha = 10$ and $\epsilon = 0.01$. In this experiment, agents always report their observed signals. Firstly, these signals are drawn from a uniform distribution. For generating random predictions, we use the algorithm proposed by Stafford [43]. This algorithm creates random, uniformly distributed vectors with fixed sum. We observe the mean, the standard deviation and the range of the resulting shares. Table 6.1 shows these values. Figure 6.1, 6.2 and 6.3 visually show the same results.

Table 6.1: Results of the proposed mechanisms with different values for M. The other parameters are fixed: $n = 100$, V $= 10000$, $\alpha = 0.5$ (peer-prediction), $\alpha = 10$ (BTS), $\epsilon = 0.01$. Truthful evaluations are uniformly distributed.

	Peer-Evaluation			**Peer-Prediction**			**BTS**		
M	Avg.	Std.	Range	Avg.	Std.	Range	Avg.	Std.	Range
2	100.00	2.85	11.75	73.61	1.50	7.41	100.00	2.82	11.91
5	100.00	4.73	28.31	59.57	2.32	13.64	100.00	4.72	27.55
7	100.00	5.00	24.05	57.48	2.54	12.25	100.01	4.92	23.12
10	100.00	5.47	28.66	55.32	2.74	14.57	100.01	5.48	29.42
25	100.00	5.75	28.37	51.99	2.87	14.02	100.03	5.79	28.43
50	100.00	4.80	23.37	50.79	2.38	11.23	100.09	4.89	23.77
75	100.00	6.16	28.33	50.28	3.07	14.43	100.15	6.30	29.92
100	100.00	5.78	30.90	50.78	2.90	15.78	100.21	5.74	31.17

Figure 6.1: Results of the peer-evaluation mechanism with different values for M when truthful evaluations are uniformly distributed. The averages of the shares are represented by the black squares, and the standard deviations by the gray lines. The dotted line is used to facilitate visualization. The parameters are: $n = 100$, $V = 10000$.

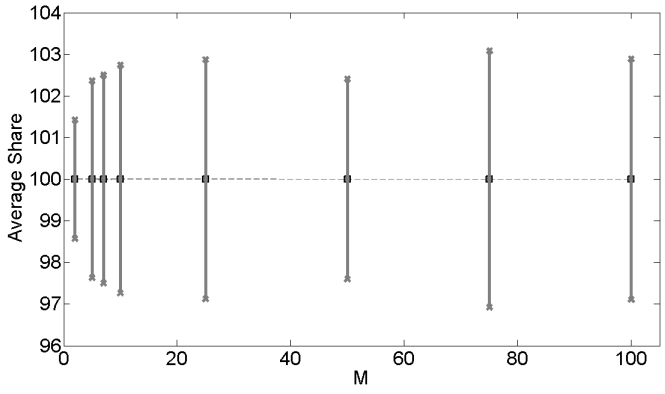


Figure 6.2: Results of the peer-prediction mechanism with different values for M when truthful evaluations are uniformly distributed. The averages of the shares are represented by the black squares, and the standard deviations by the gray lines. The dotted line is used to facilitate visualization. The parameters are: $n = 100$, $V = 10000$, $\alpha = 0.5$.

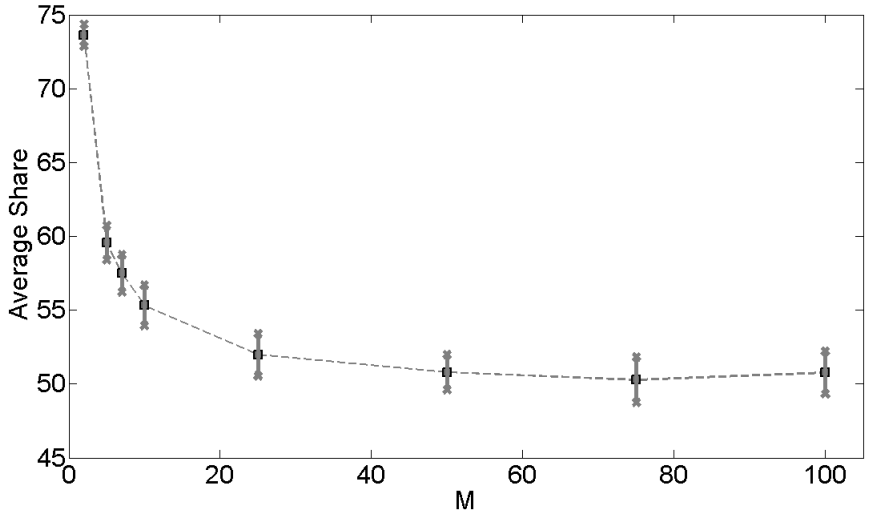
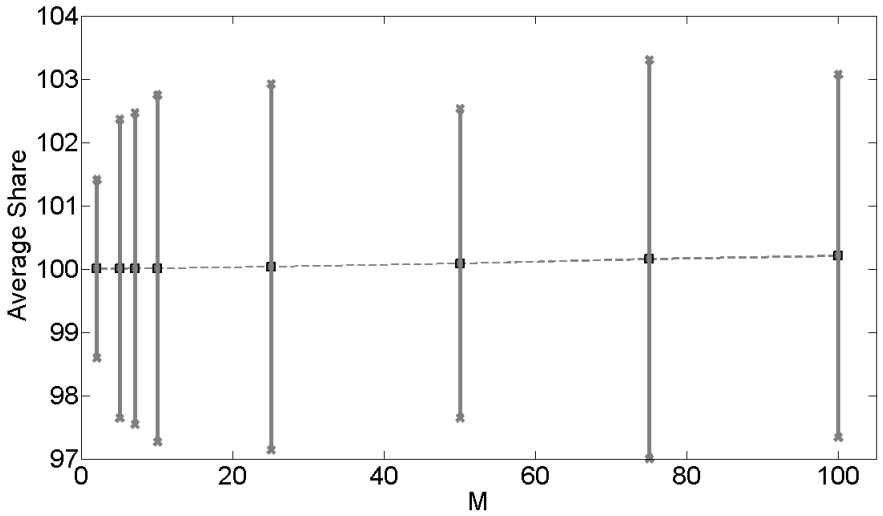


Figure 6.3: Results of the BTS mechanism with different values for M when truthful evaluations are uniformly distributed. The averages of the shares are represented by the black squares, and the standard deviations by the gray lines. The dotted line is used to facilitate visualization. The parameters are: $n = 100$, $V = 10000$, $\alpha = 10$, $\epsilon = 0.01$.



From the results of this first experiment, the influence of M on the standard deviation and range of agents' shares seems to be negligible. Intuitively, since the evaluations received by each agent are uniformly distributed, it is very difficult (in a probabilistic sense) that an agent receives a lot of extremely positive or negative evaluations. In other words, almost all agents end up with similar aggregated evaluations.

Alternatively, as M increases, the average share returned by the peer-prediction mechanism decreases, *i.e.*, the mechanism makes a larger profit. This was already expected because agents' shares are scaled by using the constant $\frac{V}{(M+2\alpha)n}$, which has the parameter M in its denominator (Equation 4.3). Consequently, the higher the value of M, the lower the average of the resulting shares. A similar situation does not occur with the other mechanisms because they are budget-balanced, *i.e.*, the average share stays (approximately) constant.

Trying to understand the influence of M on the proposed mechanisms in a (arguably) more realistic scenario, we repeat the experiment, but this time assuming that agents' truthful evaluations follow a normal distribution. In detail, for each agent in the population, we uniformly select a value inside the set $\{1, \ldots, M\}$, and use it as the mean of a normal distribution, which is employed to generate the truthful evaluations for that agent. We use a not so large variance with value $M/8$. This scenario implies that the agents observe fairly similar signals from a specific peer. If a random evaluation is less than 1 or greater than M, we reset this value to, respectively, 1 and M. We use the same values for the other parameters as before, *i.e.*, $V = 10000$, $n = 100$, $M \in \{2, 5, 7, 10, 25, 50, 75, 100\}$, $\alpha = 0.5$ for the peer-prediction mechanism, $\alpha = 10$ for the BTS mechanism, and $\epsilon = 0.01$. Agents report their observed signals and predictions are generated by using Stafford's algorithm [43]. Table 6.1 shows the mean, the standard deviation and the range of the resulting shares in this new scenario. Figure 6.4, 6.5 and 6.6 visually show similar results.

Table 6.2: Results of the proposed mechanisms with different values for M. The other parameters are fixed: $n = 100$, V = 10000, $\alpha = 0.5$ (peer-prediction), $\alpha = 10$ (BTS), $\epsilon = 0.01$. Truthful evaluations are normally distributed.

