Turning Bribes into Lemons: an optimal mechanism*

Christopher Stapenhurst† and Andrew Clausen‡

January 19, 2026

Abstract

Corruption requires a coalition to form and reach an agreement. Is there a cheap way to stop any agreement from being reached? We find an optimal mechanism that resembles Poker. The players' hands are synthetic asymmetric information, and they create a lemons problem in the market for bribes. Our Poker mechanism is robust: it thwarts bribes regardless of the negotiation procedure, including alternating offers bargaining, Dutch auctions and arbitration. In compliance settings, there is a trade-off between rewarding the agent for honesty and punishing him for non-compliance. This trade-off is resolved by rigging the Poker hand distribution against the agent and in favour of the monitor. Finally, the cost of deterring bribes is inversely proportional to the number of monitors.

Keywords: Corruption, adverse selection, screening, mechanism design, information design.
JEL: D73, D82, D86.

*We are indebted to John H. Moore, Ina Taneva, and Gabriel Ziegler for their generous support. We also thank Ramses Abul Naga, Helmuts Azacis, Peter Eso, Aggelos Kiayias, Thomas Mariotti, Eric Maskin, Tatiana Matskaya, Alfonso Montes, Humberto Moreira, Mariann Ollar, Martin Pawalczyk, Miklos Pinter, Robert Somogyi, Alessandro Spiganti, Roland Strausz, Tomasz Sulka, Balász Szentes, Jonathan Thomas, Péter Vida, Raphael Veiel, Philip Wadler, Rakesh Vohra, James Wiley, Andriy Zapechelnyuk, and Min Zhang, and seminar participants at Humboldt U. of Berlin, CERGE-EI, Corvinus U. of Budapest, Glasgow, Graz, HSE Moscow, Málaga, Vienna and participants at the Conference on Economic Design (2025), Lisbon Meetings in Game Theory and Applications (2025), and Conference on Mechanism and Institution Design (2024) for their helpful comments.

†Brandenburg University of Technology and Budapest University of Technology and Economics. christopher.stapenhurst@b-tu.de. Christopher Stapenhurst acknowledges funding from the Sustainable Development and Technologies National Programme of the Hungarian Academy of Sciences (FFT NP FTA), and from project no. EKÖP-24-4-II-BME-37 implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2024-2.1.1-EKÖP-2024-00003 funding scheme.

‡University of Edinburgh. andrew.clausen@ed.ac.uk.

1 Introduction

Across many domains, insiders and experts claim that institutions are vulnerable to ingenious corrupt schemes that are profitable and elude scrutiny. Butler (1935) and Wheeler et al. (2011) argue that military decisions are compromised by arms contractors, who provoke and prolong wars and sell expensive and impractical weapons. Barofsky (2012) and Admati and Hellwig (2024) argue that financial regulation is dominated by bankers. van der Kolk (2015) argues that the way psychiatric conditions are classified, diagnosed, and treated is dominated by the pharmaceutical industry, so that more effective non-pharmaceutical treatments are sidelined from research and practice. Herman and Chomsky (1988) argue that the media primarily publishes propaganda, due to influence from owners and advertisers (and other "filters"). Beck (2000) argues that both public and private law enforcement have often been corrupted in the US and the UK, especially through sham out of court settlements. If military discipline, the invisible hand, academia, the media, and law enforcement are unable to withstand—and be seen to withstand—corruption, this raises the question of whether any institutional design can do so.

Recent research shows how the joint design of information and transfers can produce information frictions, such as adverse selection or screening, that prevent would-be corrupt players from cutting a deal (Ortner and Chassang, 2018; Baliga and Sjöström, 1998; von Negenborn and Pollrich, 2020). Specifically, these authors show how a principal can cleverly design endogenous frictions to reduce the cost of deterring an agent from bribing a monitor to hide evidence, sometimes even to zero. Unfortunately, the practical application of the proposed mechanisms is limited by the fact that they either require the monitor to perfectly observe the agent's type or action, and/or the designer to be able to select her preferred equilibrium, and/or the players to have infinitely large budgets. Moreover, they consider neither the benefits of giving private information to both the agent and the monitor, nor whether their mechanisms can be extended to more than two players.

We propose a mechanism that overcomes these limitations and resembles the well-known game of Poker. In Poker, players with poor hands are more inclined to fold because they anticipate that any player who chooses to raise must have a better hand. This adverse selection or "lemons" effect causes play to unravel until all hands fold.[1] We use Poker to thwart corrupt agreements by sending random secret messages ("hands") to the agent and the monitor, and triggering a showdown when the monitor reports evidence against the

[1]The fact that one player folds immediately can also be seen by observing that Poker is a zero-sum game, and both players have an outside option of zero after seeing their hands.

agent. Paying the player with the highest hand a prize commensurate with the gains from evidence suppression creates a zero-sum game that deters bribes.[2] Our mechanism improves on this by introducing a novel handicapping rule: the highest hand wins only if it exceeds the lower hand by a large enough margin; otherwise nobody wins. Handicapping reduces the designer's cost of paying prizes; it continues to block bribe agreements that would be mutually beneficial, even though the game is no longer zero-sum. Nonetheless, Poker's adverse selection effect is strong enough to produce complete unravelling that leads both players to reject bribes.

Theorem 1 shows that the Poker mechanism is essentially the cheapest mechanism that blocks all corrupt side contracts, even when the players are assisted by a third party arbitrator with unlimited liability (as in Laffont and Martimort, 2000).[3] The proof requires us to overcome a transfinite induction problem in order to extend an unravelling analysis over a whole continuum of types. Theorem 2 considers two generalisations of Theorem 1. The first shows how *Rigged Poker* resolves the trade-off between rewarding an agent for honesty and punishing him for non-compliance. The second shows that adding players–such as more potential whistleblowers–in *n-Player Poker* lowers the cost of deterring corruption. Apart from providing a robust way to block corruption, the Poker mechanism partially answers some questions explored by Carroll (2016) and Brooks and Du (2024, 2025) about how much surplus from trade can be lost to information frictions, and what information structures maximise these frictions.

Section 2 describes a simple version of the mechanism, called *Constant Handicap Poker*, and how it works to defeat a simple class of corrupt side-contracts. Our focus is on deterring corruption so other aspects of the problem are kept simple. Section 3 formalises the designer's problem more generally, and presents our main mechanism, *Handicap Poker*. Section 4 proves that Handicap Poker defeats a broad class of side contracts and is an optimal solution to the designer's problem. Section 5 generalises the results to asymmetric n-player settings. Section 6 clarifies the main contributions compared to previous attempts to solve this problem, and also to worst-case information design and robust mechanism design. Section 7 outlines directions for future research. Throughout the paper we refer to the various versions of our mechanism generically as "Poker" or "the Poker mechanism".

[2]This game is isomorphic to the stylised version of Poker studied by von Neumann and Morgenstern (1953). Equation (19:21) implies that if $b = 0$, then both players will choose to "bid low", which corresponds to reporting evidence in our game. See Section 4.3.1 for details.

[3]Technically, there is no optimal mechanism. The Poker mechanism includes a small tie-breaking prize, and the cost approaches the infimum as this prize converges to zero.

2 A Corruption Problem

Dave, a developer has applied for permission to build a hotel for a profit of £10m. But Ray, the regulator, is worried that the hotel might disrupt local biodiversity. So Ray hires an auditor, Anne, to check. We assume that if the hotel is bad, then it is likely (but not necessarily certain) that Anne finds hard evidence which can be destroyed but not fabricated.[4] Ray is also worried that Dave might bribe Anne to destroy the evidence. Ray's problem is to find the cheapest way to deter bribes. Ray's first thought is a trivial mechanism that rewards Anne with £10.1m for evidence. Dave would not offer Anne a bribe, because he would have to spend more than his profit of £10m. But this is expensive: Ray has to pay out £10.1m whenever Anne produces evidence. Is there a cheaper solution, assuming limited liability for Anne and Dave?

We propose a new mechanism that resembles Poker. Ray deals cards and pays card-contingent prizes to Anne and Dave, which creates synthetic asymmetric information.[5] This frustrates bargaining between Anne and Dave by making them unsure of each other's bargaining positions, and creates a lemons market in the market for bribes with complete market failure. We prove that all negotiation procedures end in failure, including take it or leave it offers, alternating offers like Rubinstein (1982), double auctions, mediation, arbitration, and so on. And we prove that the mechanism is optimal. Any mechanism that deters bribes costs at least as much.

In the remainder of this section, we use Constant Handicap Poker to illustrate how our mechanism creates a two-sided adverse selection problem. We further simplify matters by beginning with a pure casino version of the game that ignores the distinct roles of the players by treating them symmetrically.

[4] We make the simplifying assumption that it costs nothing to destroy evidence. Our mechanism still works if destroying evidence is costly, but it can be made cheaper by deducting the cost of destroying evidence from the prizes. Section 5 discusses an extension of our mechanism that can accommodate evidence that can be fabricated, provided fabrication is costly enough.

[5] In this paper, *cards* are a metaphor for messages sent by the designer. Our assumption that the government can credibly commit to send messages according to a contractually declared distribution and make payments contingent on those cards is standard in the literature on mechanism and information design. See Section 6 for examples. Major casinos maintain credibility by using certified algorithms and undergoing regular inspections by gaming commissions. Attar et al. (2025) describe a method for using smart contracts to send private messages and react to private reports.

The Constant Handicap Poker mechanism, casino version

1. The house deals the players their hands x_1 and x_2 independently and uniformly from $[0, 1]$.
2. If the players can agree how to split a prize of size Π, then the Casino pays the split and the game ends.
3. Otherwise, play proceeds to showdown:
   (a) All cards are placed on the table facing up.
   (b) Player 1 wins $\Pi + \varepsilon$ if $\frac{1}{2}x_1 \geqslant x_2$, and similarly for player 2.
   (c) Otherwise, nobody wins.

In normal Poker, a player wins if they have the better hand, which happens with probability 1/2. Our Poker mechanism handicaps the higher player's hand by multiplying it by 1/2. As a result, each player wins with probability 1/4, and nobody wins with probability 1/2. The quantity $\varepsilon > 0$ is necessary only for breaking indifferences. Ray the regulator employs the Poker mechanism as follows:

The Constant Handicap Poker mechanism, Ray's version

1. Ray deals Anne and Dave their hands x_1 and x_2 independently and uniformly from $[0, 1]$.
2. If Anne fails to find evidence, or if Dave successfully bribes Anne to hide it, then the hotel is approved and the game ends.
3. Otherwise, Anne shows her evidence, the hotel is denied, and play proceeds to showdown:
   (a) All cards are placed on the table facing up.
   (b) Ray pays Anne £10.1m if $\frac{1}{2}x_1 \geqslant x_2$, and similarly for Dave.
   (c) Otherwise, Ray pays nothing.

In Ray's version of the Constant Handicap Poker mechanism, if Anne finds evidence, we will show that play always proceeds to showdown. Due to the handicapping, the mechanism pays out only $\frac{1}{2}\times$£10.1m = £5.05m, which is half as much as the trivial mechanism.

There are a few superficial differences between the casino and regulator version of Constant Handicap Poker. First, in the regulator version, the negotiation stage prize is not a transfer; it is earned directly by Anne and Dave ferom illicit hotel profits. This difference does not matter, because we will prove that the negotiation stage prize is never claimed, so it does not matter who pays it.

Second, in the regulator version, showdown is triggered by Anne revealing evidence or by the developer withdrawing the application, not just a negotiation failure. This difference just reflects the fact that the hotel is approved unless evidence is revealed. Showdown is triggered once the regulator knows that permission to build the hotel will be refused, because he knows that Anne and Dave did not form a coalition.

Third, in the regulator version, a deal may be impossible because Anne does not find any evidence to destroy. This means a prize is more likely to be awarded in the casino version. This creates a difference in accounting, but not incentives.

Fourth, the casino version is symmetric for both players.

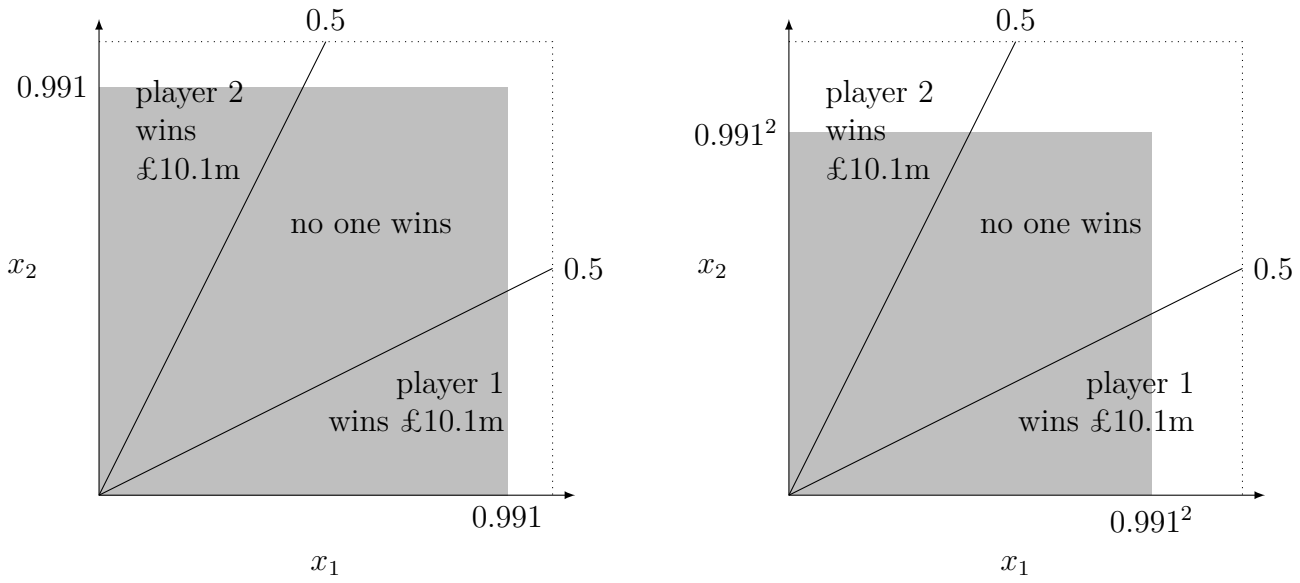
Since these differences are inconsequential and the casino version is simpler, we focus on the casino version with prize Π =£10m and tie-breaker ε =£0.1m.

