Lemons by Design: Sowing secrets to curb corruption*

[Provisional draft]

Andrew Clausen,† Christopher Stapenhurst‡

April 4, 2023

Abstract

Ortner and Chassang (2018) show how a principal can make corruption harder by randomising the monitor's incentives. We develop their idea by randomising the agent's incentives too. We find that the cheapest way to deter corruption is to enter the agent and the monitor into a special lottery. First, both players observe a secret uniform draw (their lottery number). Then, if the monitor reports hard evidence, the player with the higher draw wins a monetary prize proportional to the difference between the two draws. This creates a two-sided adverse selection problem in the market for evidence suppression. If a player expects to win the lottery, then they want the monitor to report the evidence. Thus, a player leaves the market if they draw the highest possible number. But then a player with the second highest possible number reasons that they must have the highest number on the market, so they leave too. The market unravels with each round of reasoning, and collapses altogether. This mechanism is robust to monitoring mistakes, limited liability and mediated side contracting, and exemplifies worst case information.

*Acknowledgments: We would like to thank Alex Dickson, James Dunham, Marina Halac, Ed Hopkins, Tatiana Kornienko, Raghav Malhotra, John Moore, Alfonso Montes, Mariann Ollar, Christian Lippitsch, Rachel Scarfe, Ina Taneva, Rafael Veiel, Rakesh Vohra, Yaoyao Xu, and Gabriel Ziegler for helpful comments and suggestions.

†The University of Edinburgh. andrew.clausen@ed.ac.uk

‡Budapest University of Technology and Economics. c.stapenhurst@edu.bme.hu

1 Introduction

Corruption and the perceived threat of corruption both undermine the effectiveness of real world institutions[1]. For example, Duflo et al. (2013) study factory owners (agents) who find it cheaper to bribe pollution inspectors (monitors) to report compliance, than to actually be compliant with pollution regulations. One way to deter this kind of corruption is to outbid the agent by rewarding the inspector for providing hard evidence of non-compliance[2], but this can be expensive if cost of compliance is high. Recent research by Ortner and Chassang (2018); Baliga and Sjöström (1998) and von Negenborn and Pollrich (2020) show that corruption can be deterred more cheaply by random incentives for the monitor. in this paper, we find an even cheaper way to deter corruption by jointly designing random incentives for both the monitor and the agent. We call it the *two-sided lemons scheme*. It is the cheapest in a large class of mechanisms it is robust in following senses: it fully implements compliance by the agent; it accommodates imperfect monitoring; it respects limited liability constraints for the agent and the monitor; and it deters a large class of corrupt side contracts (including bilateral bribe contracts).

As well as deterring corruption cheaply and robustly, the two-sided lemons scheme has features that are of independent interest. Firstly, and as the name suggests, it engineers a two-sided adverse selection ("lemons") problem in the market for corruption. In the classic lemons story (Akerlof, 1970), only the seller has private information about the value of the good being traded. Unravelling occurs because of a feedback loop between falls in price and falls in the buyer's expected value of the good on the market. Specifically, sellers with the most valuable goods leave the market when prices fall, which causes the average value of goods on the market to fall, causing the price to fall further, and so on. In our two-sided lemons story, the buyer and the seller both receive private information about the value of the good. Unravelling occurs at each individual price level because of a feedback loop between falls in the buyer's expected value of the good on the market, and rises in the seller's expected value. Specifically, the first seller types to leave the market are those with the highest estimation of the value of the good ('high' sellers). If they leave the market, then the buyer's expected value of the good falls, so the buyer types with the lowest expected value of the good ('low' buyers) leave the market. This increases the seller's expected value of the good, and hence the next highest sellers leave. The cycle continues until all buyer and seller types leave the market, and the market collapses. The logic is also similar to the "winner's curse" in auction theory.

Secondly, the two-sided lemons problem offers a *worst case* distribution over payoffs and signals in the sense that it maximises the expected gains from trade subject to a "no trade" constraint. This answers an open problem about how much surplus can be destroyed by information asymmetry in a bilateral trade setting with a given distribution over values. Previously, Carroll (2016) found that when the price of trade is fixed, the an-

[1] Recent examples include Tacconi and Williams (2020); Bahoo et al. (2020); Gründler and Potrafke (2019); Isaksson and Kotsadam (2018).

[2] See, for example, Tirole (1986); Kofman and Lawarrée (1993); Laffont and Martimort (2002).

swer is always 'none', because it is possible construct a public information structure that delivers the same outcome as any arbitrary information structure. We find that when the traders are able to negotiate a price, the answer can be as high as 'all of it' for some value distributions. This is because the public distributions that Carroll constructs cannot deter trade at more than one price at a time, whereas asymmetric information can deter trade at all prices simultaneously. This insight is also pertinent to the recent literature on robustness, which seeks to determine, for a given payoff structure, the worst case equilibrium that can arise under arbitrary perturbations of the information structure (Kajii and Morris, 1997; Morris and Ui, 2005).

We also make two important methodological contributions in proving our main results. Firstly, to prove that the two-sided lemons scheme deters every voluntary, incentive compatible and budget balanced side contract, we establish that each player's expected transfer can be found by integrating the cumulative increments of the transfers paid by each type. Secondly, in proving that any arbitrary trade-deterring mechanism costs at least as much as ours, we show how to overcome a transfinite induction problem in order to generalise a result of Carroll (2016) to establish that unravelling affects a whole continuum of types.

The outline of the paper is as follows. The next section demonstrates the two-sided lemons scheme in the context of a simple example. Section 3 then describes our contribution to the literatures on corruption, adverse selection, and information design. Section 4 defines an auxiliary "trade problem" that captures the essence of what it means to deter bribes. In the trade problem, a designer chooses a joint distribution over private signals and transfers for a buyer and a seller, in order to implement a "no trade" outcome (deter trade), whilst minimising the expected sum of transfers to the buyer and the seller (the cost). Section 5 contains our main results: a definition of the two-sided lemons scheme; a theorem (Theorem 1) stating that the two-sided lemons scheme deters trade, and costs less than any other mechanism that deters trade; and a proof of the theorem containing our methodological contributions (Lemma 1 and Lemma 2). Section 6 applies our main result to a moral hazard problem with corruption, limited liability and imperfect monitoring. We show that it is weakly optimal to deter bribes (Lemma 4), and the two-sided lemons scheme is the optimal way to deter bribes (Corollary 1). Finally, we compare the two-sided lemons scheme with the cheapest one-sided mechanism (Proposition 1). Section 7 concludes by discussing practical implementation of the two-sided lemons scheme, and directions for future research.

2 Illustrative examples

There are three risk neutral players: the factory (he), the inspector (she), and the government (it). Suppose for the sake of illustration that the government has committed to fine the factory \$12k whenever the inspector reports a hard, evidence of non-compliance;

and that this \$12k fine is just enough to incentivise the factory to comply with the law.[3] The bribery problem arises because the factory is always better off bribing the inspector anything up to \$12k to hide the evidence. If the factory anticipates that he can avoid the fine by bribing the inspector, then he does not have the necessary incentives to comply.

Benchmark mechanism (transfers only) One solution is for the government to buy-out the inspector by transferring the factory's fine to her when she reports the evidence. Doing so ensures that the inspector will reject any bribes less than \$12k. It follows that the factory pays at least \$12k whenever the inspector obtains evidence, so he is better off complying. Thus, the benchmark solution works, but it costs the government \$12k. Moreover, every transfer-only mechanism costs at least \$12k.

Coin-toss mechanism (simple transfers and information) The government can deter bribes at a cost of only \$6k by using information design to engineer a simple (one sided) adverse selection problem.

Firstly, the government randomly determines the player's transfers by tossing a fair coin. One side of the coin yields a \$24k fine for the factory and a \$12k prize for the inspector. We call this the 'peach outcome' because the factory avoids a \$24k fine if he successfully bribes the inspector, so bribery is a 'peach'. The other side of the coin yields no fine for the factory and no reward for the inspector. We call this the 'lemon outcome' because the factory does not avoid any fine by bribing the inspector, so bribery is a 'lemon'. These transfers are shown in Table 1. If the inspector reports a low reading then the payoffs are the same as in the benchmark: the factory gets no fine and the inspector gets no reward.

	Peach	Lemon
Probability	0.5	0.5
Factory	-24	0
Inspector	12	0

Table 1: coin toss scheme payoffs (in \$k) when the inspector reports a high reading.

Secondly, the government designs private information by showing the outcome of the coin toss to the inspector, but not to the factory. This means that the inspector knows whether or not she will be rewarded for reporting a high reading, whereas the factory only knows that he will be fined \$24k with probability one half.

The timing is as follows. First, the government announces the scheme. Then the factory chooses whether or not to comply. Then government tosses the coin and shows the outcome to the inspector, and the inspector inspects the factory (the order is not important). If she obtains a low reading, then the government make no transfers, and the game ends. Otherwise, the inspector and the factory attempt to negotiate a bribe.

[3] The full model is outlined in Section 6.

If they are successful then the factory pays the inspector a bribe, the inspector reports a low reading, the government makes no transfers, and the game ends. Otherwise, the inspector reports the evidence, the government makes the transfers according to Table 1, and the game ends.

We claim that the coin toss scheme (i) deters bribes; (ii) incentivises compliance; and (iii) costs half as much as the benchmark. The second two points are straightforward: the inspector's average prize is $\frac{1}{2} \times \$12 + \frac{1}{2} \times \$0 = \$6$k, which is half as much the benchmark; and the factory's average fine is $\frac{1}{2} \times 24 + \frac{1}{2} \times \$0 = \$12$k, so (absent bribes) the coin toss scheme incentivises compliance. It only remains to show that the coin toss scheme deters bribes.

First we show that the coin toss scheme deters 'small' bribes, i.e. bribes strictly below \$12k. The inspector will not accept small bribes in the peach outcome because she knows that she stands to receive a \$12k prize if she reports the high reading. She will accept small bribes in the lemon outcome because she knows that she will not receive anything for reporting the high reading. What about the factory? He doesn't know whether or not he will be fined, but he does know that the inspector only accepts small bribes in the lemon outcome, and this is precisely the outcome where the factory does not benefit from bribes. Therefore the factory is better off not paying small bribes. The logic is the analogous that of Akerlof's (uninformed) used car buyer:[4] if the (informed) car seller is willing to accept a low price, then it must be because they know they are selling a bad car (or 'lemon'). This inference decreases the buyer's conditional willingness to pay, so that they do not trade even at a low price. Similarly, accepting the bribe is a 'lemon' for the inspector if the coin reveals the lemon side, so she is better off rejecting. Hence the coin toss scheme deters all bribes strictly less than \$12k.

The coin toss scheme also deters 'big' bribes, i.e. those strictly greater than \$12k but less than \$24k. The inspector is better off accepting big bribes in both the peach and the lemon outcomes. The factory is better off accepting big bribes in the peach state, but he is worse off if he accepts them in the lemon state. Unfortunately for him, he can't trust the inspector not to accept the bribe in the lemon outcome for the same reason that a used car buyer can't trust a seller not to sell him a lemon — the lemon seller always wants to sell, especially at high prices! So if the factory pays the bribe, then he has to accept that the outcome is equally likely to be a peach or a lemon. But then his expected fine is only \$12k, so he is better off rejecting big bribes after all.