	Peer-Evaluation			**Peer-Prediction**			**BTS**		
M	Avg.	Std.	Range	Avg.	Std.	Range	Avg.	Std.	Range
2	100.00	13.77	34.17	85.29	8.07	20.99	100.09	13.82	34.23
5	100.00	35.08	101.31	68.34	20.23	58.42	100.21	35.01	102.35
7	100.00	41.12	127.07	61.05	22.02	68.20	100.25	41.29	125.79
10	100.00	47.88	142.13	58.31	25.42	75.42	100.30	48.05	141.31
25	100.00	47.95	158.89	57.58	26.63	88.26	100.42	47.96	158.73
50	100.00	50.16	170.39	54.87	27.02	91.80	100.50	50.40	172.30
75	100.00	53.84	188.44	49.99	26.56	92.93	100.53	53.95	188.98
100	100.00	55.55	183.23	52.60	28.94	95.46	100.57	55.52	184.35

Figure 6.4: Results of the peer-evaluation mechanism with different values for M when truthful evaluations are normally distributed. The averages of the shares are represented by the black squares, and the standard deviations by the gray lines. The dotted line is used to facilitate visualization. The parameters are: $n = 100$, $V = 10000$.

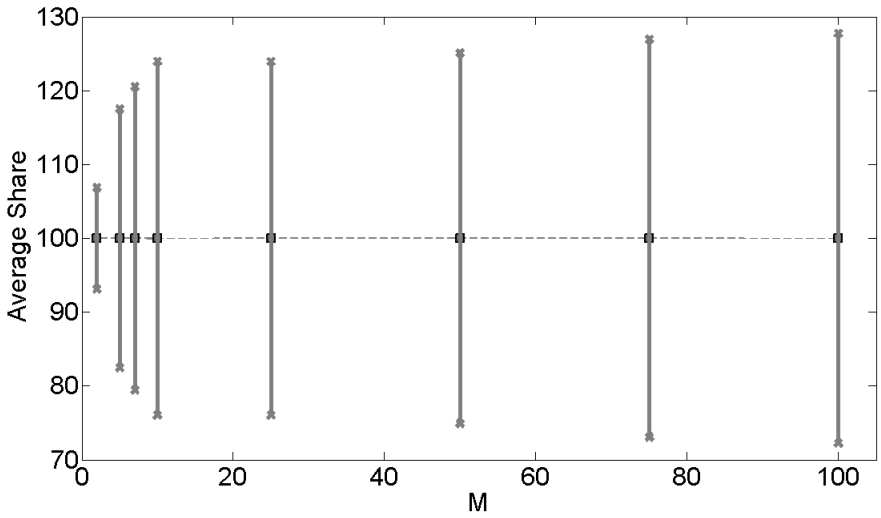


Figure 6.5: Results of the peer-prediction mechanism with different values for M when truthful evaluations are normally distributed. The averages of the shares are represented by the black squares, and the standard deviations by the gray lines. The dotted line is used to facilitate visualization. The parameters are: $n = 100$, $V = 10000$, $\alpha = 0.5$.

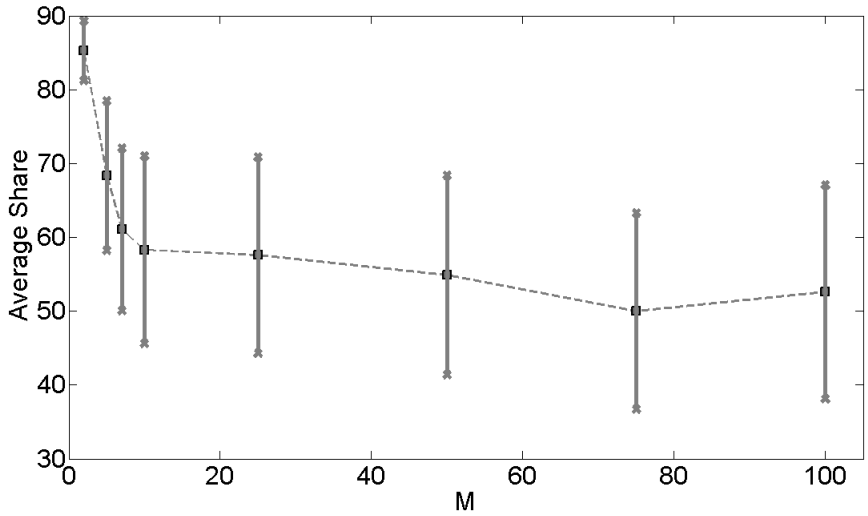
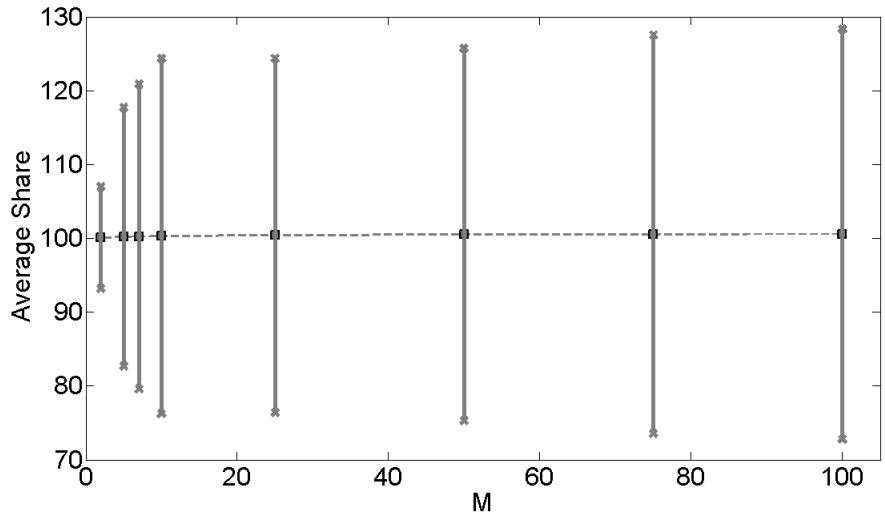


Figure 6.6: Results of the BTS mechanism with different values for M when truthful evaluations are normally distributed. The averages of the shares are represented by the black squares, and the standard deviations by the gray lines. The dotted line is used to facilitate visualization. The parameters are: $n = 100$, $V = 10000$, $\alpha = 10$, $\epsilon = 0.01$.



From the results of this second experiment, we observe that as M increases, the standard deviation and range of the resulting shares often significantly increase. Intuitively, the differences between agents are better recognized and specified by the peers, resulting in fine-grained differences in the final shares. Another way to explain these results is by noting that it is not so rare, in a probabilistic sense, that an agent receives a lot of extremely positive or negative evaluations, when these evaluations are normally distributed with a large or a small mean. Thus, differently from the previous experiment, the resulting evaluations and, consequently, shares are not very similar. Also, as M increases, the average share returned by the peer-prediction mechanism decreases, similarly to what happened in the previous experiment, and according to what is expected from the theory.

A point interesting to note is that the results, in both experiments, related to the peer-evaluation mechanism and the BTS mechanism are very similar. This happens because we deliberately choose a fairly small value for α, the constant that fine-tunes the weight given to the truth-telling scores. Consequently, the sharing functions of these mechanisms become very similar (see Equation 3.1 and Equation 5.3). The reason for doing this is that we explicitly program the agents to tell the truth. Consequently, it is not necessary to incentivise truthfulness. Another interesting point in both experiments is that when M comes closer to n, the budget-balance property of the BTS mechanism seems to dissolve since the average share is greater than V/n. We return to this point in Section 6.3.

We end this section by noting that while it might by theoretically interesting to use a

small value for M, since this may guarantee fairness (see Theorem 6, 8, 10) and a lower profit for the peer-prediction mechanism (see Equation 4.3), this may also result in a more egalitarian assignment of the shares, since it will be more difficult for the agents to express differences between themselves. We argue that the underlying application may help to determine appropriate settings for M.

6.2 Parameter α

The shares returned by the peer-prediction and BTS mechanisms have two major components. While the first one reflects the evaluations received by the agents, the second one is a truth-telling score with the complementary role of incentivising truthfulness. We need to balance these components for creating meaningful results.

Our solution is to use a constant, α, which fine-tunes the weight given to the truth-telling scores. This parameter has an important role in ensuring desirable properties for the mechanisms (see Chapter 4 and 5). In this section, we investigate this role from an empirical perspective. In detail, we look at how different values of α influence collusion-resistance, individual rationality and fairness properties.

We start by making the following experiment. For each proposed mechanism, we share the reward $V = 10000$ among 30 agents using the parameter $M = 5$ and the following values for α: $0.1, 1, 5, 10, 25, 50, 100, 500$. We use the recalibration coefficient $\epsilon = 0.01$ for the BTS mechanism. Even though the peer-evaluation mechanism does not use truth-telling scores and, consequently, it does not use the parameter α, we include it in this experiment for completeness' sake. The truthful evaluations are drawn from a uniform distribution. Predictions are randomly generated according to the algorithm proposed by Stafford [43]. We run this experiment 100 times and, for each value of α, we observe three points. First, we notice the difference between the average joint share of two fixed agents when both of them truthfully report their opinions and when they are colluding, *i.e.*, when one of them lies trying to increase the share of the other. The liar agent always reports the top possible evaluation for its collusion partner, and its truthful evaluations for the others. Everyone else in the population reports its observed signals. Thus, we are empirically looking at values of α that can prevent such collusions, *i.e.*, values where the expected joint share when both agents tell the truth is greater than when they are colluding.

Second, for each value of α, we compute the number of unfair shares returned by each mechanism throughout the experiment. In detail, for each simulation step, we make a pairwise comparison of the shares returned by a mechanism (when all agents truthfully report their opinions) determining whether they are fair or not (see Definition 4).