A lemons problem. In Constant Handicap Poker, the showdown handicapping rule means that a prize is awarded only half of the time. This means that the negotiation stage prize is £10m, whereas the showdown stage prize is only £5.05m on average. So the players have much to gain from negotiating a deal.

But we show that Constant Handicap Poker creates a lemons problem. We prove that for every exogenous split of £10m, in every equilibrium the players reject the split almost surely. This is a weak result, because of the limited scope for negotiation. In fact, we will only illustrate how negotiations unravel when the players can only choose to accept or reject a 50-50 split of £10m. We resolve this in Section 4 when we prove that Handicap Poker deals with all possible negotiation procedures.

Despite the simple accept/reject choice, the strategy spaces are large because there are so many possible hands. But notice that the best response to any strategy is a cut-off rule, with strong hands rejecting the 50-50 split because they have a high chance of winning a showdown. This reduces the strategy space to a single number. We begin with a cut-off of 1 (always accept).

Suppose player 2 always accepts, and player 1 has the best possible hand of one. If player 1 accepts the split, he receives £5m. If he rejects the split, he wins the resulting showdown half of the time, and receives £5.05m in expectation. So he rejects the split and



(a) first round

(b) second round

Figure 1: Unravelling of the market for bribes.

forces a showdown. Similarly, player 1 rejects if his hand is better than $10/10.1 \approx 0.991$ and player 2 does too. Thus, players with hands better than 0.991 leave the negotiating table. In Figure 1a, the shaded area shows the remaining hands.

It turns out that the next round of deletion is similar. Suppose player 2 accepts if his hand is 0.991 or worse, and player 1's hand is exactly 0.991. This is no longer the best possible hand. But it is the best possible hand still at the negotiating table. So player 1 conditions his choice on $x_2 \leqslant 0.991$, i.e. on playing against a hand that is at the negotiating table. If player 1 accepts the split, he receives £5m, as before. If he rejects the split, he wins a showdown half of the time, and he receives £5.05m in expectation. He therefore rejects the deal too. Like before, players with hands better than 0.991^2 reject the split, and leave the negotiating table. In Figure 1b, the shaded area shows the remaining hands.

The unravelling continues until all hands reject the 50-50 split. We conclude that Constant Handicap Poker deters the players from accepting a 50-50 split. The logic is similar for other exogenous splits, e.g. 40-60.

The unravelling process resembles Akerlof's (1970) analysis of "lemons" markets. But the two-sided nature of Constant Handicap Poker means that we do not trace a path of decreasing prices as the quality becomes more diluted. Instead, we keep the price fixed (50-50 split), and trace the dwindling set of hands that agree to a deal. Another similarity is the

nature of a "lemon". The crux of a lemons market is that one player wants to cut a deal the most when the other player benefits the least. In Constant Handicap Poker, a weak hand is likely to lose a showdown, so a deal is attractive. But a weak hand for one player means that the other player is more likely win a showdown, so he benefits the less from a deal.

How might the players try to cheat the mechanism? The easiest way would be to reveal their cards to one another. The house can prevent this by giving the players a deck of fake cards. This way, even though Anne might reveal her true card to Dave, she would not be credible since Dave can never be sure that she has shown her true hand, and not a fake.

Another way to cheat would be to commit not to look at their own hands. This would undo the lemons problem. A player with a strong hand would no longer know that he would have a high chance of winning a showdown. So the unravelling would never get started. Thus the house must insist that the players look at their hands.

Similarly, the players would also like to promise to refuse their showdown prizes. This would force the players to collect negotiation stage prizes instead, which are larger. To prevent this, the house must insist that the players take their showdown prizes home securely.

Other side contracts. There is a large literature (surveyed by Tirole, 1993) about corruption with various bargaining procedures. Laffont and Martimort (1997, 2000) point out that the proposed mechanisms are highly sensitive to the assumed bargaining procedure. They argue that good institutions for deterring collusion and corruption ought to withstand any form of negotiation.

Does Constant Handicap Poker deter all other forms of negotiation beyond accepting or rejecting an exogenous split? The answer is no. If the players enlist the help of the mafia to enforce a side contract, they can extract all of the prize money between them.

Win-or-split side contract

1. The mafia pays the players £1m each.
2. The players tell the mafia their hands, x_1 and x_2.
3. If neither player has a winning hand (neither $\frac{1}{2}x_1 \geqslant x_2$ nor $\frac{1}{2}x_2 \geqslant x_1$), then the mafia advises the players to agree to split and give the money to the mafia. The mafia gets £10m and the players get £0.
4. If either player has a winning hand, then the mafia advises them to proceed to showdown. The winner gets £5.1m, the loser and the mafia get £0.

Under the mafia's win-or-split side contract, either one player wins a showdown, or they agree to a split. So the casino always awards a prize. Win-or-split is incentive compatible because neither player can benefit from misreporting their hands: initiating extra showdowns is useless because no extra prizes would be awarded; initiating fewer showdowns is useless too because players get no extra money from splitting. It is individually rational because it gives the players a £1m fixed payment in addition any prize they would win if they rejected the contract. The mafia makes a profit of $£\frac{1}{2} \times 10 - 2 = 3$m, and each player receives $1 + \frac{1}{4} \times$ £10.1m = £3.525m on average. Together, the two players and the mafia extract £10.05m, which is all of the prize money on the table. We conclude that the win-or-split side contract defeats the Constant Handicap Poker mechanism. We need a better mechanism.

Handicap Poker. Constant Handicap Poker was vulnerable to the win-or-split side contract because the mafia was able to predict when a showdown would be successful. We propose Handicap Poker, which adds an extra layer of randomness to thwart this. Specifically, instead of a fixed handicap of $\frac{1}{2}$, the handicap is drawn uniformly from $[0, 1]$. In Poker terminology, the handicap is a *community card*, as appears in Hold'em varieties of Poker.

The Handicap Poker mechanism, casino version

1. The house deals the players their hands x_1 and x_2 independently and uniformly from $[0, 1]$. The house also draws a community card independently and uniformly from $[0, 1]$, and places it face down on the table.
2. If the players can agree how to split a prize of size Π, then the Casino pays the split and the game ends.
3. Otherwise, play proceeds to showdown:
   (a) All cards are placed on the table facing up.
   (b) Player 1 wins $\Pi + \varepsilon$ if $yx_1 \geqslant x_2$, and similarly for player 2.
   (c) Otherwise, nobody wins.

As before, nobody wins during showdown 50% of the time, so Handicap Poker costs the same £5.05m as Constant Handicap Poker.

What is the fundamental difference between Constant Handicap Poker and Handicap Poker? The answer lies in the information frictions. Constant Handicap Poker creates a pure adverse selection problem. If the players can only accept or reject an exogenous split, then Constant Handicap Poker ensures that no matter which hands decide to agree, some of

their participation constraints are violated. However, the win-or-split side contract solves the adverse selection problem. The mafia helps by aggregating information. It gives the surplus back to the players, and thus satisfies all of the participation constraints. And the players have no incentive to lie to the mafia about their hands, because the mafia lets the players keep all of the (potential) showdown prizes. The discrete win rule in Constant Handicap Poker aids the mafia by producing a steep (indeed, discontinuous) change in player i's payoff at the threshold $2x_{-i}$. This sudden jump makes it risky for a player with a low hand to "bluff" the mafia by reporting a high hand: if the threshold lies between the true hand and the report, then the mafia will "call i's bluff" by triggering a showdown, and i will lose with certainty. The fact that small bluffs can have big payoff consequences makes it easy for the mafia to screen the players. i By contrast, the random handicap in Handicap Poker means that i's expected payoff changes continuously (hence more gradually) in her hand, which facilitates bluffing. If the mafia were to propose the same win-or-split side contract during a Handicap Poker game, then a player with any hand (except 0) would try to trick the mafia into a showdown by pretending to have the best possible hand. Showdown is strictly better than agreeing to a split because every hand has a strictly positive chance of winning. Consequently, the mafia can no longer screen players costlessly; they must grant information rents to elicit truthful reporting.

The following sections formalise this argument and prove that Handicap Poker is optimal.

3 Model

There is one designer, one mafia, and two players, $i = 1, 2$ (see Section 5 for an n-player extension). All players are risk neutral.[6] At the interim stage, the mafia will try to facilitate an agreement between the players. If the players reach an agreement, then they collectively get a surplus of $\Pi = 1$. In the context of the corruption problem of Section 2, this means that they destroy the evidence and obtain planning permission. The players' outside options from reaching agreement are determined by the designer's choice of mechanism in the first stage.

[6]Risk aversion can be a problem for our mechanism because it makes the certainty of bribes relatively more attractive than the uncertainty of the showdown payoffs. However, if the players' transfers for *not* reporting evidence are also endogenous, then they can be used as a source of variation, instead of their transfers *for* reporting evidence. Doing so reverses the effects of risk aversion, since reporting evidence becomes the less risky action. Alternatively, the designer can compensate the players for their risk premium by increasing the showdown transfers.

Definition 1. A *mechanism* $\mathcal{M} = (X, Y, \Sigma, P, t)$ consists of:

- a set of possible message profiles $X = X_1 \times X_2$, where X_i is the set of messages that the designer can send to player $i = 1, 2$, and
- a set of community messages, Y, that can determine payoffs but are not observed by any player.
- a σ-algebra, Σ, on $X \times Y$.
- a probability measure $P : \Sigma \to [0, 1]$ over $X \times Y$.
- a pair of transfer functions $t = (t_1, t_2)$, where each function $t_i : X \times Y \to \mathbb{R}_+$ specifies the transfers from the designer to player i, in the event that the players fail to reach an agreement.

If the players do reach an agreement, then the designer pays them nothing.

The community message set Y is not strictly necessary, since we can replace the transfer functions with their expectation over Y without affecting the players' incentives. We include it only because our Poker mechanism is more neatly described this way. At times, it will be convenient to write transfers without the community messages as $t_i(x) = \int_Y t_i(x, y)\, dP_Y(y)$.

Our requirement that transfers be positive embodies an assumption that the players have binding limited liability constraints, which we normalise to zero.[7]

Mechanisms must also satisfy the following technical properties:

1. The transfers t_i are Σ-measurable, i.e. pre-images of Borel sets are Σ-measurable.
2. Σ is a product of σ-algebras Σ_i that contain subsets of X_i, and Σ_Y that contains subsets of Y.
3. P is absolutely continuous with respect to the product measure $P_{X_1} \times P_{X_2} \times P_Y$, where $P_{X_i} : \Sigma_i \to [0, 1]$ is player i's marginal distribution, and $P_Y : \Sigma_Y \to [0, 1]$ is the marginal distribution on Y. It is also convenient to define Σ_X to be the projection of Σ on X and P_X to be the marginal distribution on X.

[7] If one or other of the players has unlimited liability, then the designer can deter corruption for free using a mechanism of the type described in Appendix C.

4. There exist conditional probability distributions $P_{X_{-i}|X_i}$ for every player i.[8]

5. Player i's expected transfer exists and equals

$$\bar{W}_i(x_i) := \int_{X_{-i}\times Y} t_i(x_i, x_{-i}, y)\, dP_{X_{-i}|X_i}(x_{-i}, y|x_i). \tag{1}$$

Once the designer has specified a mechanism, the mafia proposes a side contract to help the players reach agreement. We make the conservative assumption that the mafia can commit to enforce side contracts and can do partial implementation, i.e. the mafia succeeds in executing a bribe if one of the equilibria is successful. This has two advantages. First, it ensures that the designer's mechanisms are robust to a wider range of side contracts. Second, it means that the revelation principle applies: the mafia can partially implement any outcome of any negotiation procedure or side contract (such as Rubinstein bargaining, double auctions, and arbitration), in a truth-telling equilibrium of an equivalent direct side contract. Hence, there is no loss of generality in restricting attention to direct side contracts. The mafia does this by promising to simulate the players' strategies on their behalf. This promise amounts to an allocation rule (when should the players agree to a split), and transfers ("bribes"), as a function of the players' hands.

Definition 2. A *direct side contract* $\mathcal{S} = (a, b)$ against $\mathcal{M}$ consists of

- a Σ-measurable agreement rule $a : X \to [0, 1]$ that specifies the probability of agreeing a split for each message profile x reported by the players; and
- a pair of Σ-measurable bribe functions $b_i : X \to \mathbb{R}$ that specifies the (possibly negative) bribe that the mafia pays the players for each message profile x.