The knife edge cases where the factory pays a bribe exactly equal to \$0 or \$12k can be ruled out by paying the inspector some arbitrarily small amount (e.g. \$1) in addition to her random prize. Finally, bribes greater than \$24k are clearly never going to be attractive to the factory because he never gets fined more than this anyway.

Thus, the coin toss scheme engineers an adverse selection problem that is severe enough to deter all bribes. It strictly improves on the benchmark because the government no longer has to completely outbid the factory. Although bribes create surplus for

[4]The key differences is that Akerlof's buyer faces uncertainty about the value of trading, whereas the factory faces uncertainty about the value of *not* trading.

the factory and the inspector in the peach outcome, they are deterred because (i) the inspector's prize is biggest in the peach outcome, so if she's willing to accept a given bribe in the peach outcome, then she is also willing to accept it in the lemon outcome; (ii) the inspector knows whether the outcome is a peach or a lemon, so she can select to accept small bribes only in the lemon outcome; (iii) the factory doesn't know the outcome, so he can't pay bribes in the peach outcome without also risking that he pays them in the lemon outcome. The numbers in the coin toss scheme are then chosen so that the smallest bribe that the inspector is willing to accept in the peach outcome is too big for the factory to risk wasting it in the lemon outcome. So no bribes take place in equilibrium. These same general properties will hold for both players in the two-sided lemons scheme.

The two-sided lemons scheme (optimal transfers and information) Two-sided private information creates scope for strategic contagion to amplify the effects of adverse selection.

In the two-sided lemons scheme, the government gives the factory a baseline fine of \$16k when the inspector reports hard evidence of non-compliance[5] Instead of tossing a coin, the government runs a lottery. The factory and the inspector each receive a random draw between 0 and 1, which constitutes their private information. Call them x_F and x_I respectively. Every number is equally likely (uniform). If the inspector reports the high reading then the player with the higher number receives a prize equal to 16 times the difference between the two numbers, divided by the larger number, i.e.

$$16\frac{|x_F - x_I|}{\max\{x_F, x_I\}}.$$

Finally, the inspector gets paid a fixed reward of \$8 to break indifferences (and simplify algebra). There is no prize when the inspector reports no evidence, and the timing is same as the coin toss scheme.

This scheme can be depicted in a unit square with the inspector's number on the horizontal axis, and the factory's number on the vertical axis (Figure 1). Each outcome of the lottery corresponds to a point in this square. The players' private information is embodied by the fact that the inspector knows only the horizontal coordinate of the true outcome, whereas the factory knows only the vertical coordinate. If the outcome is in the top left (bottom right) triangle, then the factory (the inspector) gets the prize. The size of the prize is depicted by the shading: in red outcomes, the prize is paid to the factory, in blue outcomes it is paid to the inspector. Darker shades (near the axes) indicate a higher prize, and lighter shades (near the diagonal and the top right corner) indicate a lower prize.

Every prize between 0 and \$16k is equally likely, so average prize is exactly \$8k. The inspector wins with probability 1/2 so her average payoff when she reports evidence is $\frac{1}{2}8 = 4$k. Similarly, the factory's average fine is $16 - \frac{1}{2}8 = 12$k, so he is incentivised to

[5] Any number greater than 12 works — we choose 16 only because it ensures that the cost is an integer. If chose 24 then the cost would be approximately \$2.1k.

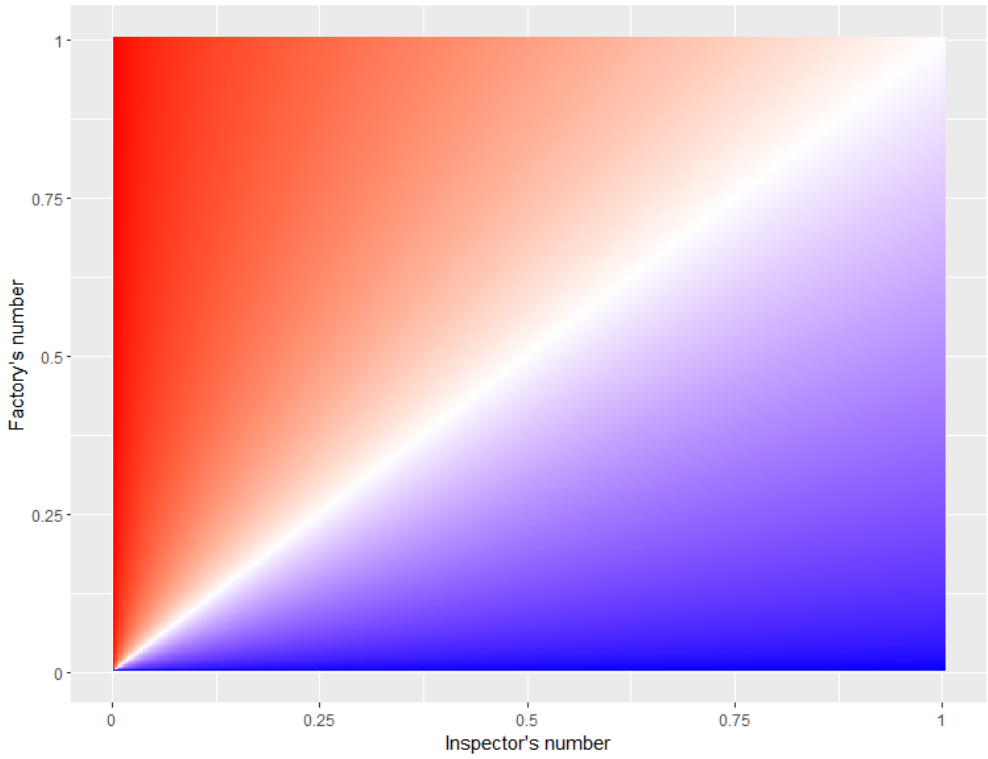


Figure 1: Darker red includes a higher prize for the factory; dark blue indicates a higher prize for the inspector.

comply. It only remains to show that the scheme deters bribes. We illustrate the main intuition by showing that it is always rational for both players to reject a \$8k bribe.[6]

First, suppose that inspector believes the factory will accept the \$8k bribe, no matter what number he receives. Given that the inspector does not know the factory's number, and that she only wins when her number x_I is highest, her expected reward is

$$\int_0^{x_I} 16\frac{x_I - x_F}{x_I}\, dx_F = 8x_I$$

thousand dollars, plus her \$8 fixed wage. So her best response must would be to reject the \$8k bribe if

$$\begin{aligned} 8000x_I + 8 &\geq 8000 \\ x_I &\geq 0.999, \end{aligned}$$

and to accept if her number is below the threshold 0.999.[7]

[6]There's nothing special about 8 except that it's halfway between 0 and 16, and therefore preserves symmetry.

[7]What she does when her number is exactly 0.999 is of no consequence because this occurs with 0 probability.

It is rational for the factory to accept the bribe if his payoff from accepting the bribe is bigger than the payoff he would get from rejecting the bribe. If the inspector rejects the bribe, then the factory's payoff is the same whether or not he accepts it. Therefore, just as in the lemons problem, the factory calculates his expected payoffs *conditional* on the inspector accepting the bribe. Conditional on the inspector accepting, the factory's payoff from accepting is -$8k (the cost of paying the bribe); his expected payoff from rejecting depends on his number x_F. If $x_F \leq 0.999$, then, given the inspector's threshold, his conditional expected payoff from rejecting is

$$\frac{1}{0.999}\left(\int_0^{x_F} 16\frac{x_F - x_I}{x_F} - 16\, dx_I + \int_{x_F}^{0.999} -16\, dx_I\right) = 8\frac{x_F}{0.999} - 16$$

thousand dollars. If $x_F \geq 0.999$, then it is

$$\frac{1}{0.999}\int_0^{0.999} 16\frac{x_F - x_I}{x_F} - 16\, dx_I = -8\frac{0.999}{x_F}$$

thousand dollars. These expressions are both increasing in x_F and, if $x_F = 0.999$, they both equal -8, which is exactly equal to the factory's payoff from accepting. It follows that the factory's best response to the inspector's threshold strategy is to accept bribes when his number is below 0.999, and reject bribes when his number is above 0.999. So his best response is to adopt the same threshold that the inspector uses.

However, the inspector's threshold was chosen as a best response to the belief that the factory accepts the bribe, no matter what number he receives. But if the inspector anticipates that the factory adopts a threshold strategy, then her conditional expected reward when she rejects will be the same as the factory's, except that she also receives the $8 fixed wage. So her best response to the factory's threshold is to accept the bribe whenever

$$\begin{aligned} 8000\frac{x_I}{0.999} + 8 &\leq 8000 \\ x_I &\leq 0.999^2 \end{aligned}$$

and to reject otherwise. Thus, she reduces her threshold to 0.999^2.

Similar reasoning shows that the factory's best response is to adopt the whatever threshold he believes the inspector is adopting, and that, the inspector's best response is to adopt 0.999 times whatever threshold she believes inspector is adopting. After n rounds of such reasoning, the players only accept the bribe only if their numbers are below 0.999^n. In the limit, 0.999^n converges to 0, so it is never rational for the them to accept the bribe. Or, to put it differently, neither player wants to agree to the bribe if they believe they have the highest number. But in equilibrium they can't both believe they have the highest number, so they can never agree the bribe. An analogous argument applies for any size of bribe, for any initial belief, and no matter how small the inspector's fixed wage is. We conclude the two-sided lemons scheme deters bribes.

The previous argument shows that the two-sided lemons scheme not only creates an adverse selection problem (like the coin toss scheme), but that this adverse selection problem is also contagious. Each player wants to accepts a bribe only when their expected value of the prize is low, which leads them to reject when they receive a high number. This creates a strategic externality, because each player's expected value is increasing in the other player's number, so their expected value falls if the other player chooses to reject when they receive high numbers. Therefore neither player ever accepts, even though there are almost always strictly positive gains from trade.

Our main result, Theorem 1, and Corollary 1 show that the two-sided lemons scheme not only deters bribes of the type described here, but it also *deters a more general class of bribe contracts* mediated by a mafia. They also show that two-sided lemons schemes, of the type described here, are the *cheapest way to to provide the correct incentives* for the firm.

3 Literature

We contribute to an extensive literature on corruption (Tirole, 1986; Laffont and Martimort, 1997; Strausz, 1997; Baliga and Sjöström, 1998). The closest paper to ours is Ortner and Chassang (2018). They are the first (to the best of our knowledge) to study the use of endogenous asymmetric information to deter bribes. They show that a principal (the government) can benefit from paying the monitor (inspector) a random wage (privately observed by the monitor) according to a public distribution, known to the agent (the factory). Doing so endows the monitor with private information about their outside option, and thereby creates an informational-friction in subsequent collusive negotiations between the monitor and a would-be criminal agent. In their model, the agent chooses between being criminal and bribing the monitor on the one hand, or being innocent on the other. Paying the monitor a random wage creates a trade-off for the agent: he can either offer a high bribe which guarantees a high probability of successfully corrupting the monitor, or he can offer a low bribe which guarantees a low probability of successfully corrupting the monitor. The principal saves money by paying random wages because low wage monitors can mimic the high wage monitors and demand high bribes.