Finally, we compute the number of negative shares returned by each mechanism throughout the experiment when all agents truthfully report their opinions. Table 6.3, 6.4, and

6.5 show, respectively, the results of this experiment for the peer-evaluation, the peer-prediction and the BTS mechanisms. Figure 6.7, 6.8, and 6.9 visually show the results related to the collusion-resistance property.

To show the statistical significance of the results obtained for the collusion-resistance property, we perform the directional t-test. Our null hypothesis is that the average joint share when the fixed agents are telling the truth is equal to the average joint share when they are colluding. Our alternative hypothesis is that the average joint share when both agents are telling the truth is greater than the average joint share when they are colluding. Since each agent that is not colluding is always reporting its observed signals, then the average joint share when the fixed agents are telling the truth and the average joint share when they are colluding are correlated. Thus, we use the t-test for correlated samples (also known as paired t-test). This test allows us to remove irrelevant and extraneous information from the analyzed shares. The resulting p-values can be seen in Table 6.6.

Focusing first on the results related to the collusion-resistance property, from Table 6.3 we observe that the average joint share returned by the peer-evaluation mechanism when the fixed agents are colluding is always greater than when they are telling the truth. Further, these values do not change with different values for α. These points were already expected since this mechanism does not use truth-telling scores and, consequently, it does not use α.

Moving to the results for the peer-prediction mechanism, we first note that according to the theory (Proposition 4, Section 4.2), if $\alpha > \frac{(M-1)(M+1)^2}{2}$, then the mechanism is guaranteed to be collusion-resistant. In our experiment, this means that α must be greater than 72. The results in Table 6.4 show that even using lower values for α, *e.g.* 50, the average joint share can be greater when the fixed agents are telling the truth than when they are colluding. However, we note that when $\alpha = 50$, this result is not statistically significant for a confidence level greater than 92%. On the other hand, for $\alpha \geq 100$, the results are statistically significant with a confidence level of 99%. An interesting point here is that when the value of α increases, the mechanism becomes stronger against collusions, *i.e.*, the loss by lying increases. This value seems to converge to 0.

Finally, looking at the results for the BTS mechanism, we notice that high values for α result in lower gains by colluding. Empirically, it seems that this mechanism can be collusion-resistant by using a very high value for α. For lower values of α, the results for the BTS mechanism are very similar to the results for the peer-evaluation mechanism. This happens because when α is small, then the sharing functions of these mechanisms become very similar (see Equation 3.1 and Equation 5.3).

Table 6.3: Results of the peer-evaluation mechanism with different values for α. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. Truthful evaluations are uniformly distributed. "Avg. 1" and "Std. 1" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they truthfully report their opinions. "Avg. 2" and "Std. 2" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they are colluding. "Unfair shares" and "Negative shares" are, respectively, the total number of unfair shares and the total number of negative shares that occurred throughout the simulation.

α	**Avg. 1**	**Std. 1**	**Avg. 2**	**Std. 2**	**Unfair shares**	**Negative shares**
0.1	667.72	43.74	675.38	42.44	0	0
1	667.72	43.74	675.38	42.44	0	0
5	667.72	43.74	675.38	42.44	0	0
10	667.72	43.74	675.38	42.44	0	0
25	667.72	43.74	675.38	42.44	0	0
50	667.72	43.74	675.38	42.44	0	0
100	667.72	43.74	675.38	42.44	0	0
500	667.72	43.74	675.38	42.44	0	0

Table 6.4: Results of the peer-prediction mechanism with different values for α. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. Truthful evaluations are uniformly distributed. "Avg. 1" and "Std. 1" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they truthfully report their opinions. "Avg. 2" and "Std. 2" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they are colluding. "Unfair shares" and "Negative shares" are, respectively, the total number of unfair shares and the total number of negative shares that occurred throughout the simulation.

α	**Avg. 1**	**Std. 1**	**Avg. 2**	**Std. 2**	**Unfair shares**	**Negative shares**
0.1	399.84	25.46	404.42	24.64	0	0
1	398.04	19.03	401.36	18.45	0	0
5	395.28	9.80	396.64	9.63	0	0
10	394.31	7.28	394.99	7.26	0	0
25	393.52	6.10	393.64	6.16	0	0
50	393.21	5.98	393.11	6.05	0	0
100	393.04	6.01	392.82	6.08	0	0
500	392.90	6.08	392.58	6.15	0	0

Table 6.5: Results of the BTS mechanism with different values for α. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$, $\epsilon = 0.01$. 100 simulations are used. Truthful evaluations are uniformly distributed. "Avg. 1" and "Std. 1" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they truthfully report their opinions. "Avg. 2" and "Std. 2" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they are colluding. "Unfair shares" and "Negative shares" are, respectively, the total number of unfair shares and the total number of negative shares that occurred throughout the simulation.

α	**Avg. 1**	**Std. 1**	**Avg. 2**	**Std. 2**	**Unfair shares**	**Negative shares**
0.1	667.72	43.74	675.38	42.44	0	0
1	667.71	43.73	675.36	42.43	0	0
5	667.65	43.66	675.30	42.39	0	0
10	667.59	43.59	675.22	42.35	0	0
25	667.38	43.46	674.98	42.30	0	0
50	667.04	43.46	674.59	42.46	0	0
100	666.36	44.33	673.80	43.65	0	0
500	660.89	78.92	667.46	79.75	0	0

Figure 6.7: Results of the peer-evaluation mechanism with different values for α when truthful evaluations are uniformly distributed. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. In the x-axis, there are different values for α. The y-axis contains the difference between the average joint share of two fixed agents when they are truthfully reporting their opinions and when they are colluding. The dotted line is used to facilitate visualization.

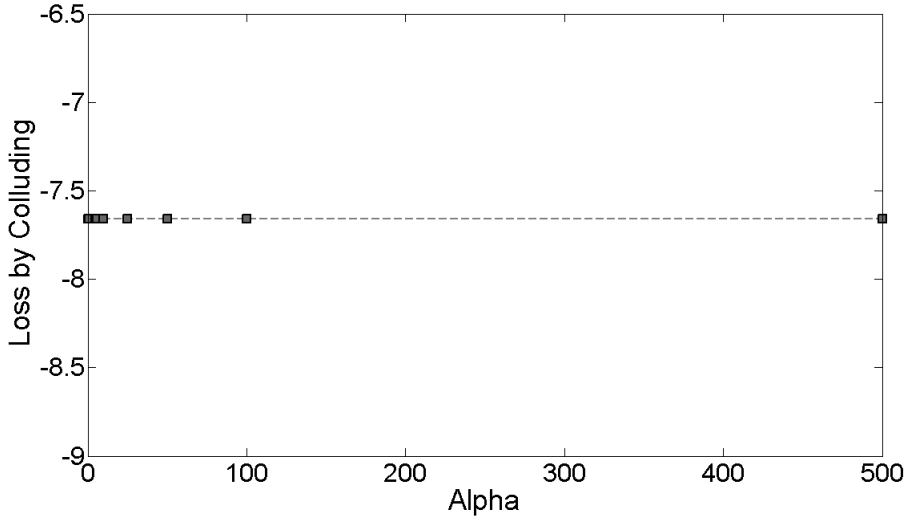


Figure 6.8: Results of the peer-prediction mechanism with different values for α when truthful evaluations are uniformly distributed. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. In the x-axis, there are different values for α. The y-axis contains the difference between the average joint share of two fixed agents when they are truthfully reporting their opinions and when they are colluding. The dotted line is used to facilitate visualization.

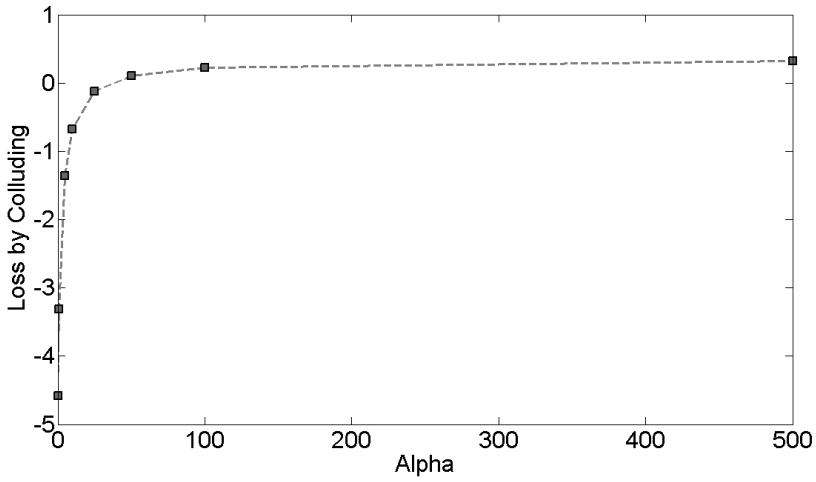


Figure 6.9: Results of the BTS mechanism with different values for α when truthful evaluations are uniformly distributed. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$, $\epsilon = 0.01$. 100 simulations are used. In the x-axis, there are different values for α. The y-axis contains the difference between the average joint share of two fixed agents when they are truthfully reporting their opinions and when they are colluding. The dotted line is used to facilitate visualization.