If both players accept the side contract, then they make reports x_1 and x_2 to the mafia and receive $b_1(x)$ and $b_2(x)$. With probability $1 - a(x)$, the mafia tells the players to reject agreement and collect transfers t_i from the designer. Then, with probability $a(x)$, the mafia tells the players to accept a split of the prize. Rather than explicitly specify *which* split they accept, we follow Myerson (1981) by assuming that money from the split is paid directly to

[8] Dudley (2002, page 343) defines $P_{X_{-i}|X_i}$ to be a conditional probability distribution if (i) for each $x_i \in X_i$, $P_{X_{-i}|X_i}(\cdot|x_i)$ is a probability measure on (X_{-i}, Σ_{-i}), (ii) for each $A_{-i} \in \Sigma_{-i}$, the function $x_i \mapsto P_{X_{-i}|X_i}(A_{-i}|x_i)$ is Σ_i-measurable from X_i into $\mathbb{R}$, and (iii) for each $A = A_i \times A_{-i} \in \Sigma_X$, $P(A) = \int_{A_i} P_{X_{-i}|X_i}(A_{-i}|x_i)\, dP_{X_i}(x_i)$. By Dudley (2002, Theorem 10.2.1), these conditions ensure that $\int g(x)\, dP(x) = \int\int g(x_i, x_{-i})\, dP_{X_{-i}|X_i}(x_{-i}|x_i)\, dP_{X_i}(x_i)$.

the mafia and redistributed according to the b_i functions. We only consider side contracts in which the players can calculate their expected payoffs, i.e.

$$V_i(x_i, x_i') := \int_{X_{-i} \times Y} \left[(1 - a(x_i', x_{-i}))t_i(x, y) + b_i(x_i', x_{-i})\right] \, dP_{X_{-i}|X_i}(x_{-i}, y|x_i) \tag{2}$$

must exist for all messages x_i and all reports x_i'. If player i declines to participate in the side contract, he receives $\bar{W}_i(x_i)$ from the designer's mechanism.

Definition 3. Given a mechanism $\mathcal{M}$, a direct side contract $\mathcal{S}$ is *feasible* if it satisfies all of the following constraints:

1. (Side incentive) Each player i maximises their payoff by truthfully reporting their private message to the mafia, i.e. for all $x_i \in X_i$,

$$V_i(x_i, x_i) = W_i(x_i) := \max_{x_i'} V_i(x_i, x_i'). \tag{SI}$$

2. (Side player participation) Each player i prefers to accept the side contract

$$W_i(x_i) \geqslant \bar{W}_i(x_i) \quad \forall x_i \in X_i. \tag{SPP}$$

3. (Side mafia participation) In expectation, the mafia does not lose money, i.e.

$$\pi := \int_X a(x) \, dP_X(x) - \int_X \left[b_1(x) + b_2(x)\right] \, dP_X(x) \geqslant 0. \tag{SMP}$$

4. (Side surplus) There is a strictly positive surplus so that the side contract Pareto dominates non-participation, i.e.

$$\int_X \sum_{i \in \{1,2\}} \left[W_i(x_i) - \bar{W}_i(x_i)\right] dP_X(x) + \pi > 0. \tag{SS}$$

A mechanism $\mathcal{M}$ *blocks* a side contract $\mathcal{S}$ if $\mathcal{S}$ is not feasible.

The left side of the mafia participation constraint comes from the fact that the players generate a combined project profit of 1 when they reach an agreement, which happens with probability $a(x)$. We assume that a feasible side contract must generate strictly positive surplus in expectation, because innocuous side contracts such as the null side contract need

not be blocked. The surplus need not be strictly positive ex post. Requiring the side contract to generate a strictly positive surplus ex post would only weaken our main result, since it would no longer guarantee that Handicap Poker is robust to mafias with big budgets.

Definition 4 (The Designer's Problem)**.** The (unweighted) *cost of a mechanism* is given by the expected value of the transfers, $c(\mathcal{M}) := \sum_{i \in N} \int_{X \times Y} t_i(x, y)\ dP(x, y)$, where $\mathcal{M} = (X, Y, \Sigma, P, t)$. The *designer's problem* is to find the cheapest mechanism that blocks all side contracts i.e.

$$c^* = \inf_{\mathcal{M}}\ c(\mathcal{M}) \text{ s.t. } \mathcal{M} \text{ blocks every feasible side contract } \mathcal{S}. \tag{P1}$$

In this notation, Handicap Poker with a showdown bonus of ε, denoted $\mathcal{M}^*(\varepsilon)$ is defined by $Y^* = X_i^* = [0, 1]$ for $i = 1, 2$, Σ^* equal to the set of Lebesgue measurable subsets of $[0, 1]^3$, P^* is the Lebesgue measure, i.e. uniform IID draws, and $t_i^*(x, y) = I(yx_i \geqslant x_j)(1+\varepsilon)$, where $I(\cdot)$ denotes the indicator function. Additionally, it admits multiple strategically equivalent formulations. We highlight three variants that prove useful for the extensions in Section 5:

- Personal community cards: $Y = Y_1 \times Y_2 = [0, 1]^2$ and $t_i(x, y) = I(y_i x_i \geqslant x_{-i})(1 + \varepsilon)$.
- No community cards: $Y = \varnothing$ and $t_i(x) = \max\{0, 1 - x_{-i}/x_i\}(1 + \varepsilon)$.
- General distributions: P is the product of three marginal distributions with invertible CDFs F_1, F_2 and F_Y, and $t_i(x, y) = I(F_Y^{-1}(y)F_{-i}^{-1}(x_{-i}) \geqslant F_i^{-1}(x_i))(1 + \varepsilon)$.

4 Results

Theorem 1. *The infimum cost of blocking side contracts in problem* (P1) *is* $c^* = \frac{1}{2}$*. Handicap Poker with showdown reward* $\varepsilon > 0$ *blocks all feasible side contracts, and its cost approaches the infimum, i.e.* $\lim_{\varepsilon \to 0} c(\mathcal{M}^*(\varepsilon)) = c^*$*. Each player receives an expected transfer of* $\frac{1+\varepsilon}{4}$*.*

The proof of Theorem 1 has three parts: the first shows that Handicap Poker blocks all side-contracts when $\varepsilon > 0$; the second shows that Handicap Poker's cost approaches $\frac{1}{2}$ as $\varepsilon \to 0$; the third shows that any mechanism that blocks all side contracts costs strictly more than $\frac{1}{2}$. The details are given in the remainder of this section.

4.1 Handicap Poker Blocks All Side Contracts

We follow Myerson's (1981) classic proof strategy, borrowing improvements from Krishna (2009). First, we fix an arbitrary (direct) side contract. Second, we apply the envelope theorem to calculate the player's values as a function of their hands (Lemma 1). Third, we calculate the minimal bribes necessary to screen the players' types for a given agreement rule (Lemma 2). Finally, we show that these bribes are too expensive for the mafia to make a profit. In other words, the cost of screening the players exceeds the gains from reaching an agreement.

Lemma 1. *For any feasible direct side contract $\mathcal{S} = (a, b)$ to Handicap Poker, the marginal value of a better hand for a player with hand x_i is*

$$W_i'(x_i) = \frac{1+\varepsilon}{x_i} \int (1 - a(x_i, x_{-i}))I(x_i \geqslant x_{-i} \geqslant yx_i)\, d(x_{-i}, y). \tag{3}$$

The value of hand x_i can be calculated as

$$W_i(x_i) = W_i(1) - \int_{x_i}^{1} W_i'(s)\, ds, \tag{4}$$

which has an expected value of

$$\int W_i(x_i)\, dx_i = W_i(1) - (1+\varepsilon) \int (1 - a(x))I(x_i \geqslant x_{-i} \geqslant yx_i)\, d(x, y). \tag{5}$$

Proof. We write the proof for player 1. The proof for player 2 is the same. If player 1 accepts the side contract, and reports x_1' when his true hand is x_1, then his expected payoff is

$$V_1(x_1, x_1') = \int_0^1 \underbrace{b_1(x_1', x_2)}_{\text{bribe}}\, dx_2 + \int_0^1 \underbrace{(1 - a(x_1', x_2))}_{\text{showdown?}} \underbrace{\int_0^1 I(yx_1 \geqslant x_2)(1+\varepsilon)\, dy}_{\text{showdown prize}}\, dx_2. \tag{6}$$

We would like to calculate the marginal value of a better hand. Indicator functions are not differentiable, so we rewrite $\int_0^1 I(yx_1 \geqslant x_2)\, dy$ as $\max\{0, 1 - x_2/x_1\}$, and player 1's value as

$$V_1(x_1, x_1') = \int_0^1 b_1(x_1', x_2)\, dx_2 + (1+\varepsilon) \int_0^{x_1} (1 - a(x_1', x_2)) \left(1 - \frac{x_2}{x_1}\right) dx_2. \tag{7}$$

By Leibnitz rule, the marginal value of a better hand, holding the report to the mafia

fixed, is

$$\frac{\partial}{\partial x_1} V_1(x_1, x_1') = (1+\varepsilon)(1 - a(x_1', x_2)) \left(1 - \frac{x_1}{x_1}\right) + (1+\varepsilon) \int_0^{x_1} (1 - a(x_1', x_2)) \frac{x_2}{(x_1)^2} \, dx_2 \quad (8)$$

$$= \frac{1+\varepsilon}{x_1} \int (1 - a(x_1', x_2)) \underbrace{I(x_1 \geqslant x_2 \geqslant y x_1)}_{\text{better, but not by enough}} \, d(x_2, y). \quad (9)$$

If $\mathcal{S}$ satisfies player 1's side incentive constraint, then x_1 is an optimal choice of x_1', so the envelope theorem implies that the marginal value of a better hand is

$$W_1'(x_1) = \left[\frac{\partial}{\partial x_1} V(x_1, x_1')\right]_{x_1' = x_1} = \frac{1+\varepsilon}{x_1} \int (1 - a(x_1, x_2)) I(x_1 \geqslant x_2 \geqslant y x_1) \, d(x_2, y). \quad (10)$$

The value of hand x_1 then follows from the fundamental theorem of calculus.[9] Player 1's expected payoff is[10]

$$\int W_1(x_1) \, dx_1 = W_1(1) - \int \int_{x_1}^1 \frac{1+\varepsilon}{s} \int (1 - a(s, x_2)) I(s \geqslant x_2 \geqslant y s) \, d(x_2, y) \, ds \, dx_1 \quad (11)$$

$$= W_1(1) - (1+\varepsilon) \int (1 - a(x)) I(x_1 \geqslant x_2 \geqslant y x_1) \, d(x, y). \quad (12)$$

□

Lemma 2. *For any feasible direct side contract $\mathcal{S} = (a, b)$ to Handicap Poker,*

$$\int b_i(x) \, dx \geqslant (1+\varepsilon) \left[\frac{1}{2} - \int (1 - a(x)) I(x_i \geqslant x_{-i}) \, dx\right] \text{ for each player } i = 1, 2. \quad (13)$$

Proof. If $\mathcal{S}$ satisfies player 1's side participation constraint, then $W_1(x_1) \geqslant \bar{W}_1(x_1)$ for all $x_1 \in [0, 1]$. For the strongest hand, we have $W_1(1) \geqslant \bar{W}_1(1) = \frac{1+\varepsilon}{2}$. Lemma 1 then implies that

$$\int W_1(x_1) \, dx_1 \geqslant (1+\varepsilon) \left[\frac{1}{2} - \int (1 - a(x)) I(x_1 \geqslant x_2 \geqslant y x_1) \, d(x, y)\right]. \quad (14)$$

[9] Notice that $V_1(x_1, x_1')$ is convex in x_1. So W_1 is an upper envelope of convex functions and is therefore also convex. Thus W_1 is absolutely continuous by (Royden and Fitzpatrick, 1988, Corollary 17), so the fundamental theorem of calculus (Royden and Fitzpatrick, 1988, Theorem 10) applies.

[10] We make use of the following fact to simplify the triple integral: If $f : [0, 1] \to \mathbb{R}$ is Lebesgue integrable, then $\int_0^1 \int_x^1 f(s) \, ds \, dx = \int_0^1 x f(x) \, dx$. To see this, let $g(x, s) = I(s \geqslant x) f(s)$, which is Lebesgue integrable on $[0, 1]^2$. By construction, $\int_0^1 \int_x^1 f(s) \, ds \, dx = \int_0^1 \int_0^1 g(x, s) \, ds \, dx$. By Fubini's theorem, $\int_0^1 \int_0^1 g(x, s) \, ds \, dx = \int_0^1 \int_0^1 g(x, s) \, dx \, ds = \int_0^1 f(s) \int_0^1 I(s \geqslant x) \, dx \, ds = \int_0^1 f(s) s \, ds$.

The side incentive constraint, together with (6), gives

$$W_1(x_1) = V_1(x_1, x_1) = \int b_1(x)\, dx_2 + (1+\varepsilon) \int (1 - a(x)) I(yx_1 \geqslant x_2)\, d(x_2, y). \tag{15}$$

Substituting this into (14) gives a lower bound on bribes paid to player 1, with

$$\int b_1(x)\, dx \tag{16}$$

$$\geqslant (1+\varepsilon) \left[\frac{1}{2} - \int (1 - a(x)) \{ I(x_1 \geqslant x_2 \geqslant yx_1) + I(yx_1 \geqslant x_2) \}\, d(x, y) \right] \tag{17}$$

$$= (1+\varepsilon) \left[\frac{1}{2} - \int (1 - a(x)) \{ I(x_1 \geqslant x_2) - I(yx_1 \geqslant x_2) + I(yx_1 \geqslant x_2) \}\, d(x, y) \right] \tag{18}$$

$$= (1+\varepsilon) \left[\frac{1}{2} - \int (1 - a(x)) I(x_1 \geqslant x_2)\, dx \right]. \tag{19}$$

The same logic applies to player 2. □

Summing the inequalities in Lemma 2, we find that the mafia's total transfers to both players are at least

$$\int b_1(x)\, dx + \int b_2(x)\, dx \geqslant (1+\varepsilon) \int a(x)\, dx \geqslant \int a(x)\, dx. \tag{20}$$

If an agreement is reached with positive probability so that $\int a(x)\, dx > 0$, then the last inequality is strict, the mafia makes a loss, and would not participate, violating (SMP). Otherwise, $\int a(x)\, dx = 0$, and the side contract creates no surplus for the parties, because

$$\int \sum_{i \in \{1,2\}} [W_i(x_i) - \bar{W}_i(x_i)]\, d(x, y) + \pi \tag{21}$$

$$= \int \sum_{i \in \{1,2\}} [(1 - a(x)) t_i(x, y) + b_i(x) - t_i(x, y)]\, d(x, y) + \int \left[a(x) - \sum_{i \in \{1,2\}} b_i(x) \right] dx \tag{22}$$

$$= 0. \tag{23}$$

This violates the side surplus constraint, (SS). Since the choice of side contract was arbitrary, we conclude that Handicap Poker deters bribes, regardless of the negotiation procedure.