Our model differs from theirs in two important respects. Firstly, Ortner and Chassang (2018) assume perfect monitoring so they can rule out bribes on the equilibrium path (when no incriminating evidence arises) without needing to rule them out off the equilibrium path (when the monitor receives incriminating evidence with certainty). By contrast, we allow for monitoring mistakes so we have to consider the impact of bribes both on and off the equilibrium path. If, like Ortner and Chassang, we pay rewards according to a distribution that pays rewards that are smaller than the agent's punishment, then we inevitably get on-path bribery because the agent is always weakly better off accepting bribes smaller than the punishment, and there will be a strictly positive probability that the monitor is willing to accept such bribes. This difficulty motivates our second main departure from their model, which is to endogenise the agent's fines. Doing so allows us to replicate the agent's trade-off in their model, because we can use the

changes in the agent's fine to imitate his choice to commit crime or not. This gives our result a qualitatively different interpretation from theirs: our agent faces a lemons problem because his fine depends on the monitor's private information. Despite these differences, our informed inspector scheme (Proposition 1), which is the closest to theirs conceptually, produces the same distribution of rewards and has the same cost.[8] We showed in Section 5 that the two-sided lemons scheme costs strictly less than the informed inspector scheme, and costs half as much in the limit as the size of the maximal punishment increases.

Our results speak to the literature on the robustness of equilibria to contagion.[9] Carroll (2016) obtains an upper bound on the amount of surplus lost due to contagion in a game with two agents either accepting or rejecting a proposed deal, where both agents have private information about the payoff outcomes of the deal. Surprisingly, Carroll finds that contagion does not prevent the agents from realising joint surplus, so long as they have common knowledge that their ex-post joint surplus from the deal is weakly positive. He concludes by asking "What change[s] if we consider ... mechanism[s] that determine not only whether a deal takes place but which deal is chosen? ... Is it possible to describe the worst-case information structure?" (pp. 355–356). Our two-sided lemons scheme entails common knowledge that the ex-post joint surplus from bribery is weakly positive, and yet we find that contagion does play an important role in this scheme. We conclude that contagion does become problematic when the players are trying to negotiate the terms of a deal, because the players' types adversely select which terms to accept. The two-sided lemons scheme has a worst case information structure which leads to all deals being rejected for a particular distribution of payoffs (payoffs are endogenous in our setting, but exogenous in his). This worst case information structure has independent and uniform signals that quantify the severity of the lemons problem faced by the recipient.

Our problem fits into a larger class of general mechanism design problems in which the designer chooses both transfers and information.[10] A particularly relevant and recent paper is Halac et al. (2021)'s *Ranking Uncertainty in Organisations*. They show how 'ranking schemes' can create strategic uncertainty and thereby induce a team of workers to exert complementary effort on a project. Ranking schemes are superficially similar to ours in two respects. Firstly, all the players receive a private message. Secondly, the distribution of payoffs is chosen so that work is a dominant strategy for the players with the highest possible message realisation, and each player finds it optimal to work conditional on the belief that all players with the same message or higher will work. Thus, like ours, their scheme produces an inductive chain that causes working to be a higher order best response for all other workers. However, the mechanism underlying their ranking scheme is qualitatively different from ours. Endogenous private information benefits their de-

[8] Garrett et al. (2021) obtain the same distribution as a solution to a similar problem in which an agent chooses their distribution of costs to maximise their information rent.

[9] See e.g. Kajii and Morris (1997); Morris and Ui (2005).

[10] See Bergemann and Morris (2019); Mathevet et al. (2020); Taneva (2019).

signer because each worker's incentives to work are strictly concave in their belief that other workers will work. Therefore, a given incentive is created more cheaply by randomising over beliefs. There is no lemons problem in their scheme because workers have complete information about their own payoffs — other workers' types only affect them indirectly through the other workers' decisions to exert effort. By contrast, asymmetric information only benefits us because it inhibits our players from negotiating bribes (which are not considered in Halac et al. (2021)). We engineer a lemons problem by designing a scheme in which each player's payoffs depend directly on the message received by the other player.

Morris and Shin (2012)

4 Model

This section presents a general model of deterring trade. We have already introduced the designer, and the traders — the buyer B and the seller S. To ensure that our proposed scheme deters all possible bribes, we follow Laffont and Martimort (1997) and introduce a fourth player, the mafia. The mafia helps the traders negotiate bribes by proposing and enforcing side contracts. The seller values the good at 0 and the buyer values it at κ.

The designer moves first by proposing a scheme to the buyer and seller.

Definition 1. A *scheme* $\mathcal{S}$ consists of (X, Y, Σ, p, t).

- X is a set of messages that the designer can privately send to the buyer.
- Y is a set of messages that the designer can privately send to the seller.
- Σ is a σ-algebra on $X \times Y$.
- $p : \Sigma \to [0,1]$ is a probability measure over $X \times Y$, so $p(E)$ is the probability of sending a pair of messages from a measurable set $E \in \Sigma$.
- $t_B, t_S : X \times Y \to \mathbb{R}_+$ specify the transfers from the designer to the buyer and the seller respectively, if they don't trade.
- If the traders do trade, then the designer pays them both zero.

We require a scheme to satisfy the following technical properties:

- The transfers t_B and t_S are Σ-measurable, i.e. pre-images of Borel sets are Σ-measurable.
- There exist marginal and conditional probability measures.

We implicitly assume that transfers are a deterministic function of the messages (x, y). This is without loss of generality, because the players are risk neutral.

If the traders agree to the scheme, the mafia proposes a side contract. We assume that the mafia 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 schemes are robust to a wider range of side contracts. Second, it means that the revelation principle applies, which means that for any equilibrium that an arbitrary side contract might implement, the mafia can adopt an equivalent direct side contract that partially implements truth-telling. Hence there is no loss of generality in restricting attention to direct side contracts.

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

- a Σ-measurable allocation rule $a : X \times Y \to [0, 1]$, which specifies the probability of trade for pair each pair of messages, (x, y) reported by the traders; and
- a pair of Σ-measurable bribe functions $b_B, b_S : X \times Y \to \mathbb{R}$ that specifies the (possibly negative) bribe that each trade pays to the mafia for each pair of messages (x, y).

The buyer's expected value from telling the mafia x' when his true message is x is

$$\int_Y [a(x', y) + (1 - a(x', y))t_B(x, y) - b_B(x', y)]\, dp_{S|B}(y|x). \tag{1}$$

If the buyer declines to participate, he receives the expected transfer

$$\int_Y t_B(x, y)\, dp_{S|B}(y|x). \tag{2}$$

It will be convenient to normalise the outside option to zero by studying the buyer's net value of reporting x' when his true message is x, i.e.

$$V_B(x, x') := \int_Y [a(x', y)(1 - t_B(x, y)) - b_B(x', y)]\ dp_{S|B}(y|x). \tag{3}$$

Similarly, the seller's net value from telling the mafia y' when his true message is y is

$$V_S(y, y') := \int_X [-a(x, y')t_S(x, y) - b_S(x, y')]\ dp_{B|S}(x|y). \tag{4}$$

These integrals exist because the integrands are non-negative and the function a is Σ-measurable. Their values of participating in the side contract are

$$W_B(x) := \max_{x'} V_B(x, x') \tag{5}$$

$$W_S(y) := \max_{y'} V_S(y, y'). \tag{6}$$

Definition 3. A side contract is *feasible* if it satisfies the following constraints:

1. (Side incentives) Each trader prefers to truthfully report their private message to the mafia, i.e. for all $x \in X$ and for all $y \in Y$

$$W_B(x) = V_B(x, x) \tag{$\mathrm{SI_B}$}$$

$$W_S(y) = V_S(y, y). \tag{$\mathrm{SI_S}$}$$

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

$$W_B(x) \geq 0 \quad \forall x \in X \tag{SP$_B$}$$
$$W_S(y) \geq 0 \quad \forall y \in Y. \tag{SP$_S$}$$

3. (Profitable) The mafia makes a strictly positive profit in expectation, i.e.

$$\int [b_S(x,y) + b_B(x,y)]\, dp(x,y) > 0. \tag{P}$$

The reason for focusing on strictly profitable side contracts is purely technical — it ensures that an optimal scheme exists. Without it, we would have to add ε to our schemes to destroy any unwanted equilibria.

The designer wants to deter these side contracts.

Definition 4. A scheme $\mathcal{S}$ *blocks trade* if there is no feasible side contract.

For any given weight $\lambda \geq 0$, the *weighted cost* of a scheme $\mathcal{S}$ is given by the weighted expected value of the transfers, $c(\mathcal{S};\lambda) := \int_{X\times Y} [\lambda t_B(x,y) + t_S(x,y)]\, dp(x,y)$. We use $\lambda \neq 1$ when we study moral hazard in Section 6.

Definition 5. The *trade problem* is to find the scheme that blocks trade at the lowest weighted cost, i.e.

$$\min_{\mathcal{S}} c(\mathcal{S};\lambda) \text{ s.t. } \mathcal{S} \text{ blocks trade.} \tag{7}$$

5 Blocking trade

Definition 6. We propose the *two-sided lemons scheme* $\mathcal{S}^*(\lambda) = (X^*, Y^*, \Sigma^*, p^*, t^*)$ as follows.

- The designer draws the messages (x,y) uniformly from $[0,1]^2$. Thus, $X^* = Y^* = [0,1]$, Σ^* is the Lebesgue σ-algebra on $X^* \times Y^*$, and p^* is the Lebesgue measure.
- If the buyer and seller do not trade, then the designer rewards them

$$t_B^*(x,y) = \frac{\kappa}{x} \max\left\{0, x - y^{1/\lambda}\right\} \tag{8}$$
$$t_S^*(x,y) = \frac{\kappa}{y} \max\left\{0, y - x^{\lambda}\right\}. \tag{9}$$

Theorem 1. *The two-sided lemons scheme $\mathcal{S}^*(\lambda)$ solves the trade problem at a weighted cost of $c(\mathcal{S}^*;\lambda) = \kappa\frac{\lambda}{\lambda+1}$. The buyer and seller receive expected transfers of $\kappa\frac{1}{(\lambda+1)^2}$ and $\kappa\frac{\lambda^2}{(\lambda+1)^2}$.*

Proof. Without loss of generality, assume that $\kappa = 1$. It is straight forward to calculate the expected transfers of the two-sided lemons scheme $\mathcal{S}^*$, and to check that the weighted cost is $\frac{\lambda}{\lambda+1}$.[11] The proof has two parts. First we show that $\mathcal{S}^*$ blocks trade (feasibility). Then we show that every scheme that blocks trade costs at least $\frac{\lambda}{\lambda+1}$ (optimality).

Feasibility Pick any measurable side allocation rule $a : X \times Y \to [0,1]$. We adapt Myerson (1981)'s logic to calculate the traders' surplus under this allocation rule. We will prove that the rule gives all the gains from trade to the traders, and leave no profits for the mafia.