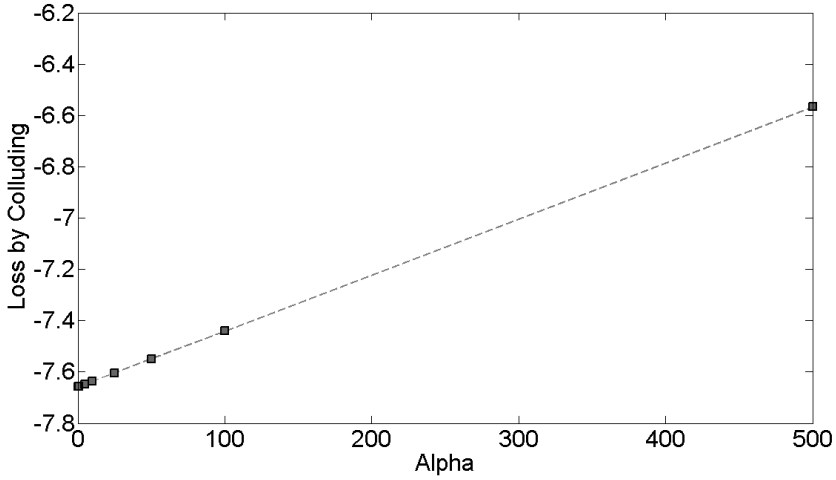


Table 6.6: The resulting p-values for the directional t-test. Our null hypothesis is that the average joint share when the fixed agents are telling the truth is equal to the average joint share when they are colluding. Our alternative hypothesis is that the average joint share when both agents are telling the truth is greater than the average joint share when they are colluding. Truthful evaluations are uniformly distributed.

α	**Peer-Evaluation**	**Peer-Prediction**	**BTS**
0.1	1	1	1
1	1	1	1
5	1	1	1
10	1	1	1
25	1	0.95	1
50	1	0.08	1
100	1	0.0016	1
500	1	< 0.0001	1

Moving to the results related to the fairness property, we start by noting that according to the theory we need to have $M < 5.38$ to guarantee that the peer-evaluation mechanism will always be fair (Theorem 6). Consequently, the peer-evaluation mechanism cannot return unfair shares in our experiment, because we use $M = 5$. The results shown in Table 6.3 are in agreement with the theory. For the other mechanisms, we need to have $\alpha < 0.5$ to guarantee that the peer-prediction mechanism will always be fair (Theorem 8), and we need both $M < 5.38$ and $\alpha < 0.12$ (Theorem 10) to guarantee that the BTS mechanism will always be fair. We observe that, even using much higher values for α than the ones recommended by the theory, none of these two mechanisms returned a single unfair share. Intuitively, the reason for this fact is that it is very unlikely, in a probabilistic sense, that an agent unanimously receives better evaluations than a peer when both the received evaluations are drawn from a uniform distribution and the underlying population of agents is fairly large, which is the setting of our experiment. Consequently, the fairness property holds (almost) trivially.

Analyzing the results related to the individual rationality property, we note that both the peer-evaluation mechanism and the peer-prediction mechanism are always individually rational, but we need to have $\alpha < 5.36$ to guarantee that the shares returned by the BTS mechanism will always be greater than zero (Proposition 6). From Table 6.5, we observe that even with much higher values for α, all the shares returned by the BTS mechanism in this experiment are greater than zero. Intuitively, the reason for this fact is that when the truthful evaluations are drawn from a uniform distribution, the resulting scores from the BTS method are in a relatively small range, since those evaluations are hardly very surprisingly common or very surprisingly uncommon. Thus, there exists negative truth-telling scores, but they are not large enough to create negative shares.

Trying to understand the influence of α on the proposed mechanisms in a (arguably) more realistic scenario, we repeat the experiment, but this time assuming that agents' truthful evaluations follow a normal distribution. In detail, for each agent in the population, we uniformly select a value inside the set $\{1, \ldots, M\}$, and use it as the mean of a normal distribution, which is employed to generate the truthful evaluations for that agent. We use a small variance, 0.625, and the same values for the other parameters as before: $V = 10000$, $n = 30$, $M = 5$, $\alpha \in \{0.1, 1, 5, 10, 25, 50, 100, 500\}$, and $\epsilon = 0.01$. We run this experiment 100 times. As before, we observe how different values of α influence the collusion-resistance, individual rationality and fairness properties of the proposed mechanisms. Table 6.7, 6.8, and 6.9 show the results of this new experiment. Figure 6.10, 6.11, and 6.12 visually show the results related to the collusion-resistance property. To show the statistical significance of the results obtained for the collusion-resistance property in this new scenario, we perform the directional t-test for correlated samples. Our null hypothesis is that the average joint share when the fixed agents are telling the truth is equal to the average joint share when they are colluding. Our alternative hypothesis is that the average joint share when both agents are telling the truth is greater than the average joint share when they are colluding. The resulting p-values can be seen in Table 6.10.

Focusing first on the results related to the collusion-resistance property, from Table 6.7 we observe that the average joint share returned by the peer-evaluation mechanism when the fixed agents are colluding is always greater than when they are telling the truth. This is quite obvious because this mechanism does use truth-telling scores.

Table 6.7: Results of the peer-evaluation mechanism with different values for α. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. Truthful evaluations are normally distributed. "Avg. 1" and "Std. 1" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they truthfully report their opinions. "Avg. 2" and "Std. 2" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they are colluding. "Unfair shares" and "Negative shares" are, respectively, the total number of unfair shares and the total number of negative shares that occurred throughout the simulation.

α	**Avg. 1**	**Std. 1**	**Avg. 2**	**Std. 2**	**Unfair shares**	**Negative shares**
0.1	677.97	171.95	682.68	169.21	0	0
1	677.97	171.95	682.68	169.21	0	0
5	677.97	171.95	682.68	169.21	0	0
10	677.97	171.95	682.68	169.21	0	0
25	677.97	171.95	682.68	169.21	0	0
50	677.97	171.95	682.68	169.21	0	0
100	677.97	171.95	682.68	169.21	0	0
500	677.97	171.95	682.68	169.21	0	0

Table 6.8: Results of the peer-prediction mechanism with different values for α. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. Truthful evaluations are normally distributed. "Avg. 1" and "Std. 1" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they truthfully report their opinions. "Avg. 2" and "Std. 2" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they are colluding. "Unfair shares" and "Negative shares" are, respectively, the total number of unfair shares and the total number of negative shares that occurred throughout the simulation.

α	**Avg. 1**	**Std. 1**	**Avg. 2**	**Std. 2**	**Unfair shares**	**Negative shares**
0.1	453.16	108.67	456.32	106.71	0	0
1	446.41	80.96	448.62	79.56	0	0
5	436.00	38.59	436.76	38.06	0	0
10	432.36	24.19	432.61	23.96	0	0
25	429.39	13.43	429.22	13.44	26	0
50	428.20	10.17	427.87	10.24	660	0
100	427.57	9.03	427.15	9.10	2114	0
500	427.04	8.60	426.54	8.64	3843	0

Table 6.9: Results of the BTS mechanism with different values for α. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$, $\epsilon = 0.01$. 100 simulations are used. Truthful evaluations are normally distributed. "Avg. 1" and "Std. 1" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they truthfully report their opinions. "Avg. 2" and "Std. 2" are, respectively, the average and the standard deviation of the joint share of two fixed agents when they are colluding. "Unfair shares" and "Negative shares" are, respectively, the total number of unfair shares and the total number of negative shares that occurred throughout the simulation.

α	**Avg. 1**	**Std. 1**	**Avg. 2**	**Std. 2**	**Unfair shares**	**Negative shares**
0.1	677.97	171.95	682.68	169.21	0	0
1	678.02	171.95	682.70	169.24	0	0
5	678.23	171.95	682.79	169.33	0	0
10	678.49	171.97	682.90	169.47	0	0
25	679.28	172.09	683.22	169.95	0	0
50	680.59	172.57	683.76	171.04	0	0
100	683.22	174.51	684.84	174.23	0	0
500	704.23	228.76	693.50	237.67	172	25

Figure 6.10: Results of the peer-evaluation mechanism with different values for α when truthful evaluations are normally distributed. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. In the x-axis, there are different values for α. The y-axis contains the difference between the average joint share of two fixed agents when they are truthfully reporting their opinions and when they are colluding. The dotted line is used to facilitate visualization.