4.2 Handicap Poker Costs 1/2

The cost of Handicap Poker with showdown bonus ε is straightforward to calculate. Since $\int t_1^*(x,y)\,dy = \max\{0, 1-x_2/x_1\}(1+\varepsilon)$, we get $W_1(x_1) = (1+\varepsilon)\int_0^{x_1}(1-x_2/x_1)\,dx_2 = (1+\varepsilon)\frac{x_1}{2}$, and hence $\int W_1(x_1)\,dx_1 = \frac{1+\varepsilon}{4}$. There are two symmetric players, so the total cost is $\frac{1+\varepsilon}{2}$, which converges to $c^* = \frac{1}{2}$ as $\varepsilon \to 0$.

4.3 Handicap Poker is Optimal

To prove that Handicap Poker is optimal, this section establishes that any mechanism $\mathcal{M} = (X, Y, \Sigma, p, t)$ that blocks all side contracts must cost at least 1/2. Our proof is based on Carroll (2016)'s theory of contagious infections, which is similar in spirit to Akerlof (1970) and Rubinstein (1989). The logic involves all types rejecting trade in several rounds.

We begin with two simplifying reformulations. The first is based on the observation that equal split bribes are the hardest to block; asymmetric splits can be blocked by rewarding the worse-off player more than their share. We relax the problem to blocking the equal split bribe only. Second, we introduce a simple *bribe game* in which players choose whether to accept or reject the equal split bribe. Lemma 3 establishes that if $\mathcal{M}$ is feasible then no bribes are agreed in any strict equilibrium of the bribe game. This allows us to recast the blocking constraint in terms of players rejecting bribes in all equilibria, which is simpler than blocking any conceivable side contract. This setting is now suitable for applying Carroll's insights on informationally robust trade.

The second part uses the bribe game to construct lower bounds on the cost of blocking bribes. If under some beliefs about the other player's type, a player finds rejecting the bribe is strictly profitable, then he must expect an even bigger reward from the regulator. Thus, summing up gains from profitable rejections is isomorphic to summing up the costs for blocking various types from agreeing on bribes. To avoid double counting, we define *costing sequences* of disjoint subsets of X. Each costing sequence traces out a possible order of contagious infections. The *size* of a costing sequence is the measure of the union of all the sets along the sequence. Lemma 4 proves that each set in the sequence has the property that at least one player's conditional expected transfer exceeds 1/2. It follows that the cost of $\mathcal{M}$ must exceed 1/2 times the size of the largest costing sequence.

In the final part of the proof, Lemma 5 shows that players reject bribes in every equilibrium of the bribe game only if the size of the largest costing sequence is 1, i.e. if all types get infected. Together with part 2, this implies that $\mathcal{M}$ must cost at least 1/2. If X is finite, then

this is an immediate consequence of Claim 1, which applies Carroll's (2016, Propositions 3.1 and 3.2) logic to extend any costing sequence that is smaller than 1. Like Carroll (2016), the proof starts with all types initially accepting bribes, but gradually choosing to reject with each round of iterated reasoning.

But we do not assume a finite type space. Rather, we modify Carroll's technique to accommodate uncountable type spaces. First, Claim 1 costs a positive measure of types in each round, not just one. Second, Carroll's proof relies on Nash's equilibrium existence theorem. But with uncountable type spaces, the strategy spaces are more complex, so Claim 1 uses Balder's (1988) equilibrium existence theorem instead. Third, Claim 2 shows that costing can be completed in a countable number of rounds, which rules out transfinite induction problems.

4.3.1 The Bribe Game

In the *bribe game* of a mechanism, each player receives their message, and chooses a probability with which to accept the equal split bribe, $b_i = \frac{1}{2}$. If both players accept, then each player gets $\frac{1}{2}$ from the side contract. If either player rejects, then each player gets their transfer from $\mathcal{M}$.

Definition 5. Given a mechanism $\mathcal{M} = (X, Y, \Sigma, P, t)$, the *bribe game* of $\mathcal{M}$ is defined by a (behavioural) strategy $a_i : X_i \to [0,1]$, and an ex ante utility function

$$U_i(a_i, a_{-i}) = \int_X \left[\tfrac{1}{2} a_1(x_1)a_2(x_2) + (1 - a_1(x_1)a_2(x_2))\, t_i(x)\right] dP_X(x) \tag{24}$$

for each player $i = 1, 2$. Strategy profile (a_1^*, a_2^*) forms a *strict (weak) Nash equilibrium in behavioural strategies* if:

- each a_i^* is measurable, i.e. pre-images of Borel sets are Σ_i measurable;
- player i is strictly (weakly) better off playing a_i^* than any essentially different[11] strategy a_i', i.e. $U_i(a_i^*, a_{-i}^*) > (\geqslant) U_i(a_i', a_{-i}^*)$ for all a_i' with $P_{X_i}(x_i | a_i'(x_i) \neq a_i^*(x_i)) > 0$.

We need to consider strict equilibria because of our requirement that side contracts create a strictly positive surplus. Strategies that are the same almost everywhere yield the same expected utility, so the definition of strict equilibrium is only meaningful if we restrict attention to essentially different strategies.

[11] Two measurable functions are *essentially different* if they differ on a non-zero measure set.

Lemma 3. *If mechanism $\mathcal{M}$ blocks all side contracts, then the corresponding bribe game has no strict equilibrium (a_1^*, a_2^*) in which bribes are accepted with strictly positive probability, i.e. $\int_X a_1^*(x_1)a_2^*(x_2)\,dP_X(x) > 0$.*

Proof. We prove the contrapositive using the logic of the revelation principle (Laffont and Martimort, 2000). Suppose (a_1^*, a_2^*) is a strict equilibrium of the bribe game of $\mathcal{M}$ with $\int_X a_1^*(x_1)a_2^*(x_2)\,dP_X(x) > 0$. It suffices to show that there exists some side contract $(\hat{a}, \hat{b})$ that is not blocked by $\mathcal{M}$.

Consider the side contract $(\hat{a}, \hat{b})$ in which the mafia promises to execute the equilibrium acceptance strategies (a_1^*, a_2^*) on the players' behalf and share the resulting surplus equally, i.e. $\hat{a}(x) = a_1^*(x_1)a_2^*(x_2)$ and $\hat{b}_i(x) = \frac{\hat{a}(x)}{2}$ for $i = 1, 2$. By a standard delegation argument, truthful reporting and participation are weakly optimal for the two players.

The side contract gives the mafia a profit of 0, and a total surplus to the players of

$$\int_X \sum_{i\in\{1,2\}} [W_i(x_i) - \bar{W}_i(x_i)]\,dP_X(x) = \int_X \sum_{i\in\{1,2\}} [U_i(a_i^*(x_i), a_{-i}^*(x_{-i})) - U_i(0, a_{-i}^*(x_{-i}))]\,dP_X(x). \tag{25}$$

Since we assumed (a_1^*, a_2^*) reaches agreement with strictly positive probability, it is essentially different from the always-reject strategy 0. We deduce that player i strictly prefers a_i^* over 0, and that the total surplus is strictly positive. We conclude that $\mathcal{M}$ does not block $(\hat{a}, \hat{b})$. □

4.3.2 Costing Sequences

Definition 6. A *finite costing sequence* of R rounds is a finite sequence of costed types $(C_r)_{r\leqslant R}$, where $C_r = (C_{1r}, C_{2r})$ and $C_{ir} \subseteq X_i$ for all $r = 1, \ldots, R$, that satisfies the following properties for $i = 1, 2$:

1. At the start, no types are costed, i.e. $C_{i0} = \varnothing$.

2. Types can only be costed once, i.e. $C_{ir} \subseteq U_{ir-1}$ for all $r \leqslant R$, where $U_{ir-1} := X_i \backslash \cup_{s\leqslant r-1} C_{is}$ is the set of *uncosted* types in round $r-1$.

3. In round r, only one player has any newly costed types, i.e. if $C_{ir} \neq \varnothing$ then $C_{jr} = \varnothing$ for $j \neq i$.

4. Types can only be costed by a *costing strategy profile*. Specifically, if $C_{ir} \neq \varnothing$, then there exists a costing strategy profile a_{-i} such that

(a) uncosted types of the other player accept the bribe, i.e. $a_{-i}(x_{-i}) = 1$ for all $x_{-i} \in U_{-ir}$;

(b) each newly costed type is strictly better off rejecting the bribe, i.e.

$$\int_{X_{-i}} \left(t_i(x) - \tfrac{1}{2}\right) a_1(x_1)a_2(x_2)\, dP_{X_{-i}|X_i}(x_{-i}|x_i) > 0 \qquad \forall x_i \in C_{ir}. \tag{26}$$

The set of *newly costed type profiles* by i in round r is $\mathbb{C}_{ir} := C_{ir} \times U_{-ir}$. The *size* of costing sequence $(C_r)_{r\leqslant R}$ is $s((C_r)_{r\leqslant R}) := P_X(\cup_{\substack{r\leqslant R\\ i=1,2}} \mathbb{C}_{ir})$. *Infinite costing sequences* and their sizes are defined analogously.

Figure 2 depicts an example of costing sequence. Costing sequences give a lower bound

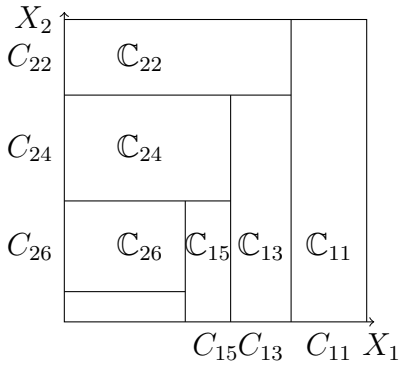


Figure 2: A two player costing sequence.

on the cost of a mechanism.

Lemma 4. *Let $(C_r)_{r\leqslant R}$ be any finite costing sequence of the bribe game of a mechanism $\mathcal{M}$ that blocks all side contracts. Then the cost of $\mathcal{M}$ exceeds 1/2 times the size of $(C_r)_{r\leqslant R}$, i.e. $c(\mathcal{M}) > s((C_r)_{r\leqslant R})/2$.*

Proof. For any $x \in X$, if $x_1 \in C_{1r}$ then

$$\int_{X_2} t_1(x)\ dP_{X_2|X_1}(x_2|x_1) \geqslant \int_{X_2} t_1(x)a_2(x_2)\ dP_{X_2|X_1}(x_2|x_1) \tag{27}$$

$$> \tfrac{1}{2}\int_{X_2} a_2(x_2)\ dP_{X_2|X_1}(x_2|x_1) \tag{28}$$

$$\geqslant \tfrac{1}{2}P_{X_2|X_1}(U_{2r}|x_1), \tag{29}$$

where the first inequality comes from the fact that $t_i(x_i) \geqslant 0$ and $a_2(x_2) \in [0,1]$, the second from (26), and the third from the fact that $a_2(x_2) = 1$ for all uncosted types $x_2 \in U_{2r}$. Integrating the first and last expressions over C_{1r} gives

$$\int_{C_{1r}\times X_2} t_1(x)\ dP_X(x) > \tfrac{1}{2}P_X(\mathbb{C}_{1r}). \tag{30}$$

The left side of (30) sums up the transfers over all newly costed types. The right side sums up the foregone bribes. Similarly, $\int_{X_1\times C_{2r}} t_2(x)\ dP_X(x) > \frac{1}{2}P_X(\mathbb{C}_{2r})$ for all C_{2r}.

Summing up the inequalities over both players and rounds $r \leqslant R$ gives

$$c(\mathcal{M}) = \int_X t_1(x) + t_2(x)\ dP_X(x) \tag{31}$$

$$\geqslant \int_{\cup_{r\leqslant R} C_{1r}\times X_2} t_1(x)\ dP_X(x) + \int_{X_1\times \cup_{r\leqslant R} C_{2r}} t_2(x)\ dP_X(x) \tag{32}$$

$$= \sum_{r\leqslant R}\int_{C_{1r}\times X_2} t_1(x)\ dP_X(x) + \int_{X_1\times C_{2r}} t_2(x)\ dP_X(x) \tag{33}$$

$$> \tfrac{1}{2}\sum_{\substack{r\leqslant R\\ i=1,2}} P_X(\mathbb{C}_{ir}) \tag{34}$$

$$= \tfrac{1}{2}s((C_r)_{r\leqslant R}). \tag{35}$$

The first inequality results from our assumption that transfers are positive. The second results from (30) and the fact that the C_{ir} and $\mathbb{C}_{ir}$ sets are disjoint[12]. □

[12]To see that they are disjoint for all $r \leqslant R$, suppose that $x \in \mathbb{C}_{ir} \cap \mathbb{C}_{i'r'}$. Definition 6 rules out the

4.3.3 All Types Get Costed

Lemma 5. *If players reject bribes in every equilibrium of the bribe game of $\mathcal{M}$, then there is an infinite costing sequence of size one.*

The main idea of the proof of Lemma 5 is given by Claim 1. Claim 2 is required for infinite type spaces.