The buyer's value function, $W_B(x)$, is convex and therefore absolutely continuous.[12] The Fundamental Theorem of Calculus (Royden and Fitzpatrick, 1988, Theorem 10) implies that the value of type x can be calculated by summing the marginal values,

$$W_B(x) = W_B(1) - \int_x^1 \frac{d}{ds} W_B(s)\, ds, \tag{10}$$

using type $x = 1$ as our reference point. These marginal values can be calculated with the envelope theorem,[13]

$$\frac{d}{ds} W_B(s) = \frac{\partial}{\partial s} V_B(s,s')\Big|_{s'=s} = -\int_0^{s^\lambda} a(s,y) \frac{y^{1/\lambda}}{s^2}\, dy = -\frac{1}{s}\int_0^{s^\lambda} a(s,y)[1 - t_B^*(s,y)]\, dy. \tag{11}$$

Substituting (11) and the side participation constraint $W_B(1) \geq 0$ into (10), we find that the buyer's expected surplus is at least

$$\int_0^1 W_B(x)\, dx \geq \int_0^1 \int_x^1 \frac{1}{s} \int_0^{s^\lambda} a(s,y)[1 - t_B^*(s,y)]\, dy\, ds\, dx. \tag{12}$$

We use the following lemma to simplify the triple integral.

Lemma 1. *If $f : [0,1] \to \mathbb{R}$ is Lebesgue integrable, then $\int_0^1 \int_x^1 f(s)\, ds\, dx = \int_0^1 x f(x)\, dx$.*

Proof. Let $g(x) = \int_x^1 f(s)\, ds$. Notice that $g(1) = 0$ and that, by Leibniz' rule, $g'(x) = -f(x)$. Since g is absolutely continuous (it is an indefinite integral) and 1 is Lebesgue integrable, integration by parts implies

$$\int_0^1 g(x) \times 1\, dx = g(1) \times 1 - g(0) \times 0 - \int_0^1 g'(x) x\, dx = -\int_0^1 -f(x) x\, dx = \int_0^1 x f(x)\, dx. \qquad \square$$

[11] Specifically $\int_0^1 t_B^*(x,y)\, dy = \int_0^{x^\lambda} \frac{x - y^{1/\lambda}}{x}\, dy = x^\lambda - \frac{1}{x}\frac{\lambda}{\lambda+1} x^{\lambda+1} = \frac{1}{\lambda+1} x^\lambda$ so that $\int_{[0,1]^2} t_B^*(x,y)\, d(x,y) = \int_0^1 \frac{1}{\lambda+1} x^\lambda\, dy = \frac{1}{(\lambda+1)^2}$. Similarly, $\int_0^1 t_S^*(x,y)\, dx = \frac{\lambda}{\lambda+1} y^{1/\lambda}$ and $\int_{[0,1]^2} t_S^*(x,y)\, d(x,y) = \frac{\lambda^2}{(\lambda+1)^2}$.

[12] Notice that $V_B(x,x')$ is convex in x. So W_B is an upper envelope of convex functions so it is also convex. Thus W_B is absolutely continuous by (Royden and Fitzpatrick, 1988, Corollary 17).

[13] Specifically, $V_B(x,x') = \int_0^1 a(x',y)\, dy + \int_0^{x^\lambda} a(x',y) \frac{y^{1/\lambda}}{x}\, dy - \int_0^1 b_B(x',y)\, dy$ so the envelope theorem implies $\frac{\partial}{\partial x} V_B(x,x') = a(x',y) \frac{y^{1/\lambda}}{x}\Big|_{y=x^\lambda} \frac{dx^{1/\lambda}}{dx} - \int_0^{x^\lambda} a(x',y) \frac{y^{1/\lambda}}{x^2}\, dy = -\int_0^{x^\lambda} a(x',y) \frac{y^{1/\lambda}}{x^2}\, dy$.

This lemma has the following intuition. Suppose that the agent's type x is drawn uniformly from $[0,1]$. If an agent's first-order condition is that the marginal transfer $t'(x) = f(x)$, then each side of the equation gives a way to calculate the expected transfers. On the left side, the inner integral calculates $t(x)$, and the outer integral calculates the expectation. On the right side, $f(x)$ is added to all transfers from $t(0)$ to $t(x)$, so $xf(x)$ represents the cumulative increment. The integral sums these increments.

Applying Lemma 1, the bound on the buyer's expected surplus (12) becomes

$$\int_0^1 W_B(x)\,dx \geq \int_{x>y^{1/\lambda}} a(x,y)(1-t_B^*(x,y))\,d(x,y). \tag{13}$$

The calculations for the seller's expected surplus are similar, and give

$$\int_0^1 W_S(y)\,dy \geq \int_{x<y^{1/\lambda}} a(x,y)(1-t_S^*(x,y))\,d(x,y). \tag{14}$$

Summing the two inequalities, we deduce that the traders' joint surplus is at least as big as the gains from trade, i.e.

$$\int_0^1 W_B(x)\,dx + \int_0^1 W_S(y)\,dy \geq \int a(x,y)(1-t_B^*(x,y)-t_S^*(x,y))\,d(x,y). \tag{15}$$

Since there is no surplus left for the mafia, the two-sided lemons scheme blocks trade.

Optimality We now prove that the two-sided lemons scheme $\mathcal{S}^*(\lambda) = (X^*, Y^*, \Sigma^*, p^*, t^*)$ has the lowest possible weighted cost, i.e. every scheme $\mathcal{S}$ that blocks trade costs at least $\frac{\lambda}{\lambda+1}$. We proceed in three steps. We follow the literature by referring to x and y as the traders' types. First we define a special type of side contract, called a *simple bribe*. Next, we define what it means for a scheme to 'infect' a player-type in the resulting game (in the spirit of Akerlof/Rubinstein-style unravelling), and show that any scheme that blocks all simple bribes necessarily infects all the types of at least one player. Finally, we show that any scheme that infects all the types of either player necessarily costs at least $c(\mathcal{S}^*, \lambda)$. Hence any scheme that blocks all side contracts costs at least $c(\mathcal{S}^*, \lambda)$.

Let $\mathcal{S} = (X, Y, \Sigma, p, t)$ be a any scheme.

Definition 7. A *simple bribe* is a side contract of the form

$$a(x,y) = a_B(x)a_S(y) \tag{16}$$
$$b_B(x,y) = a(x,y)b \tag{17}$$
$$b_S(x,y) = -a(x,y)b \tag{18}$$

where $a_B : X \to [0,1]$ and $a_S : Y \to [0,1]$ are measurable acceptance policies for the buyer and the seller respectively, and $b > 0$ is a constant price paid by the buyer to the seller.

The traders do not need the mafia's help to execute a simple bribe. The buyer would offer a bribe of b directly to the seller if $a_B(x) = 1$, and the seller would accept this if $a_S(y) = 1$.

If a buyer type x best responds to a seller strategy a_S by rejecting the bribe, then we say that x is 'infected by' a_S.

Definition 8. A *finite outbreak* of length N is a pair of finite sequences of infected types $(X_n, Y_n)_{n=0}^N$ where $X_n \subseteq X$ and $Y_n \subseteq Y$ that satisfies the following properties:

1. At the start, no types are infected, i.e. $X_0 = Y_0 = \emptyset$.
2. Once infected, types stay infected, i.e. $X_n \subseteq X_{n+1}$ and $Y_n \subseteq Y_{n+1}$ for all $n < N$.
3. In round n, at most one player has a non-empty set of *newly infected types*, i.e. either $\bar{X}_n := X_n \setminus X_{n-1} \neq \emptyset$ or $\bar{Y}_n := X_n \setminus X_{n-1} \neq \emptyset$, but not both.
4. Newly infected buyer types get infected by a 'adverse seller strategy' and newly infected seller types get infected by a 'adverse buyer strategy', i.e. if $\bar{X}_n \neq \emptyset$ then there exists a strategy a_S such that
   (a) uninfected seller types accept, i.e. $a_S(y) = 1$ for all $y \notin Y_n$;
   (b) each newly infected type is better off rejecting the deal, i.e.

$$\int_Y [t_B(x,y) - (1-b)] a_S(y)\, dp(y|x) > 0 \qquad \forall x \in \bar{X}_n. \tag{19}$$

   Similarly, if $\bar{Y}_n \neq \emptyset$ then there exists a strategy a_B such that $a_B(x) = 1$ for all $x \notin X_n$, and

$$\int_X [t_S(x,y) - b] a_B(x)\, dp(x|y) > 0 \qquad \forall y \in \bar{Y}_n. \tag{20}$$

An *infinite outbreak* $(X_n, Y_n)_{n\in\mathbb{N}}$ is defined similarly. The *size* of an outbreak is equal to the measure of infected buyer types, $p_B(\cup_n X_n)$ (the argument would be completely analogous if we defined it to be the measure of infected seller types).

Lemma 2. *There exists an infinite outbreak $(X''_n, Y''_n)_{n\in\mathbb{N}}$ that infects all of the buyer's types, i.e. $p_B(X_n) \to 1$.*

Proof of Lemma 2. We would like to prove that every buyer type gets infected. If there were a finite number of types, then we could use Carroll (2016, Propositions 3.1 and 3.2)'s logic. He shows that any outbreak that falls short of infecting all types can be extended to infect more types. By induction, we can show that, one by one, all types get infected. However, this argument does not directly generalise to infinite types. Infecting an infinite number of types one by one would lead to a transfinite induction problem.

Our approach has two differences. Instead of infecting types one-by-one, we infect a positive measure of types at each step. And we show that if a type can be infected

after a countable number of rounds of infection, then finite rounds would also suffice. To these ends, Claim 1 proves that there exists an infinite outbreak that is the same size as the supremum of the size of all the finite outbreaks. Claim 2 proves that any infinite outbreak that fails to infect all of the buyer types is smaller than some finite outbreak. It follows that the supremum of the size of all the finite outbreaks is 1 — otherwise we could first apply Claim 1 to obtain an infinite outbreak that attains the supremum, and then apply Claim 2 to obtain a larger finite outbreak that exceeds the supremum, giving a contradiction. Claim 1 then implies that there exists an infinite outbreak of size 1, which is precisely what it means to infect almost all buyer types. All that remains is to formally state and prove Claim 1 and Claim 2.

□

Claim 1. *Let O denote the set of finite outbreaks. Let*

$$r^* = \sup_{(X_n, Y_n)_{n=0}^N \in O} p_B(X_N)$$

denote the supremum of the size of all the finite outbreaks. There exists an infinite outbreak of size r^*.