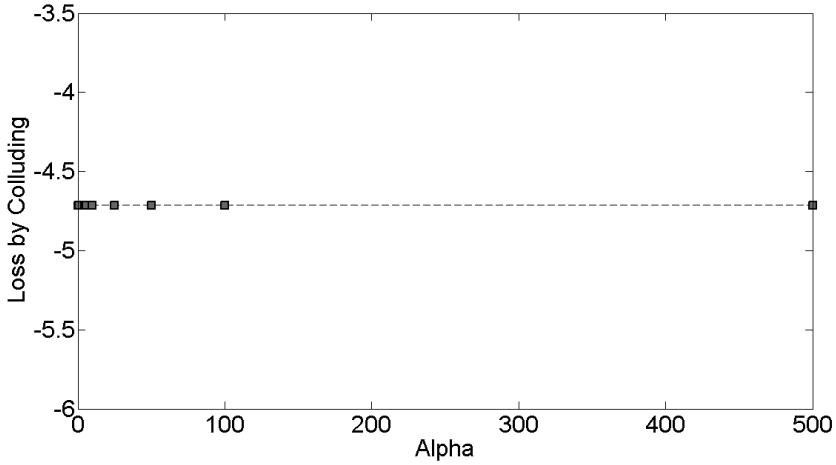


Figure 6.11: Results of the peer-prediction mechanism with different values for α when truthful evaluations are normally distributed. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$. 100 simulations are used. In the x-axis, there are different values for α. The y-axis contains the difference between the average joint share of two fixed agents when they are truthfully reporting their opinions and when they are colluding. The dotted line is used to facilitate visualization.

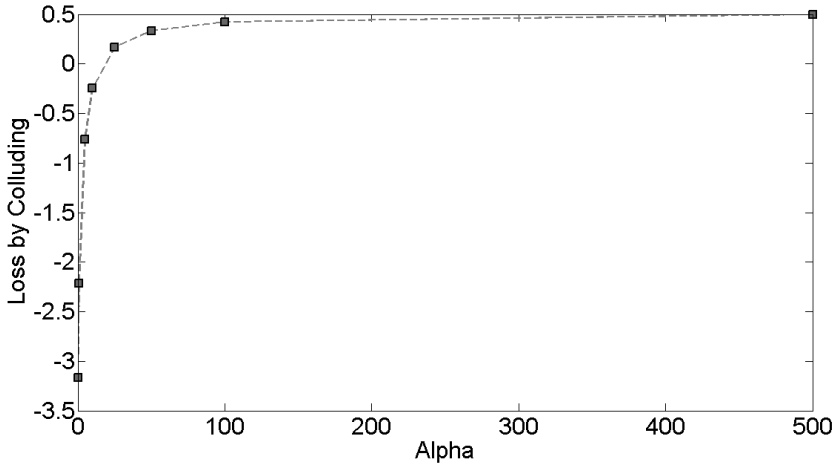


Figure 6.12: Results of the BTS mechanism with different values for α when truthful evaluations are normally distributed. The other parameters are fixed: $n = 30$, $V = 10000$, $M = 5$, $\epsilon = 0.01$. 100 simulations are used. In the x-axis, there are different values for α. The y-axis contains the difference between the average joint share of two fixed agents when they are truthfully reporting their opinions and when they are colluding. The dotted line is used to facilitate visualization.

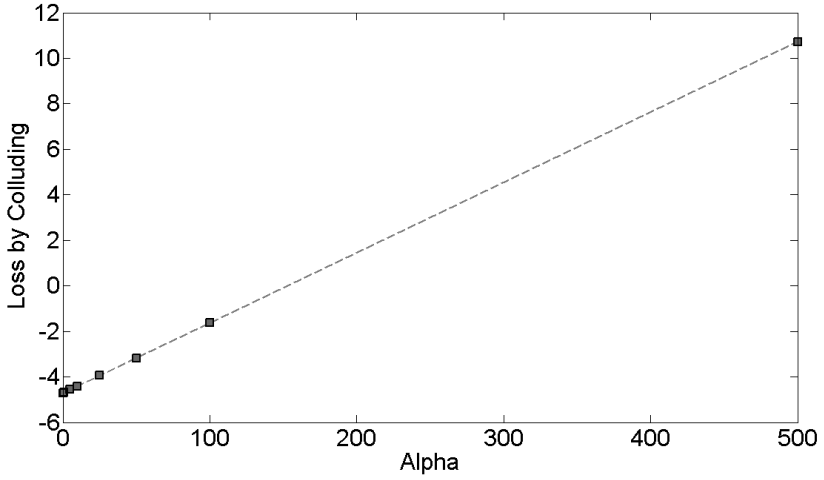


Table 6.10: The resulting p-values for the directional t-test. Our null hypothesis is that the average joint share when the fixed agents are telling the truth is equal to the average joint share when they are colluding. Our alternative hypothesis is that the average joint share when both agents are telling the truth is greater than the average joint share when they are colluding. Truthful evaluations are normally distributed.

α	**Peer-Evaluation**	**Peer-Prediction**	**BTS**
0.1	1	1	1
1	1	1	1
5	1	1	1
10	1	0.9955	1
25	1	0.0362	1
50	1	0.0004	1
100	1	< 0.0001	1
500	1	< 0.0001	< 0.0001

Moving to the results for the peer-prediction mechanism, Table 6.8 reinforces that by using lower values for α than the ones recommended by the theory, it is still possible to obtain collusion-resistance. For example, when $\alpha = 25$ this fact occurs, and it is statistically

significant with a confidence level of 96%. For $\alpha \geq 50$, this happens again and the results are statistically significant even with a confidence level of 99.9%. These results are a way better than the results when truthful evaluations are uniformly distributed. Intuitively, the reason for that is because when evaluations are normally distributed, if the mean of the distribution of the evaluations received by an agent is small, then a liar agent will probably not receive a high score when it lies by increasing its evaluation to the top possible one. Moreover, when the value of α increases, the mechanism becomes stronger against collusions, *i.e.*, the loss by lying increases. This value seems to converge to 0.5.

Finally, looking at the results from the BTS mechanism, we remember that in our previous experiment high values for α resulted in lower gains by colluding. This is not only true now, but a very high value for α can actually make the BTS collusion-resistant. For example, when $\alpha = 500$ this happens and this result is statistically significant with a confidence level of 99.99%. Thus, the evidence supporting the hypothesis that a high value of α can help prevent collusion in the BTS mechanism is strong.

Analyzing the results related to the fairness property, the trade-off between fairness and collusion-resistance is visible. When α increases, the peer-prediction and the BTS mechanisms become stronger against collusions, but the number of unfair shares also increases. Intuitively, this happens because these mechanisms are putting more weight on the truth-telling scores rather than on the aggregated evaluations. The peer-evaluation mechanism cannot return any unfair share since $M < 5.38$ (see Theorem 6).

A similar trade-off also exists between the collusion-resistance and the individually rationality properties for the BTS mechanism. Since the scores returned by the BTS method can be negative, the resulting shares from the BTS mechanism can be negative when α increases, as shown in Table 6.9. We note that according to the theory, α must be less than 5.36 to mathematically ensure that the BTS mechanism is individually rational (Theorem 10). However, this experiment shows that in practice we can use much higher values for α, and the mechanism still does not return negative shares. The peer-evaluation and the peer-prediction mechanisms are always individually-rational.

Summarizing, the results of the experiments presented in this section show that in practice we can use values for α different than the ones recommended by the theory and still get desirable properties. This was already expected because most of the propositions and theorems are proved with the worst-case scenario in mind. Also, we show that practice agrees with theory. We empirically demonstrate that collusions can be successfully avoided by fine-tuning α, and that when the value of this parameter increases, the number of unfair shares returned by the peer-prediction and BTS mechanisms also increases, as well as the number of negative shares returned by the BTS mechanism.

6.3 Parameter n

Our last experiment investigates how the mechanisms behave when dealing with populations of different sizes. In detail, we study how the size of the population affects the budget of the mechanisms. From the theory, the peer-prediction mechanism is weakly budget-balanced. Here, we empirically study what happens to the profit of this mechanism when the population grows. Further, the BTS mechanism is budget-balanced under some extra assumptions. One of these assumptions is that the underlying population is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations. In Chapter 5, we argued that more important than having a very large population of agents is a guarantee that each empirical distribution of received evaluations is balanced, *i.e.*, that there exists a significant number of endorsements for each available evaluation. In this section, we empirically verify this claim. Despite the fact that the peer-evaluation mechanism is budget-balanced without extra assumptions, we include it in our experiment for completeness' sake.

For each proposed mechanism, we share the reward $V = 10000$ using the parameter $M = 10$, and populations with sizes $5, 7, 10, 25, 50, 100$. We deliberately choose high values for α to make the resulting shares more diverse. For the peer-prediction mechanism, we set $\alpha = 10$, and for the BTS mechanism we set $\alpha = 100$. We use the recalibration factor $\epsilon = 0.01$. In this experiment, agents always report their observed signals. These signals are drawn from a uniform distribution. For generating random predictions, we use the algorithm proposed by Stafford [43]. We execute this experiment 100 times. In each execution, we compute the sum of the returned shares. At the end of the experiment, we compute the average and the standard deviation of these sums. Table 6.11 shows these results. Figure 6.13, 6.14 and 6.15 visually present the same results. To find out whether the sum of the resulting shares from each mechanism have a common mean, we perform the statistical test ANOVA. Table 6.12 shows the resulting p-values.