Claim 1. *Suppose the bribe game of $\mathcal{M}$ has no strict equilibrium in which bribes are accepted with strictly positive probability. For any pair of positive measure sets of uncosted types $U_{1*} \subseteq X_1$ and $U_{2*} \subseteq X_2$, either player 1 has a strategy a_1 that costs a positive measure of player 2's types $C_{2*} \subseteq U_{2*}$, or vice versa.*

Proof. Consider the "constrained" bribe game in which uncosted types of both players are constrained to accept bribes with probability 1, so that $a_i(x_i) = 1$ for all $x_i \in U_{i*}$, for $i = 1, 2$. Costed types $x_i \in X_i \backslash U_{i*}$ are unconstrained. Our assumptions ensure that those of Balder's equilibrium existence theorem (Balder, 1988, Theorem 3.1) apply to this constrained game,[13] so it has a weak equilibrium (a_1^*, a_2^*). There are no profitable deviations within this restricted strategy space. But a positive measure of constrained (uncosted) types accept bribes, so it cannot be a strict equilibrium of the unconstrained bribe game (this would violate the premise of the claim). Therefore, at least one player must have a weakly profitable deviation by a strictly positive measure of uncosted types $C_{i*} \subseteq U_{i*}$. Thus, a_{-i}^* is a valid costing strategy profile that induces C_{i*} to reject bribes, thereby getting costed, i.e. (26) holds for all $x_i \in C_{i*}$. □

If X is finite then all type profiles must all get costed in a finite number of rounds, giving a finite costing sequence of size 1, so Lemma 5 holds.

For the general case where X is not finite, we require Claim 2. Let O denote the set of finite costing sequences. Let $s^* = \sup_{(C_r)_{r \leqslant R} \in O} P_{X_1}(\cup_{r \leqslant R} C_{1r}) + P_{X_2}(\cup_{r \leqslant R} C_{2r})$ denote the supremum across finite costing sequences of the sum of the measure of costed player types.

Claim 2. *Suppose the bribe game of $\mathcal{M}$ has no strict equilibrium in which bribes are accepted with strictly positive probability. There exists an infinite costing sequence $(C_r)_{r \in \mathbb{N}}$ with $P_{X_1}(\cup_{r \in \mathbb{N}} C_{1r}) + P_{X_2}(\cup_{r \in \mathbb{N}} C_{2r}) = s^*$.*

possibility that $i = i'$ and $r \neq r'$, or that $r = r'$ and $i \neq i'$. If $i \neq i'$ and $r < r'$ then $x_i \in \cup_{s \leqslant r'} C_{is}$ so $x_i \notin U_{ir'}$. But then $x \notin C_{i'r'}$. The only remaining possibility is that $r = r'$ and $i = i'$.

[13] Balder's equilibrium existence theorem requires that the players' expected utility functions are $(C1')$ measurable with respect to types and actions; $(C1'')$ continuous with respect to actions; $(C1''')$ integrable over types and actions; that the prior is $(C2)$ absolutely continuous with respect to its marginals; and that the action spaces are $(C3)$ compact metric spaces.

Proof. Since s^* is the supremum, there must exist a sequence of finite costing sequences $((C_r^m)_{r\leqslant R_m})_{m\in\mathbb{N}}$, each of length R_m, such that $P_{X_1}(\cup_{t\leqslant R_m} C_{1r}^m) + P_{X_2}(\cup_{r\leqslant R_m} C_{2r}^m)$ converges to s^*. For each $i = 1, 2$, let $(C'_{ir})_{r\in\mathbb{N}}$ be the concatenation of the finite sequences $(C_{ir}^m)_{r\leqslant R_m}$, and let $(C_{ir})_{r\in\mathbb{N}}$ be the sequence defined by $C_{i0} = C'_{i0}$ and $C_{ir+1} = C'_{ir+1} \cup C_{ir}$. It is straightforward to check that $(C_r)_{r\in\mathbb{N}}$ is an infinite costing sequence, and that

$$P_{X_1}(\cup_{r\leqslant\sum_{q\leqslant m} R_q} C_{1r}) + P_{X_2}(\cup_{r\leqslant\sum_{q\leqslant m} R_q} C_{2r}) = P_{X_1}(\cup_{r\leqslant R_m} C_{1r}^m) + P_{X_2}(\cup_{r\leqslant R_m} C_{2r}^m) \xrightarrow{m\to\infty} s^*. \tag{36}$$

Hence $P_{X_1}(\cup_{r\in\mathbb{N}} C_{1r}) + P_{X_2}(\cup_{r\in\mathbb{N}} C_{2r}) = s^*$. □

Proof of Lemma 5. By Claim 2, there exists a costing sequence $(C_r)_{r\in\mathbb{N}}$ with $P_{X_1}(\cup_{r\in\mathbb{N}} C_{1r}) + P_{X_2}(\cup_{r\in\mathbb{N}} C_{2r}) = s^*$. Suppose for the sake of a contradiction that $s((C_r)_{r\in\mathbb{N}}) < 1$. Then both players must have a positive measure of uncosted types, U_{1*} and U_{2*}, so Claim 1 implies (without loss of generality) that player 1 has a strategy a_1^* that costs a set of positive measure of 2's uncosted types, i.e. there exists $C_{2*} \subseteq U_{2*}$ such that $a_1(x_1) = 1$ for all $x_1 \in U_{1*}$, and (26) holds for all $x_2 \in C_{2*}$. Construct a sequence of strategies $(a_1^r)_{r\in\mathbb{N}}$ for player 1 by amending a_1^* so that uncosted types accept, i.e. $a_1^r(x_1) := a_1^*(x_1) + I(x_1 \in U_{1r})(1 - a_1^*(x_1))$. This sequence converges pointwise to a_1^* because $a_1^*(x_1) = 1$ for all $x_1 \in U_{1*}$. So the Monotone Convergence Theorem and (26) imply that

$$\lim_{r\to\infty} \int_{X_1} \left(t_2(x) - \tfrac{1}{2}\right) a_1^r(x_1)\, dP_{X_1|X_2}(x_1|x_2) = \int_{X_1} \left(t_2(x) - \tfrac{1}{2}\right) a_1^*(x_1)\, dP_{X_1|X_2}(x_1|x_2) > 0 \tag{37}$$

for all $x_2 \in C_{2*}$. The Severini-Egorov Theorem tells us that the left side converges uniformly on a set $B_{2*} \subset C_{2*}$ with measure $P_{X_2}(B_{2*}) > 0$. It follows that there exists $S \in \mathbb{N}$ such that the types in B_{2*} can be costed in S rounds. There exists $R \geqslant S$ such that the difference between s^* and $P_{X_1}(\cup_{r\leqslant R} C_{1r}) + P_{X_2}(\cup_{r\leqslant R} C_{2r})$ is less than $P_{X_2}(B_{2*})$. This implies that $P_{X_1}(\cup_{r\leqslant R} C_{1r}) + P_{X_2}(\cup_{r\leqslant R} C_{2r}) + P_{X_2}(B_{2*}) > s^*$.

Define the finite sequence $(C'_r)_{r\leqslant R}$ that sequence extends $(C_r)_{r\in\mathbb{N}}$ by costing the B_{2*} types in period S: $C'_{1r} = C_{1r}$, and

$$C'_{2r} := \begin{cases} C_{2r} \cup B_{2*} \text{ if } r \geqslant S, \\ C_{2r} \text{ otherwise.} \end{cases} \tag{38}$$

Letting a_2^r denote player 2's costing strategy in round r of costing sequence $(C_{2r})_{r\in\mathbb{N}}$, the

sequence $(C'_r)_{r\leqslant R}$, together with strategy profiles $(a_1^r, a_2^r)_{r\leqslant R}$, satisfies the conditions of Definition 6. Hence, $(C'_r)_{r\leqslant R}$ is a well defined, finite costing sequence with

$$P_{X_1}(\cup_{r\leqslant R}C'_{1r}) + P_{X_2}(\cup_{r\leqslant R}C'_{2r}) \geqslant P_{X_1}(\cup_{r\leqslant R}C_{1r}) + P_{X_2}(\cup_{r\leqslant R}C_{2r}) + P_{X_2}(B_{2*}) > s^*, \quad (39)$$

thereby contradicting our assumption that s^* is the supremum. We conclude that $(C_r)_{r\in\mathbb{N}}$ is a costing sequence of size one. □

It follows from Lemma 4 and Lemma 5 that $c(\mathcal{M}) > 1/2$.

5 Extensions

In this section, we (simultaneously) generalise Theorem 1 to accommodate asymmetric roles among the players, and more than two players.

5.1 Rigged Poker

The basic Poker mechanism treats both players symmetrically, but this may not be desirable in practical applications. For example,

- If there are monitoring costs, then the designer might need to bias the mechanism in favour of the auditor to ensure that her payoff is large enough to incentivize her to bear these costs.
- If evidence can be fabricated, then the designer might need to bias the mechanism against the auditor to ensure that she does not find it profitable to fabricate evidence.
- If the agent can exert effort to comply (and hence decrease the chance the auditor obtains evidence), then rewarding the agent for not suppressing evidence undercuts the incentive to exert effort.

Each of these situations adds a constraint to the basic problem in (P1) that a player's expected transfer is not too high or too low, i.e., that $\bar{t}_i \geqslant \int_X t_i(x)\, dP(x) \geqslant \underline{t}_i$ for upper and lower bounds $\bar{t}_i$ and $\underline{t}_i > 0$. Any solution to this constrained problem is also a solution to an unconstrained problem in which the designer minimises a weighted sum of transfers with

weights λ_1 and λ_2:

$$\inf_{\mathcal{M}} \int_{X\times Y} [t_1(x,y)/\lambda_1 + t_2(x,y)/\lambda_2]\ dP(x,y) \text{ s.t. } \mathcal{M} \text{ blocks every side contract } \mathcal{S}. \qquad \text{(P2)}$$

Assigning a higher welfare weight λ_i has the effect of increasing the expected transfers that player i receives. Problem (P2) is solved by the following *Rigged Poker* mechanism.[14] Here, $\mathcal{P}(\lambda_i)$ denotes the power distribution with shape parameter λ_i.[15]

Rigged Poker, casino version

1. The house deals the players their hands x_1 and x_2 independently from $\mathcal{P}(\lambda_1)$ and $\mathcal{P}(\lambda_2)$.
2. If the players can agree how to split a prize of Π, then the Casino pays the split and the game ends.
3. Otherwise, play proceeds to showdown:
   (a) All cards are placed on the table facing up.
   (b) The house independently draws a pair of community handicap cards $y_1 \sim \mathcal{P}(\lambda_1)$ and $y_2 \sim \mathcal{P}(\lambda_2)$.
   (c) Player 1 wins $\Pi + \varepsilon$ if $y_1 x_1 \geqslant x_2$, and similarly for player 2.
   (d) Otherwise, nobody wins.

It is without loss of generality to normalise the weights so that $\lambda_1 + \lambda_2 = 1$. Theorem 2 shows that player i gets an expected payoff of $\lambda_i^2(\Pi+\varepsilon)$, and that the cost of the mechanism is $(\lambda_1^2 + \lambda_2^2)(\Pi + \varepsilon)$. If player i's upper constraint is binding, then the solution to (P2) is obtained by setting $\lambda_i = \sqrt{\overline{t}_i/(\Pi+\varepsilon)}$ and $\lambda_{-i} = 1-\lambda_i$. Similarly, $\lambda_i = \sqrt{\underline{t}_i/(\Pi+\varepsilon)}$ if player i's lower constraint is binding.[16]

Suppose we amend our running example from Section 2 so that Ray the Regulator wants to incentivise Dave the Developer to provide habitats for local species at a cost of £2m. If he does provide habitats, then there is a 10% chance that Anne finds hard evidence that the hotel is bad for biodiversity (i.e., monitoring is imperfect). If she reports evidence whenever

[14] The proof is a corollary of Theorem 2.

[15] A power distribution with shape parameter λ_i has support $[0,1]$ and CDF $F(z) = z^{\lambda_i}$.

[16] If none of the transfer constraints binds then $\lambda_1 = \lambda_2 = 1/2$. If both bind then they must both be lower bound constraints, in which case a solution can be obtained by setting $\lambda_1 = \lambda_2 = 1/2$ and increases the transfers by the smallest constant necessary to satisfy the lower bounds.

she finds it, then Dave's expected payoff is $10\% \times \int_{X \times Y} t_2(x,y) dP(x,y) + 90\% \times £10.1m - £2m$. If he does not provide habitats, then there is a 35% chance that Anne finds evidence, so his expected payoff is $35\% \times \int_{X \times Y} t_2(x,y) dP(x,y) + 65\% \times £10.1m$. Dave therefore provides habitats if and only if

$$\begin{aligned} &10\% \times \int_{X \times Y} t_2(x,y) dP(x,y) + 90\% \times £10.1m - £2m \\ &\geqslant 35\% \times \int_{X \times Y} t_2(x,y) dP(x,y) + 65\% \times £10.1m, \end{aligned} \tag{40}$$

which simplifies to $£2.1m \geqslant \int_{X \times Y} t_2(x,y) dP(x,y)$. Hence $\bar{t} = £2.1m$ The symmetric Poker mechanisms do not work because they give Dave an expected payoff of £2.525m: rewarding Dave for not bribing Anne undermines his incentive to provide habitats. However, if we choose $\lambda_D = \sqrt{2.1/10.1}$, then his expected payoff is exactly $10.1\lambda_D^2 = 2.1m$. Hence, the following Rigged Poker mechanism effectively preserves Dave's incentive to provide habitat by limiting his rewards:

Rigged Poker, moral hazard version

1. Ray announces the mechanism.
2. Dave chooses whether or not to conserve the local species.
3. Ray deals Anne and Dave their hands x_1 and x_2 independently from $\mathcal{P}(1 - \lambda_D)$ and $\mathcal{P}(\lambda_D)$ respectively.
4. If Anne fails to obtain evidence, or if Dave bribes Anne to hide any evidence that she does find, then Ray grants permission and the game ends.
5. Otherwise, if Anne reveals evidence that Dave did not conserve, then play proceeds to showdown:
   (a) All cards are placed on the table facing up.
   (b) Ray independently draws a pair of community handicap cards $y_1 \sim \mathcal{P}(1 - \lambda_D)$ and $y_2 \sim \mathcal{P}(\lambda_D)$.
   (c) Ray pays Anne a £10.1m reward if $y_1 x_1 \geqslant x_2$, and similarly for Dave.
   (d) Otherwise, Ray pays nothing.