Proof. Since r^* is the supremum, there exists a sequence of finite outbreaks $(X^m, Y^m)_{m\in\mathbb{N}}$, each of length N^m, such that $r^m = p_B(X^m_{N^m})$ converges to r^*. Let X'_n be the concatenation of these sequences, i.e. $X' = X^0 \,\|\, X^1 \,\|\, \cdots$. Let X''_n be the sequence $X''_0 = X'_0$ and $X''_{n+1} = X'_{n+1} \cup X''_n$. Construct the sequences Y'_n and Y''_n in the same way. The sequence $(X''_n, Y''_n)_{n\in\mathbb{N}}$ is an infinite outbreak because:

1. $X''_n = X'_0 = X^0_0 = \emptyset$ and $Y''_n = Y'_0 = Y^0_0 = \emptyset$.
2. $X''_{n+1} = X'_{n+1} \cup X''_n \supseteq X''_n$. Similarly, $Y''_{n+1} \supseteq Y''_n$.
3. We will prove that if $X''_{n+1} \neq X''_n$ then $Y''_{n+1} = Y''_n$. The reverse case when $Y''_{n+1} \neq Y''_n$ is similar. Suppose $X''_{n+1} \neq X''_n$. If these straddle a boundary between two finite outbreaks, $X^m \,\|\, X^{m+1}$, then $Y'_{n+1} = \emptyset$ so $Y''_{n+1} = Y''_n \cup \emptyset = Y''_n$, as required. Otherwise, $X''_{n+1} \neq X''_n$ lie within a single outbreak X^m. In this case, $X'_{n+1} \neq X'_n$ and hence $Y'_{n+1} = Y'_n$. We conclude that $Y''_{n+1} = Y''_n \cup Y'_{n+1} = Y''_n \cup Y'_n = Y''_n$ as required.
4. Suppose a non-empty set of buyer types $\bar{X}''_n$ are newly infected in step n. Let a_S be the adverse seller strategy from the underlying finite outbreak X^m_k.
   (a) We must check that the adverse seller strategy accepts on the uninfected types $Y \setminus Y''_n$. This follows from the fact that there are fewer uninfected types, i.e. $Y^m_k = Y'_n \subseteq Y''_n$ so $Y \setminus Y''_n \subseteq Y \setminus Y^m_k$.
   (b) We must check that the newly infected buyer types prefer to reject the deal. This follows from the fact that there are fewer newly infected types, i.e. $\bar{X}''_n \subseteq \bar{X}^m_k$.

   The logic for newly infected seller types is analogous.

Finally, every infection set at the end of each outbreak, $X^m_{N^m}$, is contained as a subset of some set X''_n, it follows that $p_B(X''_n) \to r^*$. □

Claim 2. *Let $(X_n, Y_n)_{n\in\mathbb{N}}$ be an infinite outbreak of size $r := \lim_{n\to\infty} p_B(X_n)$ that falls short of infecting all of the buyer types, so that $r < 1$. There exists a finite outbreak $(X'_n, Y'_n)_{n\leq N}$ with size $p_B(X'_N) > r$.*

Proof. Let $(X^*, Y^*) = (\cup_{n\in\mathbb{N}} X_n, \cup_{n\in\mathbb{N}} Y_n)$ denote the set of types infected by the outbreak, so that $p_B(X^*) = r$. Consider the amended game in which the uninfected types $(X\backslash X^*, Y\backslash Y^*)$ are constrained to accept trade, but the remaining types may choose to either accept or reject the trade. [14] The assumptions above ensure that of Balder's equilibrium existence theorem (Balder, 1988, Theorem 3.1) apply,[15] so an equilibrium (a^*_B, a^*_S) exists in the constrained game. There are no profitable deviations within this restricted strategy space. But trade between uninfected types occurs, so (a^*_B, a^*_S) is not an equilibrium in the unconstrained game. Therefore, the buyer (without loss of generality) must have a profitable deviation by a strictly positive measure of uninfected types $\bar{X} \subseteq X \setminus X^*$. These types get infected in the unconstrained game, so they would rather reject trade, i.e. for all $x \in \bar{X}$,

$$\int_Y [T_B(x) - (1-b)] a^*_S(y) dp(y|x) > 0. \tag{21}$$

Summing up, we deduce

$$\int_{\bar{X}\times Y} [T_B(x) - (1-b)] a^*_S(y) dp(x,y) > 0. \tag{22}$$

Since uninfected seller types $y \in Y\backslash Y^*$ accept, we can write

$$\int_{\bar{X}\times Y^*} a^*_S(y)[T_B(x) - (1-b)] dp(x,y) + \int_{\bar{X}\times(Y\backslash Y^*)} [T_B(x) - (1-b)] dp(x,y) > 0. \tag{23}$$

Now, for each period $N \in \mathbb{N}$ in which buyer types get infected (not seller types), consider the finite sequence $(X'_n, Y'_n)_{n\leq N}$ which equals (X_n, Y_n) up until period $N-1$, and equals $(X_{N-1} \cup \bar{X}, Y_N)$ in period N. In period N, the seller plays a strategy a^N_S which is equal to 1 (trade) on the set $Y\backslash Y_N \supseteq Y\backslash Y^*$ (so it satisfies the constrainr in round N), and which is equal to a^*_S on the set Y_N. The payoff for buyer types in $\bar{X}$ under a^N_S differs from their payoff under a^*_S by

$$\int_{Y^*\backslash Y_N} |1 - a^*_S(y)| [T_B(x) - (1-b)] \, dp(y|x), \tag{24}$$

[14] We owe this proof technique to Carroll (2016)'s study of the finite case.

[15] Balder's equilibrium existence theorem requires that the player's utility functions are (C1') measurable with respect to types and actions; (C1") conutinuous with respect to actions, for each type; (C1") bounded; and that the prior distribution over types is (C2) measurable with respect to it's marginals.

which is at most $\mathbb{P}[Y^* \setminus Y_N] \times [T_B(x) - (1-b)] \xrightarrow{N\to\infty} 0$. Types in $\bar{X}$ get a strictly prefer to reject trade under a_S^*, so there must exists some N large enough that they strictly prefer to reject trade under a_S^N as well. This means that types in $\bar{X}$ can get infected in round N, so $(X'_n, Y'_n)_{n\leq N}$ is a well defined finite outbreak.

Moreover, X_{N-1} and $\bar{X}$ are disjoint, so the size of $(X'_n, Y'_n)_{n\leq N}$ is

$$p_B(X_{N-1} \cup \bar{X}) = p_B(X_{N-1}) + p_B(\bar{X}) \to r + p(\bar{X}),$$

which is strictly greater than r. Thus, there exists a large enough N that defines a finite outbreak $(X'_n, Y'_n)_{n\leq N}$ with size greater than r. □

Lemma 3. *If $\mathcal{S}$ deters a bribe of b, then its weighted cost is at least* $\min\{\lambda(1-b), b\}$. *Hence, if $\mathcal{S}$ deters a bribe of $b^* = \frac{\lambda}{\lambda+1}$, its weighted cost is at least $\frac{\lambda}{\lambda+1}$.*

Proof. By Lemma 2, there is an outbreak $(X_n, Y_n)_{n\in\mathbb{N}}$ that infects almost all buyer types. At each round n of the outbreak, either some buyer or seller types are infected. Suppose the buyer types $\bar{X}_{n+1}$ are infected in round n. When faced with an adverse seller strategy a_S^n, they reject the bribe in favour of the transfers, i.e.

$$\int_{\bar{X}_{n+1}\times Y} a_S^n(y)[t_B(x,y) - (1-b)]dp(x,y) > 0. \tag{25}$$

Suppose that all of the infected seller types reject the bribe, leaving only the uninfected seller types $Y\setminus Y_n$, who must accept. We deduce that the newly infected buyers also reject these bribes, and with more vigor, i.e.

$$\int_{\bar{X}_{n+1}\times Y} t_B \, dp > (1-b) \int_{\bar{X}_{n+1}\times Y} a_S^n(y) \, dp(x,y) \geq (1-b)p(\bar{X}_{n+1} \times (Y\setminus Y_n)). \tag{26}$$

Similarly, in a round n where seller types $\bar{Y}_{n+1}$ are infected, they reject bribes when only the uninfected buyers $X\setminus X_n$ accept, i.e.

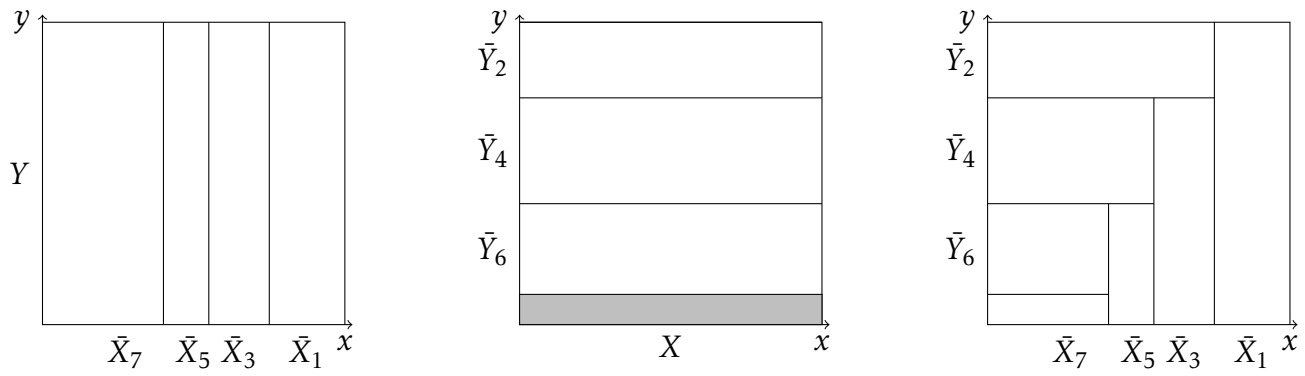
$$\int_{X\times\bar{Y}_{n+1}} t_S \, dp \geq bp((X\setminus X_n) \times \bar{Y}_{n+1}). \tag{27}$$

These inequalities involve three ways of splitting the type space up, depicted in Figure 2:

(a) Partitioning by buyer infections in each round n, i.e. $\{\bar{X}_n \times Y\}$.

(b) Splitting by seller infections in each round n, i.e. $\{X \times \bar{Y}_n\}$. Note that a positive measure are never infected, so these cells do not cover the type space.

(c) Partitioning by infections matched with uninfected types in each round n, i.e. $\{\bar{X}_{n+1} \times (Y\setminus Y_n)\}_{n\in\mathbb{N}} \cup \{(X\setminus X_n) \times \bar{Y}_{n+1}\}_{n\in\mathbb{N}}$.



(a) Buyer infections (b) Seller infections (c) Matched infections

Figure 2: Splitting up the type space. The shaded region is not covered.