Analyzing the results of this experiment, we note that, as expected, the size of the population does not influence the budget-balancedness property of the peer-evaluation mechanism. Surprisingly, it also does not significantly affect the profit of the peer-prediction mechanism. Intuitively, a mechanism should spend more when there are a lot of agents than when there are only few of them. However, the shares returned by the peer-prediction mechanism are scaled by using the constant $\frac{V}{(M+2\alpha)n}$ (Equation 4.3), which means that the greater the value of n, the smaller the share of the agents. In this way, this constant seems to nullify any effect caused by a large population. The interesting point here is that the standard deviation of the sum of agents' shares appears to converge to zero when the population size increases.

The most interesting result comes from the BTS mechanism. As n comes closer to $M = 10$, the loss made by this mechanism gradually increases. However, for n greater

than M, this loss gradually reduces, and it seems to converge to zero for a very large population, which agrees with the theory. When $n < M$, the resulting truth-telling scores are high. Since there are few agents to endorse a larger number of possible evaluations, the reported evaluations are very often surprisingly common. On the other hand, when $n > M$, the scores of the agents are more balanced, since there are more agents to endorse fewer evaluations, and, consequently, a reported evaluation is less likely to be surprisingly common. The standard deviation of the sum of the shares also decreases when n increases, thus supporting our claim that the scores are more balanced. Finally, the result of the ANOVA test shows, with a confidence level of 99.99%, that the sums of the shares do not have a common mean, thus suggesting that n indeed influences the resulting shares.

Summarizing, different population sizes have a negligible impact on the budget of the peer-evaluation and the peer-prediction mechanisms when truthful evaluations are normally distributed. On the other hand, the BTS mechanism works much better with large populations, since it may make a loss when dealing with small populations. Our rule of thumb presented in chapter 5, *i.e.*, to set $M \leq \sqrt{n-2}$, has a strong empirical support here, since at this point the population seems to be large enough, in relation to the number of available evaluations, so that a possible loss made by the mechanism is negligible. Clearly, this is not a sufficient condition, since a balanced distribution of the empirical evaluations is also important.

Table 6.11: Results of the proposed mechanisms with different values for n. Truthful evaluations are uniformly distributed. The other parameters are fixed: $V = 10000$, $M = 10$. 100 simulations are used. "Average" and "Std" are, respectively, the average and the standard deviation of the sums of the returned shares.

	Peer-Evaluation		**Peer-Prediction**		**BTS**	
n	Average	Std.	Average	Std.	Average	Std.
5	10000	0	5510.74	213.22	10019.00	1.42
7	10000	0	5471.22	143.77	10020.64	1.50
10	10000	0	5474.86	100.67	10021.48	1.76
25	10000	0	5476.56	39.40	10015.78	1.64
50	10000	0	5475.80	20.73	10010.92	1.11
100	10000	0	5473.03	10.15	10009.61	0.52

Table 6.12: The resulting p-values from the ANOVA test.

Peer-Evaluation	**Peer-Prediction**	**BTS**
1	0.13	< 0.0001

Figure 6.13: Results of the peer-evaluation mechanism with different values for n. The other parameters are fixed: $V = 10000$, $M = 10$. 100 simulations are used. The averages of the sums of the returned shares are represented by the black squares and the standard deviations by the gray lines. The dotted line is used to facilitate visualization.

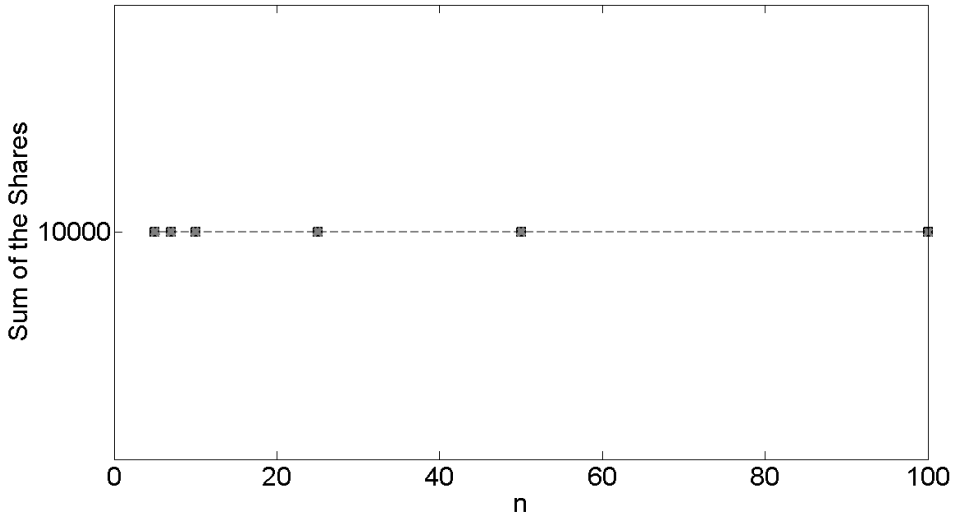


Figure 6.14: Results of the peer-prediction mechanism with different values for n. The other parameters are fixed: $V = 10000$, $M = 10$, $\alpha = 10$. 100 simulations are used. The averages of the sums of the returned shares are represented by the black squares and the standard deviations by the gray lines. The dotted line is used to facilitate visualization.

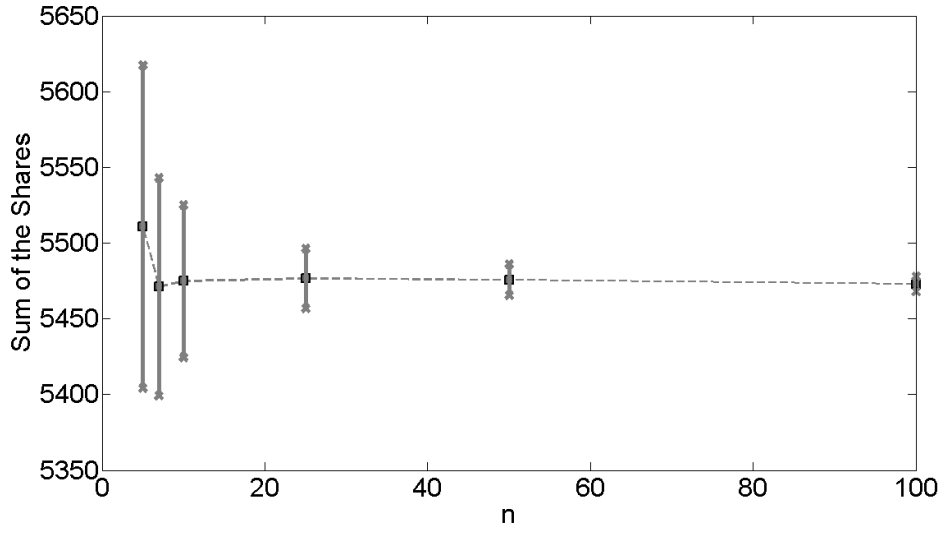
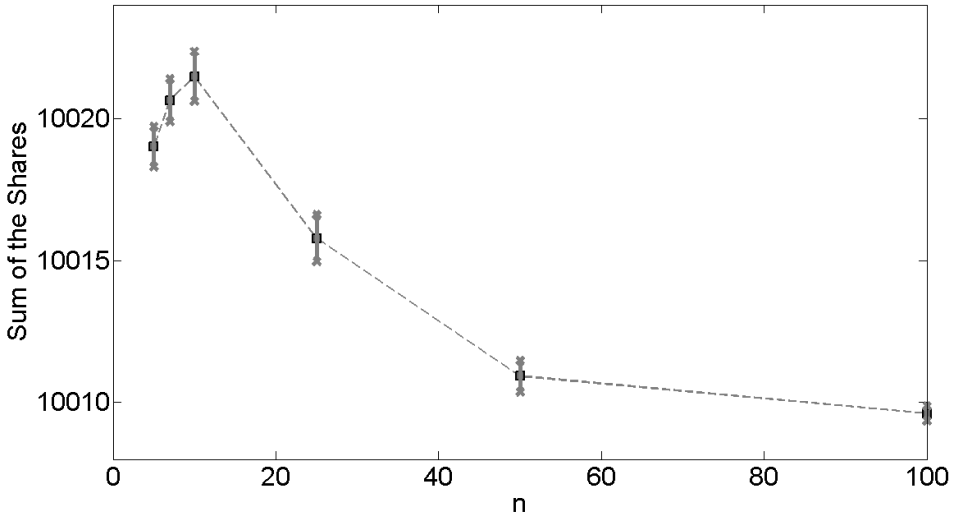


Figure 6.15: Results of the BTS mechanism with different values for n. The other parameters are fixed: $V = 10000$, $M = 10$, $\alpha = 100$, $\epsilon = 0.01$. 100 simulations are used. The averages of the sums of the returned shares are represented by the black squares and the standard deviations by the gray lines. The dotted line is used to facilitate visualization.



6.4 Concluding Remarks

In this section, we studied the influence of the mechanisms' parameters on the agents' shares. Firstly, we observed that the top possible evaluation that an agent can give or receive, M, does not seem to affect the range and standard deviation of the resulting shares from all mechanisms when truthful evaluations are uniformly distributed. However, when the truthful evaluations for a peer are normally distributed, the range and standard deviation of the shares returned by the peer-prediction and BTS mechanisms significantly increase as M increases. In this last scenario, the difference between agents are better recognized and specified by their peers, resulting in fine-grained differences in the final shares.