Anne's expected payoff is £10.1m $\times(1-\sqrt{2.1/10.1})^2 \approx$ £2.989m, so the total cost of the Rigged mechanism is £5.089m, an increase of 1% relative to the symmetric mechanism.

5.2 n-player Poker

Appendix B shows that any mechanism that gives private information to only one of the two players must cost at least $1-1/e \approx 0.6321$, which is strictly greater than the cost of our Poker mechanisms. This suggests that giving private information to more players can increase information frictions and decrease the cost of deterring bribes. So what happens if we generalise Poker to more than two players? Does the presence of multiple potential whistleblowers make it easier to deter corruption?

Our definitions of mechanisms and side contracts naturally extend from 2 to a set N of players, to give the following n-player problem:

$$C(\bar{t}) = \inf_{\mathcal{M}} c(\mathcal{M}) \text{ s.t. } \mathcal{M} \text{ blocks every side contract } \mathcal{S}, \int t_i(x,y)\, dP(x,y) \leqslant \bar{t}_i \text{ for all } i \in N. \tag{P3}$$

Our analysis can be extended to accommodate lower transfer limits of the form $\int t_i(x)dP(x) \geqslant \underline{t}_i$. As in Section 5.1, we define an auxiliary, weighted cost problem by

$$\inf_{\mathcal{M}} c(\mathcal{M};\lambda) \text{ s.t. } \mathcal{M} \text{ blocks every side contract } \mathcal{S}, \tag{P4}$$

where $c(\mathcal{M};\lambda) = \sum_{i\in N}\int_X t_i(x)/\lambda_i \; dP_X(x)$ is the weighted cost of $\mathcal{M}$ with welfare weights $(\lambda_i)_{i\in N}$ normalised such that $\sum_{i\in N}\lambda_i = 1$. The n-player generalisation of Rigged Poker with showdown reward ε, denoted $\mathcal{M}^*(\varepsilon,\lambda)$ is defined by $Y^* = \times_{i\in N} Y_i^{**}$, $Y_i^* = X_i^* = [0,1]$ for all $i \in N$, Σ^* is the set of Lebesgue measurable subsets of $[0,1]^{2n}$, P_λ^* is the product of marginals $\mathcal{P}(\lambda_i)^2$ (i.e. x_i and y_i are both drawn from $\mathcal{P}(\lambda_i)$), and $t_i^*(x,y) = I(y_i x_i \geqslant \max_{i\neq j} x_j)(1+\varepsilon)$.

Theorem 2.

1. *Rigged Poker with showdown reward $\varepsilon > 0$ and welfare weights $(\lambda_i)_{i\in N}$ blocks all feasible side contracts, and gives player i an expected payoff of $(1+\varepsilon)\lambda_i^2$. It has weighted cost $c(\mathcal{M}^*(\varepsilon,\lambda);\lambda) = 1+\varepsilon$, and unweighted cost $c(\mathcal{M}^*(\varepsilon,\lambda)) = (1+\varepsilon)\sum_{i\in N}\lambda_i^2$.*

2. *For all welfare weights $(\lambda_i)_{i\in N}$, the infimum of (P4) is 1 and $\lim_{\varepsilon\to 0} c(\mathcal{M}^*(\varepsilon,\lambda);\lambda) = 1$.*

3. Hence Problem (P3) can be reformulated in terms of n-player Rigged Poker as

$$C(\bar{t}) = \inf_{\varepsilon>0,\lambda\in[0,1]^n} c(\mathcal{M}^*(\varepsilon,\lambda)) \text{ s.t. } \sum_{i\in N}\lambda_i = 1 \text{ and } (1+\varepsilon)\lambda_i^2 \leqslant \bar{t}_i \text{ for all } i\in N. \quad (41)$$

Proof. The proof of the first two points is a straightforward generalization of the proof of Theorem 1 – see Appendix A for details.

We now prove the last point. Let $\mathcal{M}^* = \{\mathcal{M}^*(\varepsilon,\lambda) : \varepsilon > 0, \lambda \in [0,1]^n, \sum_{i\in N}\lambda_i = 1\}$ be the set of Rigged Poker mechanisms. From the first part, we know that restricting attention to Rigged Poker is without loss of generality, i.e. adding a constraint, that $\mathcal{M}\in\mathcal{M}^*$, does not change the utility possibility set, and hence adding this constraint to Problem (P3) does not change its value. Moreover, every Rigged Poker mechanism $\mathcal{M}\in\mathcal{M}^*$ blocks side contracts, so the constraint $\mathcal{M}\in\mathcal{M}^*$ can in fact replace the side contract blocking constraint, giving

$$C(\bar{t}) = \inf_{\mathcal{M}\in\mathcal{M}^*} c(\mathcal{M}) \text{ s.t. } \int t_i(x,y)\, dP(x,y) \leqslant \bar{t}_i \text{ for all } i\in N. \quad (42)$$

Reformulating the choice as (ε,λ) instead of $\mathcal{M}$, and substituting $\int t_i(x,y)\,dP(x,y) = (1+\varepsilon)\lambda_i^2$ into the constraint gives (41). □

Just like regular Poker, only the player with the best hand stands a chance of winning. Unlike regular Poker, cards are drawn from different decks with different distributions, and the player with the best hand does not necessarily win: she only wins if her hand beats the second highest hand by a large enough margin. The (unweighted) cost of the mechanism is minimised when all players are symmetric (i.e. when $\lambda_i = 1/n$ for all $i\in N$). In this case, the cost of the mechanism is $1/n$.

This means that the presence of multiple potential whistleblowers decreases the cost of deterring corruption. This is because n-Player Poker creates a more severe adverse selection problem. We discuss the implications for information design and corruption in sections 6 and 7 respectively.

6 Literature

We contribute to an extensive economic literature on corruption.[17] Tirole (1986) developed the first principal-monitor-agent model of corruption, and Laffont and Martimort (1997)

[17] See Tirole (1993) and Aidt (2003) for surveys.

were the first to study the role of (exogenous) information frictions in corrupt side-contracts. Asseyer (2020) and Mookherjee and Tsumagari (2023) study the impact of information frictions arising from the agent's private information about his type. Asseyer (2020) shows how the principal can utilise this friction by carefully designing the monitor's monitoring technology, thereby reducing the costs of deterring bribes. Mookherjee and Tsumagari (2023) show that the principal can use this friction to deter bribes and extortion by increasing the bargaining power of the monitor. But neither consider the possibility of using random contracts to create endogenous information frictions. Angelucci and Russo (2022) show that a designer can use self-reporting schemes to create an equilibrium that is free from bribes and extortion. But this equilibrium is not guaranteed to be unique. If all evidence is soft, then Strulovici (2021) finds that it is impossible to incentivise truthful reporting by corruptible monitors without resorting to unbounded rewards and punishments. But he does not consider the use of random incentives to reduce enforcement costs.

The closest papers to ours are von Negenborn and Pollrich (2020) and Ortner and Chassang (2018). von Negenborn and Pollrich (2020) use random incentives and private messages to create a lemons problem that deters agent-monitor collusion at arbitrarily small cost. Their scheme is simple—it has a single informed player with a binary message —but its practical use is limited by the fact that it uses infinitely large rewards and punishments (Appendix C describes a similar mechanism that would solve our corruption problem at arbitrarily low cost). When rewards and punishments are bounded, their mechanism is no longer feasible. Appendix C demonstrates how the their logic can be extended to produce a mechanism with bounded payoffs, which is feasible, but this mechanism is not optimal: it costs 3/4, whilst the Poker mechanism costs only 1/2. It improves on the binary scheme by giving private information to both players, and increasing the number of states from two to a continuum. Appendix C shows that no mechanism with a single informed player with a binary message can cost less than 3/4. The details of our respective settings are also different, e.g. they do partial implementation with soft evidence, whereas we do full implementation with hard evidence.

Ortner and Chassang (2018) deter collusion in a moral hazard setting by paying the monitor a random wage. Doing so creates a screening problem for the agent because the monitor can demand a large bribe by pretending to have a high wage, even if her wage is actually small. They find that the optimal wage distribution creates unit-elastic demand for hiding evidence and thereby maximises the monitor's information rent. If monitoring is perfect, then their mechanism incentivises the agent to comply, so the monitor never obtains

incriminating evidence on the equilibrium path. There is no need for them to deter bribes per se, because they only occur off-path. But their mechanism breaks down if monitoring is imperfect because the agent is always better off offering at least a small bribe whenever the monitor obtains evidence, and there is a strictly positive chance that the monitor accepts it. This can potentially upset the agent's incentive to comply in the first place. Indeed, the same is true for any wage distribution whose support strictly contains the agent's fine. We address this by allowing the designer to pay the agent a reward too. Doing so allows her to correlate the agent's reward with the monitor's wage, thereby creating a lemons problem for the agent that deters on and off path bribes. Constant Handicap Poker reduces the cost of deterring bribes even further by giving the agent some private information about his reward, so as to create a two-sided lemons problem.[18] Handicap Poker deters a larger class of arbitrated bribes by adding a two-sided screening problem with the same one-sided rent-maximising payoff distribution (based on the reciprocal of a uniform distribution) as Ortner and Chassang (2018) and others (Condorelli and Szentes, 2020; Ali et al., 2022; Garrett et al., 2023).

Our results contribute to the literature on robust mechanism design (see Carroll, 2019, for a survey) and worst-case information structures (see Brooks and Du, 2025). A large literature has shown how private information can lead to market inefficiencies through channels such as adverse selection (Akerlof, 1970), screening (Myerson and Satterthwaite, 1983), and contagion (Morris and Shin, 2012). Carroll (2016) studies the role of these frictions in two-player binary-action supermodular games (including bilateral trade at a fixed price). Surprisingly, he finds that for every private information structure, there exists a public information structure with the same lowest expected surplus across all equilibria. This suggests that strictly private information frictions need not play a role in the worst-case information structures of bilateral trade at a fixed price. We find that this result no longer holds when the terms of trade are negotiable. A worst-case public information structure for one price is typically not worst-case for other prices. By contrast, Poker generates an information structure with private messages that is simultaneously worst-case for all prices (and mediated side contracts) and exhibits all of the classical information frictions. Moreover, we adapt the proof technique of Carroll's Proposition 3.1 to circumvent a transfinite induction problem, allowing us to generalize our results from finite to arbitrary information structures.

In the context of public goods provision, Brooks and Du (2023, 2024) solve for an infor-

[18] Appendix B shows that the cost of solving (P1) with one-sided private information is $1 - 1/e \approx 0.6321$, whilst Poker costs 0.5.

mation structure that minimises the maximum attainable social surplus from any equilibrium of any mechanism. Our design problem is closely related to theirs. Their players have an unknown value for a public good, whereas ours have an unknown outside option for reaching agreement. In their model, the designer chooses a mechanism to allocate expenditure, subject to incentive constraints. In ours, the mafia chooses a mechanism to allocate the surplus generated by agreement, subject to incentive constraints. In their model, Nature chooses an information structure to adversarially minimise attainable social surplus, subject to a constraint that expected value of the public good is above a given threshold. In ours, the (grand) designer chooses an information structure to minimise the expected value of outside options (i.e. maximise the expected net value of agreement), subject to the constraint that there are no feasible contracts for the mafia to offer. Like them, we find that a worst-case information structure (i.e. Handicap Poker) that completely destroys the gains from coordination, even though it is common knowledge that the coordination generates strictly positive surplus with probability arbitrarily close to 1. Nonetheless, there are two key differences between Poker and Brooks and Du's potential minimising information structure. First, their information structure features correlated messages, whilst Poker has independent messages. Correlated messages are useless in our environment because neither the players nor the Mafia face ex post liability constraints, so the Mafia could use a Cremer and McLean (1988)-type mechanism to recover players' private information. Second, their players' values for the public good depend on the relative size of their own message to the sum of all messages, whereas our players' showdown payoffs from Poker depend on the relative size of their own message to the maximum of all other messages. This reflects the fact that our players require unanimous agreement to generate a fixed amount of surplus, so the amount of realisable surplus depends on the most reluctant player, i.e. who has the best outside option. By contrast, surplus from public good provision in Brooks and Du scales linearly with the sum of expenditures.

Our problem fits into a larger class of *general* mechanism design problems in which the designer chooses both transfers and information (See, e.g., Bergemann and Morris, 2019; Morris et al., 2022; Moriya and Yamashita, 2020). One example is the stochastic ranking scheme of Halac et al. (2021). Their goal is to eliminate all shirking equilibria among a team of workers, which has parallels with eliminating all bribe equilibria in a monitoring hierarchy. Both shirking and bribing are socially inefficient. Their ranking schemes are superficially similar to our Poker mechanism insofar as (i) all players are told their own types; and (ii) the transfers are chosen so that the highest type has a dominant strategy, and a cascade of the subsequent types leads to a unique equilibrium. However, bribes are harder to deter

because there are gains from trade, whereas shirking is purely a coordination failure. In other words, our problem is about thwarting coordination, whereas their problem is about ensuring coordination. Our schemes are therefore different. First, our Poker mechanism creates a lemons problem, which frustrates coordination. By contrast, in ranking schemes, transfers are not interdependent, and thus do not create lemons problems. Second, Poker hands are independent, and thus can not be screened by a Cremer and McLean (1988) style mechanism. In contrast, optimal ranking schemes involve correlated types.