The left sides of (26) and (27) sum up the transfers over all new infections. The right sides sum up the foregone bribes in the matched infections. Thus, taking the λ-weighted sum of the inequalities and summing up over the cells gives

$$\int_{X\times Y} (\lambda t_B + t_S)\, dp \geq \sum_{n=1}^{\infty} \int_{\bar{X}_{n+1}\times Y} \lambda t_B\, dp + \sum_{n=1}^{\infty} \int_{X\times \bar{Y}_{n+1}} t_S\, dp \tag{28}$$

$$\geq \lambda(1-b) \sum_{n=1}^{\infty} p(\bar{X}_{n+1} \times (Y \setminus Y_n)) + b \sum_{n=1}^{\infty} p((X\setminus X_n) \times \bar{Y}_{n+1}) \tag{29}$$

$$\geq \min\{\lambda(1-b), b\}. \tag{30}$$

□

□

6 Deterring Bribes

A factory chooses whether to pollute, p, or comply c. Other things equal, polluting yields a payoff of 1, whereas complying yields a payoff of 0. An inspector inspects the factory. If the factory chooses to pollute then the inspector obtains evidence of pollution with probability π_p; otherwise, she obtains no evidence. If the factory chooses to comply then the inspector obtains evidence of pollution with probability $\pi_c < \pi_p$. The fact that π_p can be strictly less than 1 implies that the monitoring technology is imperfect — the inspector may fail to obtain evidence of pollution even though the factory has been polluting. Similarly, if π_c is strictly greater than 0 then the inspector may find evidence that the factory has been polluting, even though he has been compliant. The requirement that evidence is more likely to arise when the factory pollutes than when he complies ensures that evidence is indicative of pollution. If the inspector obtains evidence of pollution then she

can either report it to the government, or she can keep silent. We assume that evidence cannot be fabricated, so if the inspector does not obtain evidence then she has no choice but to keep silent.

The government wants to incentivise the factory to comply, which it does so by designing a scheme $\mathcal{S} = ((X, \Sigma, p), (f, s, r))$, where (X, Σ, p) is a public distribution p over a message space $X = (X_I, X_F)$ with σ-algebra Σ (the factory observes the message x_F and the inspector observes the message x_I); and f, s and r are message-contingent transfers. Specifically, $f : X \to \mathbb{R}$ is a fine paid by the factory to the government and $r : X \to \mathbb{R}$ is a reward paid by the government to the inspector in the event that the inspector reports evidence; and $s : X \to \mathbb{R}$ is a subsidy paid by the government to the factory in the event that the inspector does not report evidence.

There is also a mafia who can offer feasible side contracts of the form described in Definition 2. Here, the factory takes on the role of the buyer and the inspector takes on the role of the seller. The function a is the probability that the inspector conceals the evidence and the transfers are the bribes (which can be negative). If the inspector conceals the evidence then the factory receives the subsidy $s(x)$ instead of paying the fine $f(x)$, hence his "net value of trade" is equal to $s(x) - (-f(x)) = s(x) + f(x)$ (for comparison, the buyer's net value of trade in Section 5 is $\kappa - t_B(x)$). The mafia doesn't care about compliance but rather just wants to maximise its profit from facilitating bribes. [16]

The timing is as follows: the government and the mafia commit to their chosen scheme and side contract respectively; the factory chooses its action; the government sends private messages; the factory and the inspector send reports to the mafia; nature selects whether or not the inspector receives evidence; if the inspector does receive evidence, then the mafia's side contract specifies what bribes are paid, and whether or not the inspector conceals the evidence; finally, the transfers specified by the government's scheme are made.

To be effective, the government's scheme must respect three types of constraint. Firstly, it must ensure that the factory s better off complying than polluting. If complies then he gets $\pi_c \mathbb{E}_x[a(x)s(x) + (1 - a(x))f(x)] + (1 - \pi_c)\mathbb{E}_x[s(x)]$, whereas polluting yields a payoff of $1 + \pi_p \mathbb{E}_x[a(x)s(x) + (1 - a(x))f(x)] + (1 - \pi_p)\mathbb{E}_x[s(x)]$. Thus to be *compliance incentive* compatible(CI), the government's scheme must satisfy

$$\begin{aligned} &\pi_c \mathbb{E}_x[a(x)s(x) - (1 - a(x))f(x) - t_F(x)] + (1 - \pi_c)\mathbb{E}_x[s(x)] \\ &\qquad \geq 1 + \pi_p \mathbb{E}_x[a(x)s(x) - (1 - a(x))f(x) - t_F(x)] + (1 - \pi_p)\mathbb{E}_x[s(x)] \end{aligned} \tag{31}$$

$$\mathbb{E}_x[(1 - a(x))(f(x) + s(x)) + t_F(x)] \geq \Pi := \frac{1}{\pi_p - \pi_c} \tag{32}$$

for all at least one feasible side contract $\mathcal{C} = (a, (b_F, b_I))$ (Lemma 4 will show that it does not matter *which* feasible side contract.). We refer to the left side as the factory's *incentive to comply* because this is by how much his expected payoff increases when he complies.

[16] The mafia's precise objective is not especially important since the government's goal is to ensure that its feasible set contains only null contracts.

The quantity Π is the factory's benefit from polluting divided by the marginal risk of being caught, which we refer to as his risk-adjusted benefit of polluting. Thus the CIC constraint says that the size of the factory's incentive to comply must exceed his risk-adjusted benefit of polluting.

Secondly, if the factory's expected payoff is too low then it will go out of business. Thus, if the government wants the factory to stay in business (we assume that it does) then it must ensure the factory's expected payoff is weakly positive when if it chooses to comply. This yields the factory's *compliance participation* constraint:

$$\pi_c \mathbb{E}_x[a(x)s(x) - (1-a(x))f(x) - t_F(x)] + (1-\pi_c)\mathbb{E}_x[s(x)] \geq 0 \tag{33}$$

for some side contract $\mathcal{C}$.

Thirdly, it must respect the players' limited liability constraints. The inspector cannot receive negative rewards, so the government must choose $r(x) \geq 0$ for all $x \in X$. If we relax this then the government benefit from fining the inspector whenever she fails to produce evidence. If we relax it enough then the government can incentivise the factory for free, even without using information design, because she can use fear of punishment to incentivise the monitor to report evidence, instead of relying on rewards. The factory is not protected by the same constraint, so the government can fine him, but it only has limited wealth, so the government cannot fine him more than some amount $\bar{f}$. Similarly, the government only has limited wealth so it cannot reward the factory more than $\bar{s}$. So it must choose $f(x) \leq \bar{f}$ and $s(x) \leq \bar{s}$ for all $x \in X$.

The cost of a scheme $\mathcal{S}$ is

$$c(\mathcal{S}) := \sup_{\mathcal{C} \text{ is feasible}} \mathbb{E}_x[\pi_c(1-a(x))(r(x)-f(x)) + (1-\pi_c(1-a(x)))s(x)]. \tag{34}$$

Analogous to Definition 4, we say that a scheme $\mathcal{S}$ *deters bribes* if and only if there are no feasible side contracts.

Lemma 4. *For every feasible scheme, there exists an bribery proof feasible scheme with the same cost.*

Proof. The idea is similar to Laffont and Martimort (1997) (CHECK). Let $\mathcal{S}$ be a feasible scheme. If $\mathcal{S}$ is not bribery proof then there must exist an interim efficient, non-null side contract, $\mathcal{C}$.[17] The government can create a new scheme $\mathcal{S}'$ that replicates the ex post payoffs of the side contract $\mathcal{C}$ under the original scheme $\mathcal{S}$. The fact that $\mathcal{S}$ and $\mathcal{S}'$ are ex-post payoff equivalent implies that $\mathcal{S}'$ is feasible and costs the same as $\mathcal{S}$. It only remains to show that $\mathcal{S}'$ is bribery proof.

Suppose it is not and that there exists a non-null, interim efficient side contract $\mathcal{C}'$. Let $\mathcal{C}''$ denote the side contract that delivers the same ex post payoff outcome under $\mathcal{S}$ that $\mathcal{C}'$ does under $\mathcal{S}'$, i.e. ... Then $\mathcal{C}''$ must interim dominate $\mathcal{C}$ because (i) the scheme, contract

[17] A side contract is *interim efficient* if it is feasible and there is no other feasible side contract that delivers a weakly higher ex post payoff to all types of both players, and a strictly higher payoff to at least one type of one player.

pair $(\mathcal{S},\mathcal{C}'')$ is ex post payoff equivalent to $(\mathcal{S}',\mathcal{C}')$, which interim dominates $(\mathcal{S}',\mathcal{C}_0)$, which is ex post payoff equivalent to $(\mathcal{S},\mathcal{C})$; and (ii) the fact that $\mathcal{C}'$ feasible implies $\mathcal{C}''$ is feasible. But this contradicts the fact that $(\mathcal{S},\mathcal{C})$ was assumed to be interim efficient, so it follows that $\mathcal{S}'$ must deter bribes.

□

The fact that the government can restrict attention to bribery proof schemes simplifies the problem dramatically, because it means that bribes need not feature in the player's payoffs. The cost of this simplification is that the scheme must satisfy a bribery proofness constraint. Formally, the *moral hazard problem* is

$$\begin{aligned}
\min_{\mathcal{S}=((X,\Sigma,p),(f,s,r))} \quad & \mathbb{E}[\pi_c(r(x)-f(x))+(1-\pi_c)s(x)] \\
\text{s.t. } & \mathbb{E}_x[f(x)+s(x)] \geq \Pi && \text{(CI)} \\
& \mathbb{E}_x[(1-\pi_c)s(x)-\pi_c f(x)] \geq 0 && \text{(CP)} \\
& r(x) \geq 0 \text{ and } f(x)+s(x) \leq \kappa \text{ for all } x \in X && \text{(LL)} \\
& \mathcal{S} \text{ is bribery proof.} && \text{(BP)}
\end{aligned}$$

The solution to the moral hazard problem is given by following corollary of Theorem 1.

Corollary 1 (Moral hazard)**.** *If the bound on the factory's payoffs is lower than his risk-adjusted benefit of polluting, i.e.* $\kappa < \Pi$*, then the moral hazard problem has no solution. Otherwise, the following scheme is optimal: the designer draws messages* x_F *and* x_I *uniformly and independently from the unit interval (i.e.* $X_F^* = X_I^* = [0,1]$ *and* p^* *is uniform), and pays transfers*

$$f^*(x) = (1-\pi_c)\frac{\kappa}{x_F}\max\{0, x_F - x_I^\lambda\} \tag{35}$$

$$s^*(x) = \pi_c\frac{\kappa}{x_F}\max\{0, x_F - x_I^\lambda\} \tag{36}$$

$$r^*(x) = \frac{\kappa}{x_I}\max\{0, x_I - x_F^{1/\lambda}\}, \tag{37}$$

where $\lambda = \sqrt{\frac{\kappa}{\kappa-\Pi}} - 1$*, This scheme costs* $c^* := \pi_c\left(\sqrt{\kappa} - \sqrt{\kappa-\Pi}\right)^2$.

Proof. We show that the moral hazard problem is equivalent to the trade problem with welfare weight equal to $\sqrt{\frac{\kappa}{\kappa-\Pi}} - 1$.