Related to the parameter α, which fine-tunes the weight given to the truth-telling scores, our experiments agreed with what was expected from the theory. For the peer-prediction and BTS mechanisms, we observed the trade-off between fairness and collusion-resistance, *i.e.*, high values for α make the mechanism collusion-resistant, but this also increases the number of unfair shares. Our experiments also showed that in practice we can use values for α different than the ones recommended by the theory and still get desirable properties. This happens because most of our theorems and propositions are proved based on the

worst-case scenario, which is very unlikely to happen in real applications.

Finally, we studied how the size of the population affects the budget of the mechanisms. When truthful evaluations are uniformly distributed, our experiments showed that the population size does not affect the profit of the peer-prediction mechanism. For the BTS mechanism, small populations implied losses for this mechanism. Our rule of thumb, to set $M \leq \sqrt{n-2}$, had strong empirical support here, since at this point the population seemed to be large enough, in relation to the number of available evaluations, so that a possible loss made by the mechanism was negligible.

We end this chapter by noting that all the random evaluations used in our experiments were either normally or uniformly distributed, and that the random predictions were always uniformly distributed [43]. Other interesting experiments to pursue, that are left as future work, are to understand the influence of the mechanisms' parameters on the agents' shares when both evaluations and predictions are normally distributed, and when both have a U-shape, since these scenarios are arguably closer to real ones.

Chapter 7

Related Work

Understanding how agents can work together in order to achieve some common goal is a central topic in the field of multiagent systems [42]. Questions that are typically analyzed include how and which groups of agents should form [36], how agents should coordinate their actions once they have agreed to work together [19], how to ensure that the group, once formed, does not disintegrate [7], and how any joint rewards (or costs) should be divided among the group members [30]. It is this last question that we addressed in this thesis. Commonly called fair division, the problem of dividing one or several goods among a set of agents, in a way that satisfies a suitable fairness criterion, has been studied in several literatures.

In economics, the collective welfare approach is perhaps the most influential application of the economic analysis to fair division. It uses the concepts of collective utility functions, in its cardinal version, and of social welfare orderings, in its ordinal version, for deciding what makes a reasonable allocation. Moulin [30] examines the contribution of this modern microeconomic thinking to fair division. In detail, he compares normative arguments of fair division and their relation to efficiency and collective welfare from economics.

In computer science, the fair division problem is usually studied in settings where the underlying agents not only have preferences over alternative allocations of goods, but also actively participate in computing an allocation. Chevalyere *et al.* [5] gives an overview of allocation procedures for indivisible goods, applications, preference languages, and complexity results related to those settings.

In this chapter, we review the literature most closely related to our work, pointing out the differences and similarities to our model and mechanisms. In detail, we review similar ideas from cooperative game theory, cake-cutting, and mechanism design literature.

7.1 Cooperative Game Theory

Cooperative game theory, also known as coalitional game theory, deals with the question of how self-interested agents can combine to form effective teams [31]. The canonical model of coalitional games with transferable utility assumes that there exists a characteristic function mapping each subset of agents to a payoff, which can be freely distributed among the coalition's members. The main difference between our model and the canonical model of coalitional games with transferable utility is that we do not use a characteristic function. In detail, we assume that the grand coalition, *i.e.*, the coalition of all the agents in N, is always formed, and its payoff is equal to V, *i.e.*, the reward to be shared.

The Shapley value [41] is a key concept used in cooperative game theory to distribute a joint surplus (or cost) among a set of agents in the context of production. Roughly speaking, the Shapley value assigns a share to each agent equal to its expected marginal contribution with respect to a uniform distribution over the set of all permutations on the set of agents. The Shapley value is remarkable not only for its attractive and intuitive definition, but also for its unique characterization by a set of reasonable axioms.

We note that a sharing scheme where the fairness criterion is based on marginal contributions, like the Shapley value, is not appropriate in our setting. The idea of marginal contributions is not objectively defined in our model, since the only way to determine individual contributions is through subjective opinions.

7.2 Cake-Cutting

Cake-cutting is a common metaphor for the sharing of a heterogeneous divisible resource. A cake is usually represented by the interval [0, 1]. Each agent has an additive utility function that assigns a value to every given piece of cake. The goal is to find a partition of the cake among the agents that satisfies some fairness criteria. The two most famous fairness criteria are *proportional allocation* and *envy-freeness* [3].

In a proportional allocation, the value that each agent has for its own piece of cake is at least $1/n$ of the value that it assigns to the entire cake. In an envy-free allocation, the value that each agent assigns to its own piece of cake is at least as high as the value that it assigns to any other agent's piece of cake. There is a vast body of literature on fairly cutting a cake according to these two criteria. Solid explanations of them are given by Brams and Taylor [3].

While most of the work in artificial intelligence and theoretical computer science has focused on the allocation of indivisible resources (*e.g.*, combinatorial auctions), recent years have seen an increasing interest among computers scientists in the allocation of

heterogeneous divisible resources. Questions that are typically analyzed include cake-cutting algorithms that approximate fairness [16], and bounds on the number of cuts required to fairly divide the cake [17, 35]. A compilation of cake-cutting algorithms is given by Robertson and Webb [37].

There are many points in which our work diverges from the cake-cutting literature. First, we are not dealing with a heterogeneous resource. In our model, every agent is assumed to want more of the reward. The same is not true in the canonical cake-cutting model, since there may exist piece of cakes that are not desired by some agent. Further, both the proportional allocation and the envy-free fairness criteria are pointless in the scenario studied here. There will always be an envious agent when at least two shares are different. Further, whenever all agents receive at least $1/n$ of the reward, then either the underlying mechanism is not budget-balanced or it is an egalitarian mechanism (see definition in Section 2.1). We believe that both properties are very often undesirable.

We end this section by noting that a new trend in the cake-cutting field is to design sharing schemes that are both fair and truthful. A cake-cutting algorithm is truthful if when an agent lies, then it is allocated a piece of cake that is worth, according to its real valuation, no more than the piece of cake that would be allocated if that agent had told the truth [1, 2, 44, 4]. Differently from our work, a trade-off between fairness and truthfulness does not seem to exist in such sharing schemes.

7.3 Mechanism Design

Roughly speaking, a mechanism is a protocol with specific rules to ensure that truth-telling produces a desirable outcome for the agents [27]. Seminal results from the mechanism design literature show that it is generally possible to use transfer payments to extract agents' private information. A classical example is the Vickrey-Clarke-Grove (VCG) mechanism, where each agent maximizes its expected utility by truthfully reporting its private information [45, 6, 20].

Similar to the approach used by the peer-prediction mechanism to incentivise truthfulness, some work has suggested transfer payments based on proper scoring rules. For example, Johnson *et al.* [23] show how to construct transfer payments, using proper scoring rules, that exploit correlation in agents' private information. Johnson *et al.* [22] extend those results to the case of multidimensional, continuous private information.

With the advent of the Internet, an increasing number of applications of mechanism design emerged to extract knowledge from groups of agents. For example, Jurca and Faltings [25] propose a class of mechanisms to incentivise agents to truthfully report their experience-related information in online feedback forums. In a quite different scenario, Ju-

rca and Falting [24] propose a mechanism that extracts accurate information from rational agents in online opinion polls.

The main difference between our work and the aforementioned works is that our primary objective is to share a joint reward among a set of agents, and not to incentivise truthfulness. We note that if agents always reported their opinions truthfully, then it would not be necessary to incentivise truth-telling in our setting. The similarities between our work and the mechanism design field, as a whole, only show up when we consider that agents are strategic, *i.e.*, when they are able to manipulate their opinions for increasing their shares of the reward.

Chapter 8

Conclusion and Future Work

In this thesis, we studied how to share a joint reward among a set of agents when the individual contributions are subjective. To the best of our knowledge, this was the first attempt to formalize this relatively common scenario. In our proposed game-theoretic model, agents are asked to provide opinions about the performance of their peers in accomplishing a task, for which it is granted a reward. These opinions are elicited and aggregated by a mechanism, which is also responsible for sharing the reward. We considered two kinds of opinions: evaluations and predictions

Since the most prominent fairness criteria are not suitable for our setting, *e.g.*, marginal contribution, proportional allocation and envy-freeness, we proposed a more appropriate fairness criterion, which essentially means that if an agent unanimously receives better evaluations than a peer, then this agent should also receive a greater share of the reward than that peer. We also proposed a collusion model in which a liar agent agrees to misreport its evaluations in exchange for a side-payment from the agent who benefits from the misreporting so that both agents end up with a greater expected share than if no collusion had occurred.

Besides the game-theoretic model for sharing a reward based on subjective opinions, the new fairness criterion and the collusion model, we proposed three different mechanisms to elicit and aggregate opinions, as well as for determining agents' shares, keeping the issues of truthfulness and fairness in mind. Our first mechanism, the peer-evaluation mechanism, divides the reward proportionally to the evaluations received by the agents. We showed that this mechanism is fair, budget-balanced, individually rational, and strategy-proof, but that it can be collusion-prone.