7 Conclusion

In this article, we have proposed a new, Poker-like mechanism for deterring bribes. We showed that this mechanism is theoretically optimal and deters a wide class of corrupt side contracts. It can incentivise compliance even when monitoring is imperfect, and does not rely on using arbitrarily large rewards and punishments. The mechanism also gives insights into 'worst-case' information structures and gives an upper bound on the amount of surplus lost to information frictions in a class of bargaining and public goods games. We conclude by suggesting avenues for future research.

One of the biggest challenges in dealing with corruption is that it is very diverse, and it is difficult to understand and close every possible loophole. Indeed, most real world institutions are highly vulnerable to a single dishonest judge or auditor. We have modelled the mafia as having full commitment power in side contracts, and we have ruled out a large class of loopholes. However, we assumed the mafia is incapable of violence, and incapable of committing to punishing bribe rejections. This raises the question of whether it is possible to deter a more omnipotent mafia.

One potential source of loopholes in our Poker mechanisms is that decisions are close to a knife-edge: there are always profitable deviations from signing a corrupt agreement, but the deviation profits (determined by ε) are small, to keep the cost of the mechanism down. If corruption opportunities arise regularly, then the short-term gain from a profitable deviation is outweighed by the long-term benefit of maintaining a reputation for being agreeable to corrupt deals. One solution would be to avoid repeat encounters through random assignment of auditors. However, this might be difficult if there are few auditors, or there are a small number of large firms to be regulated. There might be better solutions, such as combining multiple projects into a single Poker game.

Another possible loophole is the "casino" which administers the Poker game. We have

replaced one problematic assumption, that auditors can not be bribed, with another problematic assumption, that casinos can not be bribed. It is unclear whether the public can monitor a Poker game and rule out the casino breaking its own rules. One solution might be to eliminate the casino altogether, and replace it with a computer system based on a cryptographic communication protocol. Indeed, Goldwasser and Micali (1982) proposed a protocol for playing "mental Poker" without the help of a casino. Their proposal focuses on unilateral cheating rather than collusion, so they do not prevent the players from credibly showing each other their hands, e.g. with zero-knowledge proofs. A large literature has extended their technique to "receipt-free electronic voting", which prevents vote buying. A similar extension ought to allow Handicap Poker without a casino.

Another direction is adapting the mechanism to other scenarios. The fact that the expected cost of n-player Poker decreases inversely in the number of players suggests that our Poker mechanism may be particularly useful in regulatory contexts where large numbers of monitors or whistle blowers are available. If evidence is perfectly correlated, then the regulator version of 2-player Poker extends very naturally to n-players because there are still only two evidential outcomes: either all monitors receive evidence, or none of them do. But what if it is not perfectly correlated? What rewards or punishments should be paid if some but not all monitors report evidence?

A variant of Handicap Poker might be applicable to other scenarios involving coalition formation, such as international agreements. Efficient cooperation is often thwarted by the threat of subcoalitional deviations.[19] The literature on coalition formation therefore seeks to predict which outcomes can be made stable against such deviations (see Demuynck et al. (2019) for a recent example). But to the best of our knowledge, this literature has not yet considered endogenous private information and random incentives. By increasing information frictions within subcoalitions, a Poker-like mechanism may be able to stabilise a broader range of efficient outcomes.

References

Admati, A. and M. Hellwig (2024). *The Bankers' New Clothes: What's Wrong with Banking and What to Do about It*. Princeton University Press.

[19]For instance, the World Trade Organisation aims to maximise global welfare by reducing trade barriers, but is increasingly undermined by the formation of regional Free Trade Agreements and Customs Unions.

Aidt, T. S. (2003). Economic analysis of corruption: a survey. *Economic Journal 113*(491), F632–F652.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics 84*(3), 488–500.

Ali, S. N., N. Haghpanah, X. Lin, and R. Siegel (2022). How to sell hard information. *Quarterly Journal of Economics 137*(1), 619–678.

Angelucci, C. and A. Russo (2022). Petty corruption and citizen reports. *International Economic Review 63*(2), 831–848.

Asseyer, A. (2020). Collusion and delegation under information control. *Theoretical Economics 15*(4), 1547–1586.

Attar, A., L. Bozzoli, and R. Strausz (2025). Mediated renegotiation. TSE Working Paper n. 24–1522, Toulouse School of Economics.

Balder, E. J. (1988). Generalized equilibrium results for games with incomplete information. *Mathematics of Operations Research 13*(2), 265–276.

Baliga, S. and T. Sjöström (1998). Decentralization and collusion. *Journal of Economic Theory 83*(2), 196–232.

Barofsky, N. (2012). *Bailout: An Inside Account of How Washington Abandoned Main Street While Rescuing Wall Street*. Free Press.

Beck, J. (2000). The false claims act and the english eradication of qui tam legislation. *North Carolina Law Review 78*(3), 539–642.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Brooks, B. and S. Du (2023). Robust mechanisms for the financing of public goods. *Technical Report, The University of Chicago and University of California—San. Diego.*.

Brooks, B. and S. Du (2024). On the structure of informationally robust optimal mechanisms. *Econometrica 92*(5), 1391–1438.

Brooks, B. and S. Du (2025). Simplicity and portability in mechanism design: A case for (and against) the worst case.

Butler, S. D. (1935). *War is a Racket*. Round Table Press.

Carroll, G. (2016). Informationally robust trade and limits to contagion. *Journal of Economic Theory 166*(C), 334–361.

Carroll, G. (2019). Robustness in mechanism design and contracting. *Annual Review of Economics 11*(1), 139–166.

Condorelli, D. and B. Szentes (2020). Information design in the holdup problem. *Journal of Political Economy 128*(2), 681–709.

Cremer, J. and R. P. McLean (1988). Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions. *Econometrica 56*(6), 1247–1257.

Demuynck, T., P. J.-J. Herings, R. D. Saulle, and C. Seel (2019). The myopic stable set for social environments. *Econometrica 87*(1), 111–138.

Dudley, R. (2002). *Real Analysis and Probability*. Cambridge University Press.

Garrett, D. F., G. Georgiadis, A. Smolin, and B. Szentes (2023). Optimal technology design. *Journal of Economic Theory 209*, 105621.

Goldwasser, S. and S. Micali (1982). Probabilistic encryption and how to play mental poker keeping secret all partial information. In *14th Symposium on Theory of Computing*, pp. 365–377. ACM Press.

Halac, M., E. Lipnowski, and D. Rappoport (2021). Rank uncertainty in organizations. *American Economic Review 111*(3), 757–86.

Herman, E. and N. Chomsky (1988). *Manufacturing Consent: The Political Economy of the Mass Media*. Random House.

Krishna, V. (2009). *Auction theory*. Academic press.

Laffont, J.-J. and D. Martimort (1997). Collusion under asymmetric information. *Econometrica 65*(4), 875–912.

Laffont, J.-J. and D. Martimort (2000). Mechanism design with collusion and correlation. *Econometrica 68*(2), 309–342.

Mookherjee, D. and M. Tsumagari (2023). Regulatory mechanism design with extortionary collusion. *Journal of Economic Theory 208*, 105614.

Moriya, F. and T. Yamashita (2020). Asymmetric-information allocation to avoid coordination failure. *Journal of Economics & Management Strategy 29*(1), 173–186.

Morris, S., D. Oyama, and S. Takahashi (2022). On the joint design of information and transfers. *Available at SSRN 4156831*.

Morris, S. and H. S. Shin (2012). Contagious adverse selection. *American Economic Journal: Macroeconomics 4*(1), 1–21.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research 6*(1), 58–73.

Myerson, R. B. and M. A. Satterthwaite (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory 29*(2), 265 – 281.

Ortner, J. and S. Chassang (2018). Making corruption harder: Asymmetric information, collusion, and crime. *Journal of Political Economy 126*(5), 2108–2133.

Royden, H. L. and P. Fitzpatrick (1988). *Real analysis*, Volume 32. Macmillan New York.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica 50*(1), 97–109.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under "almost common knowledge". *American Economic Review 79*(3), 385–391.

Strulovici, B. (2021). Learning and corruption on monitoring chains. In *AEA Papers and Proceedings*, Volume 111, pp. 544–548.

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law, Economics, & Organization 2*(2), 181–214.

Tirole, J. (1993). Collusion and the theory of organizations. In J.-J. Laffont (Ed.), *Advances in Economic Theory: Sixth World Congress*, pp. 151–213. Cambridge University Press.

van der Kolk, B. (2015). *The Body Keeps the Score: Brain, Mind, and Body in the Healing of Trauma*. New York: Viking.

von Negenborn, C. and M. Pollrich (2020). Sweet lemons: Mitigating collusion in organizations. *Journal of Economic Theory 189*, 105074.

von Neumann, J. and O. Morgenstern (1953). *Theory of Games and Economic Behavior (3rd ed.)*. Princeton university press.

Wheeler, W. T. (Ed.) (2011). *The Pentagon Labyrinth: 10 Short Essays to Help You Through It*. Washington, D.C.: Center for Defense Information.

Appendix A n-player Poker

Let $\lambda_1, \ldots, \lambda_n$ be welfare weights for each of the players, and assume that $\sum_{i \in N} \lambda_i = 1$. The proof in Section 4.1 that Handicap Poker deters all possible side contracts generalises in a straightforward way. Here we give only the key equations and mathematical derivations.

A.1 n-player Poker Blocks All Side Contracts

The n-player Poker mechanism $\mathcal{M}^*(\varepsilon, \lambda)$ is equivalent to the mechanism that draws x_i and y_i uniformly and pays transfers

$$\int_Y t_i(x,y)\, dP_Y(y) = (1+\varepsilon) P_Y\left(y^{1/\lambda_i} x_i^{1/\lambda_i} \geqslant \max_{j \neq i} (x_j)^{1/\lambda_j} \right) \tag{43}$$

$$= (1+\varepsilon) P_Y\left(y \geqslant \max_{j \neq i} (x_j)^{\lambda_i/\lambda_j} / x_i \right) \tag{44}$$

$$= (1+\varepsilon) \max\left\{ 0, 1 - \max_{j \neq i} (x_j)^{\lambda_i/\lambda_j} / x_i \right\}. \tag{45}$$

Player i's expected payoff if he reports x_i' to the side contract when his true hand is x_i is

$$V_i(x_i, x_i') = \int_{X_{-i}} b_i(x_i', x_{-i}) + (1 - a(x_i', x_{-i}))(1+\varepsilon) \max\left\{0, 1 - \max_{j \neq i}\left\{x_j^{\lambda_i/\lambda_j}\right\}/x_i\right\} \, dx_{-i} \tag{46}$$

$$= \int_{X_{-i}} b_i(x_i', x_{-i}) \, dx_{-i}$$

$$+ (1+\varepsilon) \sum_{j \neq i} \int_0^{x_i^{\lambda_i/\lambda_j}} \int_{\times_{k \neq i,j}\left[0, x_j^{\lambda_k/\lambda_j}\right]} (1 - a(x_i', x_{-i}))(1 - x_j^{\lambda_i/\lambda_j}/x_i) \, dx_{-i,j} \, dx_j \tag{47}$$

$$= \int_{X_{-i}} b_i(x_i', x_{-i}) \, dx_{-i} + (1+\varepsilon) \sum_{j \neq i} \int_0^{x_i^{\lambda_j/\lambda_i}} (1 - A_{ij}(x_i', x_j))(1 - x_j^{\lambda_i/\lambda_j}/x_i) \, dx_j, \tag{48}$$

where $A_{ij}(x_i, x_j) := 1 - \int_{\times_{k \neq i,j}[0, x_j^{\lambda_i/\lambda_k}]} (1 - a(x)) \, dx_{-i,j}$ is the probability of accepting a split when i has the greatest weighted hand, followed by j. Leibnitz rule gives

$$\frac{\partial}{\partial x_i} V_i(x_i, x_i') = (1+\varepsilon) \sum_{j \neq i} \tfrac{\lambda_j}{\lambda_i} x_i^{\lambda_j/\lambda_i - 1} (1 - x_i^{\lambda_i \lambda_i^{-1}}/x_i)(1 - A_{ij}(x_i', x_j)) \, dx_j \tag{49}$$

$$+ (1+\varepsilon) \sum_{j \neq i} \int_0^{x_i^{\lambda_j/\lambda_i}} x_j^{\lambda_i/\lambda_j}/x_i^2 \; (1 - A_{ij}(x_i', x_j)) \, dx_j \tag{50}$$

$$= (1+\varepsilon) \tfrac{1}{x_i} \sum_{j \neq i} \int_0^{x_i^{\lambda_j/\lambda_i}} x_j^{\lambda_i/\lambda_j}/x_i \; (1 - A_{ij}(x_i', x_j)) \, dx_j. \tag{51}$$

The envelope theorem gives the following generalization of Lemma 1.

$$W_i'(x_i) = (1+\varepsilon) \frac{1}{x_i} \sum_{j \neq i} \int_0^{x_i^{\lambda_j/\lambda_i}} x_j^{\lambda_i/\lambda_j}/x_i \; A_{ij}(x_i, x_j) \, dx_j. \tag{52}$$