First, we note that that the (CP) constraint must hold with equality. Otherwise, if it were slack by some amount ϵ, then the government could increase the fine and reduce subsidy each by $\epsilon/2$. Doing so would clearly we cheaper. It would also be feasible because the left side of the (CP) constraint would be reduced by $\epsilon/2$, so it would still be greater than 0; and the sum $f(x)+s(x)$ would be unchanged so the (CI), (LL) and (BP) constraints

would continue to hold. Substituting the (CP) constraint into the objective, and dividing the objective by the constant π_c simplified the problem to

$$\begin{aligned}
\min_{\mathcal{S}=((X,\Sigma,p),(f,s,r))} \quad & \mathbb{E}[r(x)] \\
\text{s.t. } & \mathbb{E}_x[f(x)+s(x)] \geq \Pi && \text{(CI)} \\
& \mathbb{E}_x[(1-\pi_c)s(x)-\pi_c f(x)] = 0 && \text{(CP)} \\
& r(x) \geq 0 \text{ and } f(x)+s(x) \leq \kappa \text{ for all } x \in X && \text{(LL)} \\
& \mathcal{S} \text{ is bribery proof.} && \text{(BP)}
\end{aligned}$$

Next, we restate the problem in terms of the variables $t_B(x) = \kappa - (f(x)+s(x))$ and $t_S(x) = r(x)$. The variable f and s occur everywhere in the problem as the sum $f(x)+s(x)$ except in the (CP) constraint. But if we can solve for $t_B(x)$ then we can satisfy (CP) by setting $s(x) = \pi_c(\kappa - t_B(x))$ and $f(x) = (1-\pi_c)(\kappa - t_B(x))$. Hence, the problem reduces to

$$\begin{aligned}
\min_{\mathcal{S}=((X,\Sigma,p),(t_B,t_S))} \quad & \mathbb{E}[t_S(x)] \\
\text{s.t. } & \mathbb{E}_x[t_B(x)] \leq \kappa - \Pi && \text{(CI)} \\
& t_S(x) \geq 0 \text{ and } t_B(x) \geq 0 \text{ for all } x \in X && \text{(LL)} \\
& \mathcal{S} \text{ is bribery proof.} && \text{(BP)}
\end{aligned}$$

The only difference between this problem and the trade problem (Equation 7) is that the buyer's expected transfer $\mathbb{E}_x[t_B(x)]$ enters the moral hazard problem through the a constraints, with lower bound $\kappa - \Pi$, whereas it enters the trade problem through the objective with welfare weight λ. Setting the welfare weight equal to $\lambda^* := \sqrt{\frac{\kappa}{\kappa-\Pi}} - 1$ yields exactly

$$\mathbb{E}[t_B(x)] = \kappa \frac{1}{(\lambda^*+1)^2} = \kappa - \Pi. \tag{38}$$

A higher welfare weight results in a higher expected value of t_S, and therefore has a higher cost; a lower welfare weight results in a higher expected value of t_B, violating (CI). Therefore the solution to the this problem is the same as the solution to the trade problem with the specific welfare weight λ^*; The value of this problem is equal to the expected payoff of the seller in the trade problem, which is

$$\begin{aligned}
\kappa \frac{(\lambda^*)^2}{(\lambda^*+1)^2} &= \kappa \frac{1}{(\lambda^*+1)^2} \left(\sqrt{\frac{\kappa}{\kappa-\Pi}} - 1\right)^2 && (39) \\
&= (\kappa-\Pi)\left(\frac{\kappa}{\kappa-\Pi} - 2\sqrt{\frac{\kappa}{\kappa-\Pi}} + 1\right) && (40) \\
&= \kappa - 2\sqrt{\kappa(\kappa-\Pi)} + \kappa - \Pi && (41) \\
&= \left(\sqrt{\kappa} - \sqrt{\kappa-\Pi}\right)^2. && (42)
\end{aligned}$$

Finally, the cost of the scheme is equal to probability finding incriminating evidence in equilibrium, times the inspectors expected reward, which is $\pi_c\left(\sqrt{\kappa}-\sqrt{\kappa-\Pi}\right)^2$. □

We conclude this section by comparing the two-sided lemons mechanism with the optimal on-sided mechanism. The reasons for presenting this particular scheme are threefold. Firstly, it demonstrates that the government can attain the first best outcome if it can use infinitely large fines. We consider the case of infinitely large fines to be unrealistic, so this fact motivates us to consider cases where fines are bounded. Secondly, it is the optimal one-sided scheme (when the bound on fines is not too small[18]), so the fact that our two-sided lemons scheme costs strictly less than it motivates our interest in two-sided lemons schemes. Thirdly, it provides the best basis for comparing our main result with previous literature (Ortner and Chassang, 2018).

Proposition 1 (Optimal one-sided scheme (Informed Inspector))**.** *For any constant $k \geq \Pi$, the scheme $(p^{II}, (f^{II}, s^{II}, r^{II}))$ defined by*

$$p^{II} \text{ uniform on } [0,1]$$

$$s^{II}(x_I) = \begin{cases} 0 & \text{if } x_I \leq \frac{k-\Pi}{k} \\ (1-\pi_c)k & \text{otherwise} \end{cases}$$

$$s^{II}(x_I) = \begin{cases} 0 & \text{if } x_I \leq \frac{k-\Pi}{k} \\ \pi_c k & \text{otherwise} \end{cases}$$

$$r^{II}(x_I) = \begin{cases} 0 & \text{if } x_I \leq \frac{k-\Pi}{k} \\ k - \frac{k-\Pi}{x_I} & \text{otherwise,} \end{cases}$$

1. *deters bribes and satisfies the factory's voluntary participation and incentive compatibility constraints;*
2. *costs $c^{II} := \pi_c\left[\Pi + (k-\Pi)\ln\left(1-\frac{\Pi}{k}\right)\right] \xrightarrow{k\to\infty} 0$;*
3. *costs less than any other one-sided, informed inspector scheme that satisfies the factory's voluntary participation and incentive compatibility constraints.*

Proof. We prove here that the informed inspector scheme deters bribes. The rest of the proof is given in subsection 8.1.

Consider a bribe b. The inspector will agree to the bribe if and only if she receives a message for which her reward $r(x_I)$ is less than the bribe b. We have $r(x_I) = k - \frac{k-\Pi}{x_I}$ so the inspector agrees to the bribe if and only if $b \geq k - \frac{k-\Pi}{x_I}$, or equivalently, $x_I \leq \frac{k-\Pi}{k-b}$. This is

[18] If the bound on fines is small enough then the one-sided informed inspector scheme is undercut by a one-sided informed firm scheme.

an example of a cutoff strategy with cutoff equal to $\frac{k-\Pi}{k-b}$. the factory's expected incentive, conditional on the inspector's cutoff strategy, is equal to

$$\begin{aligned}
&\mathbb{E}\left[f^{\mathrm{II}}(x_I)+s^{\mathrm{II}}(x_I)\;\middle|\; b\geq k-\frac{k-\Pi}{x_I}\right]\\
=&\mathbb{P}\left[x_I\leq\frac{k-\Pi}{k}\;\middle|\; x_I\leq\frac{k-\Pi}{k-b}\right]0+\mathbb{P}\left[x_I\geq\frac{k-\Pi}{k}\;\middle|\; x_I\leq\frac{k-\Pi}{k-b}\right]k\\
=&\min\left\{1,\frac{\frac{1}{k-b}-\frac{1}{k}}{\frac{1}{k-b}}\right\}k\\
=&\min\{k,b\},
\end{aligned}$$

so he is indifferent about accepting bribes less than k, and strictly prefers to reject bribes greater than k. If $b=0$, then the inspector strictly prefers to take her reward if her message is $x_I>1-\frac{\Pi}{k}$, otherwise she is indifferent. Therefore, the factory's conditional expected incentive is equal to 0, so he is indifferent as well. In all cases, the government can deter the players from exchanging the zero bribe at an arbitrarily small cost (e.g. by adding ϵ to the inspector's reward). □

The key feature of this scheme is that the distribution of rewards is chosen so that the factory's probability of facing a peach conditional on a given bribe increases in proportion to the size of the bribe, so as to keep him indifferent about accepting the bribe. In other words, 'peach inspectors' enter the market at the highest rate possible without giving the factory a strict preference to enter the market. In the limit, the factory's incentive, k, becomes arbitrarily large with vanishing probability, which corresponds to the use of extreme incentives in Becker (1968). By contrast, the inspector's reward never exceeds Π. Since the government only has to pay the inspector with the same vanishing probability that she punishes the factory, the inspector's expected reward can be made arbitrarily small. This in turn means that the cost of the scheme approaches the first best cost, so no scheme can do better. We show in subsection 8.2 that, when the k is restricted to be small enough, there is an informed firm scheme that costs less than the optimal informed inspector scheme.

The two-sided lemons scheme is a substantial improvement on the optimal one-sided scheme. In the one-sided scheme, the inspector's willingness to accept bribes was decreasing in her own message, so she only wanted to agree to bribes when her message was below a certain cutoff. The factory's willingness to accept bribes was increasing in the inspector's message, so the inspector's choice of cutoff strategy made him unwilling to accept any bribes. In the two-sided lemons scheme, the inspector is both informed (about her own message) and uninformed (about the factory's message) so she inherits both of these features. Her willingness to accept bribes is decreasing in her own message, so she continues to adopt a cutoff strategy. But her willingness to accept bribes is increasing in the factory's message, so her optimal choice of cutoff is decreasing in her belief about the factory's message. Her belief about the factory's message depends on his strategy. The

distribution of transfers ensures that his best response is also a cutoff strategy, and his optimal cutoff is decreasing in his belief about the inspector's message. The distribution of transfers is chosen so that each player wants to choose a cutoff that is slightly below the cutoff that the other player chooses. Therefore there can be no equilibrium in which they both use a positive cutoff, which means that they do not agree to bribes in any equilibrium. Thus, unlike the informed inspector scheme, the two-sided lemons scheme uses contagion in higher order beliefs to amplify the adverse selection problem.

Another closely related paper to ours is von Negenborn and Pollrich (2020). They also find that engineering a lemons problem is an optimal solution to a mechanism design problem. Our main contribution relative to theirs is that we impose bounds on all transfers. Whereas their proposed mechanisms attain the first best by using large rewards and/or punishments. Therefore, they do not need to engineer an optimal lemons problem – any lemons problem would suffice. E.g., consider a biased coin with $\mathbb{P}[L] = q := 1 - \sqrt{1 - \frac{\Pi}{k}}$.

- Setting $f(P) + s(P) = qk$ deters distraction and bribes.
- Cost equals $\sqrt{k}(\sqrt{k} - \sqrt{k - \Pi})$.
- Cost *decreases* in k and converges to $\frac{\Pi}{2}$ as $k \to \infty$.

Table 2 compares the cost of the two-sided lemons scheme to the one-sided (informed inspector) scheme (and others) for a range of parameter values. We show in a subsection 8.3 that the two-sided lemons scheme costs strictly less than the informed inspector scheme at all parameter values and that the cost of the two-sided lemons scheme converges to *half* the cost of the optimal one-sided scheme as κ gets large.

Table 2: The costs of selected schemes for parameters $\pi_p = \frac{2}{3}, \pi_c = \frac{1}{3}, \Pi = 3$.