Our second mechanism, the peer-prediction mechanism, shares the reward by considering the evaluations received by the agents and their truth-telling scores, which are computed by using a proper scoring rule. Under the assumption that agents are Bayesian

decision-makers, we showed that this mechanism is weakly budget-balanced, individually rational, and incentive-compatible. Further, we presented approaches that guarantee this mechanism to be always collusion-resistant and fair.

Our last mechanism, the BTS mechanism, elicits both evaluations and predictions from agents. It considers the evaluations received by the agents and their truth-telling scores when sharing the reward. For computing the scores, it uses the Bayesian truth serum method. Under the assumptions that agents are Bayesian decision-makers, and that the population of agents is sufficiently large so that a single evaluation cannot significantly affect the empirical distribution of evaluations, we showed that this mechanism is incentive-compatible and budget-balanced. Further, we presented approaches that guarantee this mechanism to be always individually rational and fair.

A natural question to ask is when each mechanism should be used. A general rule of thumb here is to only use the peer-evaluation mechanism when there is a guarantee that the underlying agents will always tell the truth, *e.g.*, when they are softwares specifically programmed to do that. The peer-prediction mechanism should be used when the budget-balance property is not strongly desired. Finally, the BTS mechanism should be used when the underlying population is relatively large in comparison to the top possible evaluation that an agent can give or receive, *i.e.*, when $M \leq \sqrt{n-2}$, and when providing predictions is not a too heavy burden for the agents.

8.1 Future Work

This work opens up new exciting directions for future work. In this section, we outline some of them.

8.1.1 Improving The BTS Mechanism

Prelec and Seung [33] propose a very promising algorithm in which scores from the BTS method are used to find the truthful answer to a question, even when subjective opinions remain the only source of evidence and there is a possibility that most agents are providing wrong answers. The algorithm essentially works by finding agents with more accurate metaknowledge. Under some assumptions, the answer provided by those agents converges to the truthful answer. The less informed agents, in a metaknowledge sense, do not disturb the outcome, but subsidize those who are more informed.

A similar idea can be directly applied to extend the BTS mechanism. Originally, this mechanism uses the BTS method exclusively to incentivise truthfulness. However, if the evaluations of the agents with more accurate metaknowledge have greater influence on the

sharing process, we can then find the "truthful shares", even if most agents are providing "wrong evaluations". In other words, we can adjust the BTS mechanism for the fact that some agents are better informed than others. We note that the idea of using metaknowledge as an indicative of domain expertise is purely theoretical. To the best of our knowledge, the role of metaknowledge has not been thoroughly discussed in the expertise literature. We redirect the interested reader to Shanteau's work [40] for further discussion.

8.1.2 Collusion Model

Our collusion model is fairly narrow because it is defined considering only two agents, a liar and a beneficiary. Another interesting research direction is to model different kinds of collusions, and to propose new schemes for making our mechanisms collusion-resistant under these new models.

8.1.3 Exploiting Correlation

In some practical applications, we expect the truthful evaluations for a specific agent to be correlated or, more formally, *stochastically relevant* [22]. A very promising direction for future work is to design a mechanism that considers that the observed signals are dependent. To promote truthfulness in this scenario, we can use seminal results from the mechanism design literature that show that it is generally possible to exploit correlation in agents' private information to induce truthful reporting [12, 13, 10, 11, 28].

8.1.4 Real Applications

It would be interesting, for future work, to experimentally validate our model and mechanisms. There are several possible scenarios in which we can apply the ideas proposed in this thesis, for example cooperative organizations and academic group work.

Bibliography

[1] Steven J. Brams, Michael A. Jones, and Christian Klamler. Better ways to cut a cake. *Notices of the AMS*, 53(11):1314–1321, 2006.

[2] Steven J. Brams, Michael A. Jones, and Christian Klamler. Proportional pie-cutting. *International Journal of Game Theory*, 36(3):353–367, 2008.

[3] Steven J. Brams and Alan D. Taylor. *Fair division: From cake-cutting to dispute resolution*. Cambridge University Press, 1996.

[4] Yiling Chen, John K. Lai, David C. Parkes, and Ariel D. Procaccia. Truth, justice and cake cutting. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (to appear)*, 2010.

[5] Yann Chevaleyre, Paul E. Dunne, Ulle Endriss, Jérôme Lang, Michel Lemaître, Nicolas Maudet, Julian Padget, Steve Phelps, Juan A. Rodríguez Aguilar, and Paulo Sousa. Issues in multiagent resource allocation. *Informatica*, 30:3–31, 2006.

[6] Edward H. Clarke. Multipart pricing of public goods. *Public Choice*, 11(1):17–33, 1971.

[7] Vincent Conitzer and Tuomas Sandholm. Complexity of constructing solutions in the core based on synergies among coalitions. *Artificial Intelligence*, 170:607–619, 2006.

[8] Karen S. Cook and Karen A. Hegtvedt. Distributive justice, equity, and equality. *Annual Review of Sociology*, 9:217–241, 1983.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006.

[10] Jacques Cremer and Richard P. McLean. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica*, 53(2):345–61, March 1985.

[11] Jacques Cremer and Richard P. McLean. Full extraction of surplus in bayesian and dominant strategy auctions. *Econometrica*, 56(6):1247–1257, 1988.

[12] Claude d'Aspremont and Louis-Andre Gerard-Varet. Incentives and incomplete information. *Journal of Public Economics*, 11(1):25–45, February 1979.

[13] Claude d'Aspremont and Louis-Andre Gerard-Varet. Bayesian incentive compatible beliefs. *Journal of Mathematical Economics*, 10(1):83–103, June 1982.

[14] Morris H. DeGroot. *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

[15] Morton Deutsch. Equity, equality, and need: What determines which value will be used as the basis of distributive justice. *Journal of Social Issues*, 3(31):137–149, 1975.

[16] Jeff Edmonds and Kirk Pruhs. Balanced allocations of cake. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 623–634, 2006.

[17] Jeff Edmonds and Kirk Pruhs. Cake cutting really is not a piece of cake. In *Proceedings of the 17th annual ACM-SIAM symposium on Discrete algorithm*, pages 271–278, 2006.

[18] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2 edition, 2003.

[19] Barbara J. Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86:269–357, 1996.

[20] Theodore Groves. Incentives in teams. *Econometrica*, 41(4):617–631, July 1973.

[21] John C. Harsanyi. Games with incomplete information played by "Bayesian" players, parts I-III. *Management Science*, 14:159–182, 320–334, 486–502, 1967.

[22] Scott Johnson, Nolan Miller, John W. Pratt, and Richard J. Zeckhauser. Efficient Design with Interdependent Valuations and an Informed Center. Working paper, Kennedy School, RWP02-025, 2002.

[23] Scott Johnson, John W. Pratt, and Richard J. Zeckhauser. Efficiency despite mutually payoff-relevant private information: The finite case. *Econometrica*, 58(4):873–900, July 1990.

[24] Radu Jurca and Boi Faltings. Incentives for expressing opinions in online polls. In *Proceeddings of the 2008 ACM Conference on Electronic Commerce*, pages 119–128. ACM, July 2008.

[25] Radu Jurca and Boi Faltings. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research*, 34:209–253, 2009.

[26] Gary Marks and Norman Miller. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1):72–90, 1987.

[27] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[28] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

[29] H. Moulin and S. Shenker. Strategyproof sharing of submodular costs: Budget balance versus efficiency. *Economic Theory*, 18:511–533, 2001.

[30] Hervé Moulin. *Fair Division and Collective Welfare*. The MIT Press, 2004.

[31] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

[32] D. Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, October 2004.

[33] Drazen Prelec and H. Sebastian Seung. An algorithm that finds truth even if most people are wrong. Working paper, Massachusetts Institute of Technology, 2007.

[34] Drazen Prelec and Ray Weaver. Truthful answers are surprisingly common: Experimental tests of the bayesian truth serum. Working paper, Massachusetts Institute of Technology, 2009.

[35] Ariel D. Procaccia. Thou shalt covet thy neighbor's cake. In *Proceedings of the 21st international joint conference on Artificial intelligence*, pages 239–244, 2009.

[36] Talal Rahwan, Savapali Ramchurn, Nicholas Jennings, and Andrea Giovannucci. An anytime algorithm for optimal coalition structure generation. *Journal of Artificial Intelligence Research*, 34:521–567, 2009.

[37] Jack Robertson and William Webb. *Cake Cutting Algorithms: Be Fair If You Can*. A. K. Peters, 1998.

[38] Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[39] Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–61, June 1998.

[40] James Shanteau. Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, 53(2):252–266, 1992.

[41] Lloyd S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, 1953.

[42] Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2009.

[43] Roger Stafford. Random vectors with fixed sum, 2006. Available at: http://www.mathworks.com/matlabcentral/fileexchange/9700.

[44] William Thomson. Children crying at birthday parties. why? *Journal of Economic Theory*, 31:501–521, 2007.

[45] William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1):8–37, 1961.

[46] Rebecca J. Weiss. Optimally aggregating elicited expertise: a proposed application of the bayesian truth serum for policy analysis. Master's thesis, Engineering Systems Division, Technology and Policy Program, Massachusetts Institute of Technology, 2009.

[47] Robert L. Winkler. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 64(327):1073–1078, 1969.