The side player participation constraint (SPP) for the strongest hand gives

$$W_i(1) \geqslant (1+\varepsilon)\sum_{j\neq i}\int_0^1 (1-x_j^{\lambda_i/\lambda_j})\prod_{k\neq i,j} x_j^{\lambda_k/\lambda_j}\,dx_j \tag{53}$$

$$= (1+\varepsilon)\sum_{j\neq i}\int_0^1 (1-x_j^{\lambda_i/\lambda_j}) x_j^{\sum_{k\neq i,j}\lambda_k/\lambda_j}\,dx_j \tag{54}$$

$$= (1+\varepsilon)\sum_{j\neq i}\int_0^1 \left[x_j^{\sum_{k\neq i,j}\lambda_k/\lambda_j} - x_j^{\sum_{k\neq j}\lambda_k/\lambda_j}\right]\,dx_j \tag{55}$$

$$= (1+\varepsilon)\sum_{j\neq i}\left[\frac{x_j^{\sum_{k\neq i,j}\lambda_k/\lambda_j+1}}{\sum_{k\neq i,j}\lambda_k/\lambda_j+1} - \frac{x_j^{\sum_{k\neq j}\lambda_k/\lambda_j+1}}{\sum_{k\neq j}\lambda_k/\lambda_j+1}\right]_0^1 \tag{56}$$

$$= (1+\varepsilon)\sum_{j\neq i}\frac{1}{(1-\lambda_i-\lambda_j)/\lambda_j+1} - \frac{1}{(1-\lambda_j)/\lambda_j+1} \tag{57}$$

$$= (1+\varepsilon)\sum_{j\neq i}\left(\frac{\lambda_j}{1-\lambda_i} - \frac{\lambda_j}{1}\right) \tag{58}$$

$$= (1+\varepsilon)\left(\frac{1-\lambda_i}{1-\lambda_i} - (1-\lambda_i)\right) \tag{59}$$

$$= (1+\varepsilon)\lambda_i. \tag{60}$$

Together with the fundamental theorem of calculus, this gives

$$\int_0^1 W_i(x_i)\,dx_i = (1+\varepsilon)\lambda_i - (1+\varepsilon)\sum_{j\neq i}\int_0^1\int_0^{x_i^{\lambda_j/\lambda_i}} x_j^{\lambda_i/\lambda_j}/x_i\;(1-A_{ij}(x_i,x_j))\,dx_j\,dx_i. \tag{61}$$

Integrating (48) gives

$$\begin{aligned}\int_0^1 W_i(x_i)dx_i = &\int_X b_i(x)\,dx \\ &+ (1+\varepsilon)\sum_{j\neq i}\int_0^1\int_0^{x_i^{\lambda_j/\lambda_i}} (1-x_j^{\lambda_i/\lambda_j}/x_i)\;(1-A_{ij}(x_i,x_j))\,dx_j\,dx_i.\end{aligned} \tag{62}$$

Substituting this into (61) gives

$$\int_X b_i(x)\,dx \geqslant (1+\varepsilon)\lambda_i - (1+\varepsilon)\sum_{j\neq i}\int_0^1\int_0^{x_i^{\lambda_j/\lambda_i}} x_j^{\lambda_i/\lambda_j}/x_i\ (1-A_{ij}(x_i,x_j))\,dx_j\,dx_i \tag{63}$$

$$-(1+\varepsilon)\sum_{j\neq i}\int_0^1\int_0^{x_i^{\lambda_j/\lambda_i}} (1-x_j^{\lambda_i/\lambda_j}/x_i)\ (1-A_{ij}(x_i,x_j))\,dx_j\,dx_i \tag{64}$$

$$\geqslant (1+\varepsilon)\lambda_i - (1+\varepsilon)\sum_{j\neq i}\int_0^1\int_0^{x_i^{\lambda_j/\lambda_i}} (1-A_{ij}(x_i,x_j))\,dx_j\,dx_i \tag{65}$$

$$\geqslant (1+\varepsilon)\lambda_i - (1+\varepsilon)\int_{x_i^{1/\lambda_i}\geqslant \max_j x_j^{1/\lambda_j}} 1-a(x)\,dx. \tag{66}$$

So the mafia's total transfers to the players are at least

$$\int_X \sum_i b_i(x)\,dx \geqslant (1+\varepsilon)\sum_i \lambda_i - \int_X (1-a(x))\,dx \geqslant (1+\varepsilon)\int_X a(x)\,dx \geqslant \int_X a(x)\,dx. \tag{67}$$

The last inequality is strict if $\int a(x)\,dx > 0$, as per Lemma 2.

By the same reasoning that concludes Section 4.1, this implies one of two possibilities. If the side contract delivers agreements with positive probability, then the mafia makes a loss, violating their participation constraint. Otherwise, the side contract creates no surplus, violating the side surplus constraint.

A.2 The Cost of n-player Poker

Derivations analogous to equations (53)–(60) give $\bar{W}_i(x_i) = (1+\varepsilon)\lambda_i(x_i)^{1/\lambda_i - 1}$. Hence $\int_0^1 \bar{W}_i(x_i)\,dx_i = (1+\varepsilon)\lambda_i^2$. The weighted cost of n-player Poker is therefore $(1+\varepsilon)\sum_{i\in N}\lambda_i = 1+\varepsilon$. The unweighted cost is $(1+\varepsilon)\sum_{i\in N}\lambda_i^2$. If all players are weighted equally, then the cost is $\frac{1+\varepsilon}{n}$.

A.3 n-player Poker is Optimal

If $b \in \mathbb{R}^n_+$ is an exogenous split, then the weighted cost of buying out the cheapest player is $\min_{i\in N}\frac{b_i}{\lambda_i}$. The most costly split to block is $\bar{b} = \lambda$.[20] The players' ex ante utility function in

[20] The most costly split solves $\max_{b\in\mathbb{R}^n_+}[\min_{i\in N} b_i/\lambda_i]$ s.t. $\sum_{i\in N} b_i = 1$.

Definition 5 in the bribe game generalises to

$$U_i(a_i, a_{-i}) = \int_X \left[\bar{b}_i \prod_{j \in N} a_j(x_j) + \left(1 - \prod_{j \in N} a_j(x_j) \right) t_i(x) \right] dP_X(x). \tag{68}$$

Lemma 3 generalises in a straightforward way. Equation (26) in the definition of costing sequences generalises to

$$\int_{X_{-i}} [t_i(x) - \bar{b}_i] \prod_{j \neq i} a_j(x_j) \; dP_{X_{-i}|X_i}(x_{-i}|x_i) > 0 \qquad \forall x_i \in C_{ir}. \tag{69}$$

We have $\mathbb{C}_{ir} := C_{ir} \times \bigtimes_{j \neq i} U_{jr}$. Lemma 4 becomes

Lemma 4′. Let $(C_r)_{r \leqslant R}$ be any finite costing sequence of the bribe game of a mechanism $\mathcal{M}$ that blocks all side contracts. Then the weighted cost of $\mathcal{M}$ exceeds the size of $(C_r)_{r \leqslant R}$, i.e. $c(\mathcal{M}; \lambda) > s((C_r)_{r \leqslant R})$.

The inequalities in the proof generalise to

$$\int_{X_{-i}} t_i(x) \; dP_{X_{-i}|X_i}(x_{-i}|x_i) > \lambda_i P_{X_{-i}|X_i}(\times_{j \neq i} U_{jr} | x_i), \tag{70}$$

$$\int_{C_{ir} \times X_{-i}} t_i(x) \; dP_X(x) > \lambda_i P_X(\mathbb{C}_{ir}). \tag{71}$$

Taking the weighted sum of the inequalities over players $i \in N$ and rounds $r \leqslant R$ gives

$$\begin{aligned}
c(\mathcal{M}; \lambda) &= \sum_{i \in N} \int_X t_i(x)/\lambda_i \; dP_X(x) && (72)\\
&\geqslant \sum_{i \in N} \int_{\cup_{r \leqslant R} C_{ir} \times X_{-i}} t_i(x)/\lambda_i \; dP_X(x) && (73)\\
&= \sum_{\substack{i \in N \\ r \leqslant R}} \int_{C_{ir} \times X_{-i}} t_i(x)/\lambda_i \; dP_X(x) && (74)\\
&> \sum_{\substack{i \in N \\ r \leqslant R}} \lambda_i/\lambda_i \times P_X(\mathbb{C}_{ir}) && (75)\\
&= s((C_r)_{r \leqslant R}). && (76)
\end{aligned}$$

Finally, in the proof of Lemma 5, the inequality (37) becomes

$$\lim_{r\to\infty}\int_{X_{-i}}[t_i(x)-\bar{b}_i]\prod_{j\neq i}a_j^r(x_j)\,dP_{X_{-i}|X_i}(x_{-i}|x_i)$$

$$=\int_{X_{-i}}[t_i(x)-\bar{b}_i]\prod_{j\neq i}a_j^*(x_j)\,dP_{X_{-i}|X_i}(x_{-i}|x_i) \tag{77}$$

$$>0 \tag{78}$$

for all $x_i \in C_{i*}$.

Appendix B One-sided Mechanisms

Proposition 1. *Any solution to the problem in* (P1) *with the additional constraint that* $|S_2|=1$ *costs at least* $1-1/e$*, where* e *is Euler's number.*

Proof. The idea of the proof is similar to that of Proposition 1 in Ortner and Chassang (2018).[21] Consider the class of "price" side contracts of the form $a(x)=I(t_1(x)\leqslant p)$, $b_1(x)=p$, $b_2(x)=1-p-\varepsilon$, where p is a bribe (or price). By construction, price contracts satisfy (SPP) and (SMP) participation constraints for player 1, and the mafia. Player 2's expected payoff from a contract with price p is $F(p)(1-p-\varepsilon)+(1-F(p))T_2(p)$, where $F(p):=P_X(x\in X: t_1(x)\leqslant p)$ is the CDF of player 1's transfer, and $T_2(p):=\mathbb{E}[t_2(x)|t_1(x)\geqslant p]$ is player 2's expected transfer conditional on player 1's transfer exceeding the price. The fact that transfers are positive means that $F(0)=0$ and $T_2(0)=\mathbb{E}[t_2(x)]$. Thus, a mechanism blocks these price contracts if and only if it violates player 2's (SPP) constraint for all $\varepsilon>0$. This occurs only if $F(p)(1-p)+(1-F(p))T_2(p)\leqslant T_2(0)$ for all $p\geqslant 0$, which rearranges to

$$F(p)\leqslant\frac{T_2(0)-T_2(p)}{1-p-T_2(p)}. \tag{79}$$

Since the designer wants to minimise the expected value of t_1, she can do no better than choosing P so that (79) holds with equality. In this case, the support of F is bound above by $1-T_2(0)$, so the right side of (79) is increasing in $T_2(p)$. The value of $T_2(p)$ that minimises player 1's expected payoff is therefore 0 for all $p>0$. So player 1's expected payoff is bound

[21] Our mechanism nonetheless differs qualitatively from theirs because player 2's transfer t_2 is endogenous and depends on player one's message x_1; in their mechanism it is exogenous and constant.

below by

$$\mathbb{E}[t_1(x)] = \int_0^{1-T_2(0)} p\, dF(p) = T_2(0) \int_0^{1-T_2(0)} \frac{p}{(1-p)^2}\, dp. \tag{80}$$

The total cost of the mechanism is bound below by

$$\mathbb{E}[t_1(x) + t_2(x)] = T_2(0) \int_0^{1-T_2(0)} \frac{p}{(1-p)^2}\, dp + T_2(0), \tag{81}$$

which is minimised by $T_2(0) = 1/e$ and takes minimal value $1 - 1/e$. □

An optimal one-sided mechanism has P uniform over $[0,1]$, $t_1(x) = 1 - \frac{1}{ex}$ and $t_2(x_1) = I(x_1 \leqslant 1/e)$. This yields the same unit-elastic demand for agreements as the optimal distributions obtained in Ortner and Chassang (2018), Condorelli and Szentes (2020) and others.

Appendix C Binary Mechanisms

Proposition 2. *Any solution to the problem in* (P1) *with* $|S_1| = 2$ *and* $|S_2| = 1$ *costs at least 3/4.*

Proof. Suppose that the messages are labelled H (high) and L (low). Suppose without loss of generality that $t_1(H) \geqslant t_1(L)$. The designers problem reduces to choosing transfers $t_1(H)$, $t_1(L)$, $t_2(H)$, and $t_2(L)$, and a CDF $F(t_1(L))$ to minimise

$$F(t_1(L))(t_1(L) + t_2(L)) + (1 - F(t_1(L))(t_1(H) + t_2(H)) \tag{82}$$

subject the side contract blocking constraint.

The same reasoning as the proof of Proposition 1 gives us that (79) holds at $F(t_1(L))$ and $F(t_1(H)) = 1$. The latter gives $T_2(0) = 1 - t_1(H)$. The former then gives $F(t_1(L)) = \frac{1-t_1(H)}{1-t_1(L)}$. The lower bound on the designer's cost reduces to

$$\frac{1-t_1(H)}{1-t_1(L)} t_1(L) + \left(1 - \frac{1-t_1(H)}{1-t_1(L)}\right) t_1(H) + 1 - t_1(H) \tag{83}$$

$$= \frac{1-t_1(H)}{1-t_1(L)} (t_1(L) - t_1(H)) + 1 \tag{84}$$

which is minimised by setting $t_1(L) = 0$ and $t_1(H) = 1/2$. The cost must therefore be greater than 3/4. □

A feasible mechanism that attains this lower bound is the following "coin toss mechanism": $p^*(H) = 1/2$, $t_1^*(L) = t_2^*(H) = 0$, $t_1^*(H) = 1/2$, and $t_2^*(L) = 1$. The L state plays the role of a lemon: player 1 is always keen to reach an agreement in this state, but it is always bad for player 2.

If one of the players has unlimited liability then for any $M > 0$, the following mechanism is feasible and costs only $1/(M + 1)$. The player with unlimited liability is player 2, the uninformed player, and $p^*(H) = M/(1 + M)$, $t_1^*(L) = 0$, $t_1^*(H) = t_2^*(L) = 1$, and $t_2^*(H) = -M$.