Scheme	$\kappa = 3$	$\kappa = 4$	$\kappa = 6$	$\kappa = 30$	$\kappa = \infty$
deterministic	1	1	1	1	1
fair coin toss (informed firm)	—	0.667	0.667	0.667	0.667
biased coin toss (informed firm)	1	0.667	0.586	0.513	0.5
biased coin toss (informed inspector)	1	0.75	0.5	0.1	0
one-sided (informed inspector)	—	0.538	0.307	0.052	0
two-sided	1	0.333	0.172	0.026	0

7 Conclusion

We show how information design can be used together with transfer design to deter bribes by engineering a lemons problem. The optimal scheme characterised in our main result, Theorem 1, accommodates monitoring errors, costs strictly less than other schemes in

the literature, and is relatively simple to implement. This scheme also gives insights into 'worst case' information structures and gives an upper bound on the amount of surplus lost to contagion in bargaining games.

We have shown that the two-sided lemons scheme deters bribes, but bribery contracts are only a special case of an incentive compatible side-contract. Side contracts additionally allow for the possibility of message dependent bribes, $b(x)$, and correlated agreement strategies that depend on both messages, $\sigma(x)$. We conjecture that the two-sided lemons scheme deters all side contracts. One possibility for future research is to prove this by showing that the core of the cooperative game with incomplete information (Myerson, 2007; Forge and Serrano, 2013) induced by the scheme is empty.

Studying the core of the cooperative game induced by the two-sided lemons scheme would also be valuable for extending our results to more than two players. We see this as a particularly promising avenue for future research because it could help us to utilise the information held by potential whistleblowers. Specifically, if the lemons problem can be made disproportionately worse by spreading information across an even larger number of 'inspectors', then we expect to find that the costs of implementing compliance can be further reduced by offering stochastic rewards to whistleblowers. This stands in contrast to the case without information design, where hiring multiple monitors does not help to deter bribery (Stapenhurst, 2019).

Finally, the use of endogenous lemons to deter collusion may have applications beyond monitoring. For instance, it can be used to deter illegal trades, such as weapons, drugs, and human tracking. We also speculate that it could also be used to break up cartels or to deter sub-coalitions of would-be signatories from undermining international agreements.

8 Mathematical appendix

8.1 Proof of proposition 1

The cost of the scheme $\mathcal{S}^{II} = (q^{\text{II}}, (f_e^{\text{II}}, f_0^{\text{II}}, r^{\text{II}}))$ is

$$\begin{aligned}
&\mathbb{E}[\pi_c(r^{\text{II}}(x_I) - f_e^{\text{II}}(x_I)) + (1-\pi_c)(-f_0^{\text{II}}(x_I))] \\
=&\int_{\frac{k-\Pi}{k}}^{1} \pi_c r^{\text{II}}(x_I) - (\pi_c f_e^{\text{II}}(x_I) + (1-\pi_c) f_0^{\text{II}}(x_I)) dx_I \\
=&\int_{\frac{k-\Pi}{k}}^{1} \pi_c \left(k - \frac{k-\Pi}{x_I}\right) - (\pi_c(1-\pi_c)k - (1-\pi_c)\pi_c k) dx_I \\
=&\pi_c \int_{\frac{k-\Pi}{k}}^{1} k - \frac{k-\Pi}{x_I} dx_I \\
=&\pi_c \left(k\left(1 - \frac{k-\Pi}{k}\right) - (k-\Pi)\Big[\ln x_I\Big]_{\frac{k-\Pi}{k}}^{1}\right) \\
=&\pi_c \left[\Pi + (k-\Pi)\ln\left(1 - \frac{\Pi}{k}\right)\right] \\
\leq&\pi_c \left[\Pi - (k-\Pi)\frac{\Pi}{k}\right] \\
=&\pi_c \frac{\Pi^2}{k} \xrightarrow{k\to\infty} 0,
\end{aligned}$$

where the inequality comes from the fact that $\ln(1+x) \leq x$ for all $x > -1$ (Topsøe, 2004).

We now show that the scheme is optimal by showing that any feasible solution to the government's problem costs weakly more than $\mathcal{S}^{II}$. In any scheme where Ina has full information about the transfers, she will accept the bribe b whenever she receives a message x_I such that $r(x_I) < b$. Finn anticipates this, so he accepts bribe b if $\mathbb{E}[f_e(x_I) - f_0(x_I) | r(x_I) < b] > b$. Therefore, any informed inspector scheme which deters bribes must at least satisfy $\mathbb{E}[f_e(x_I) - f_0(x_I) | r(x_I) < b] \leq b$ for all bribes $b \geq 0$.

Define Ina's expected incentive when Ina rejects a bribe b by $E(b) := \mathop{\mathbb{E}}_{x\sim q}[f_e(x_I) - f_0(x_I) | r(x_I) \geq b]$, and $F_r(b) = \mathop{\mathbb{P}}_{x\sim q}[r(x_I) < b]$. We use the fact that $E(0) = \mathbb{E}[f_e(x_I) - f_0(x_I)] = \mathbb{E}[f_e(x_I) - f_0(x_I) | r(x_I) < b] F_r(b) + E(b)(1 - F_r(b))$ for any b, to rewrite the no-bribery constraint as $\frac{E(0) - E(b)(1-F_r(b))}{F_r(b)} \leq b$. This rearranges to give $F_r(b) \leq \frac{E(b)-E(0)}{E(b)-b}$. If $E(0) < b$ then $\frac{E(b)-E(0)}{E(b)-b} > 1$, so the constraint is slack. Otherwise, if $E(0) \geq b$, then $\frac{E(b)-E(0)}{E(b)-b}$ is increasing in $E(b)$ and decreasing in $E(0)$. Finn's limited liability constraint requires that $E(b) \leq k$ and his incentive compatibility constrain requires that $E(0) \geq \Pi$. It follows that every bribery-proof informed inspector scheme satisfies

$$F_r(b) \leq \frac{E(b) - E(0)}{E(b) - b} \leq \frac{k - \Pi}{k - b}.$$

In $\mathcal{S}^{II}$, the distribution of rewards is

$$F_r^{II}(b) = \mathop{\mathbb{P}}_{x_I \sim q^{II}}[r^{II}(x) < b] = \mathop{\mathbb{P}}_{x_I \sim q^{II}}[x_I < \frac{k-\Pi}{k-b}] = \frac{k-\Pi}{k-b}.$$

Hence the distribution of rewards in any bribery-proof solution must first order stochastically dominate the distribution of rewards in the scheme $\mathcal{S}^{II}$. This implies that every bribery-proof solution has a weakly higher expected reward than does $\mathcal{S}^{II}$. At the same time, $\mathcal{S}^{II}$ exactly satisfies Finn's (VP) constraint, so every feasible scheme must have a weakly lower expected fine than $\mathcal{S}^{II}$. The cost of a scheme is given by Ina's expected reward minus Finn's expected fine, so it follows that every feasible scheme must cost weakly more than $\mathcal{S}^{II}$.

8.2 Informed firm schemes

Consider the biased coin toss (informed firm) scheme described in table 3. Similar tech-

Table 3: The biased coin toss (informed firm) scheme.

	Lemon	Peach
Probability	$1-\sqrt{1-\Pi/\kappa}$	$\sqrt{1-\Pi/\kappa}$
Fine Finn	4	$\left(1-\sqrt{1-\Pi/\kappa}\right)\kappa$
Reward Ina	4	0

niques to those used in section 2 show that this scheme satisfies voluntary participation, incentive compatibility and deters bribes. The cost of the scheme is $\pi_c\sqrt{\kappa}(\sqrt{\kappa}-\sqrt{\kappa-\Pi})$. When $\kappa = 4$ and $\Pi = 3.9$ we get that this informed firm scheme costs $3.4\pi_c$, whereas the cheapest informed inspector scheme costs $\pi_c\left[\Pi + (k-\Pi)\ln\left(1-\frac{\Pi}{k}\right)\right] = 3.5\pi_c$. In general, informed firm schemes are cheaper when κ is small enough relative to Π.

The optimal informed firm scheme takes the following form,

$$q^{\text{IF}} \text{ is uniform on } [0,1]$$

$$f_e^{\text{IF}}(x_F) = (1-\pi_c)\begin{cases} k & \text{if } x_F \le \tilde{x}^{IF} \\ \frac{\tilde{x}^{IF}}{x_F} & \text{otherwise} \end{cases}$$

$$f_0^{\text{IF}}(x_F) = -\pi_c\begin{cases} k & \text{if } x_F \le \tilde{x}^{IF} \\ \frac{\tilde{x}^{IF}}{x_F} & \text{otherwise} \end{cases}$$

$$r^{\text{IF}}(x_F) = \begin{cases} k & \text{if } x_F \le \tilde{x}^{IF} \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{x}^{IF}$ solves $\tilde{x}^{IF}\ln\left(\frac{e}{\tilde{x}^{IF}}\right) = \frac{\Pi}{\kappa}$. Similar techniques to those used in appendix 8.1 show that this scheme is feasible and cheaper than any other informed firm scheme. However, it is not easy to work with because no analytical solution for $\tilde{x}^{IF}$ exists.

8.3 The two sided scheme costs strictly less than the one sided schemes.

The informed inspector scheme costs strictly more than the two-sided scheme at all parameter values:

$$\begin{aligned}
c^{\mathrm{II}}/\pi_c &= \Pi + (\kappa - \Pi)\ln\left(1 - \frac{\Pi}{\kappa}\right) \\
&\geq \Pi - (\kappa - \Pi)\frac{\frac{\Pi}{\kappa}}{\sqrt{1 - \frac{\Pi}{\kappa}}} \\
&= \Pi - (\kappa - \Pi)\frac{\frac{\Pi}{\sqrt{\kappa}}}{\sqrt{\kappa - \Pi}} \\
&= \Pi - \sqrt{\kappa - \Pi}\frac{\Pi}{\sqrt{\kappa}} \\
&= \frac{\Pi}{\sqrt{\kappa}}\left(\sqrt{\kappa} - \sqrt{\kappa - \Pi}\right) \\
&> (\sqrt{k} - \sqrt{\kappa - \Pi})^2 \\
&= c^*/\pi_c,
\end{aligned}$$

where the first inequality results from the fact that $\ln(1+x) \geq \frac{x}{\sqrt{1+x}}$ for all $x \in (-1, 1]$ (Topsøe, 2004), and the second comes from the fact that

$$\begin{aligned}
\sqrt{\kappa} &> \sqrt{\kappa - \Pi} \\
\sqrt{\kappa}\sqrt{\kappa - \Pi} &> \kappa - \Pi \\
\Pi &> \kappa - \sqrt{\kappa}\sqrt{\kappa - \Pi} \\
\frac{\Pi}{\sqrt{\kappa}} &> \sqrt{\kappa} - \sqrt{\kappa - \Pi}.
\end{aligned}$$

Now we show that the two-sided scheme costs half as much as the informed inspector scheme in the limit.

$$\begin{aligned}
\frac{c^*}{c^{\Pi}} &\leq \frac{\sqrt{\kappa}}{\Pi}(\sqrt{\kappa} - \sqrt{\kappa - \Pi}) \\
&= \frac{\kappa - \sqrt{\kappa(\kappa - \Pi)}}{\Pi} \\
&= \frac{\frac{\kappa - \sqrt{\kappa(\kappa-\Pi)}}{\Pi}(\kappa + \sqrt{\kappa(\kappa - \Pi)})}{\kappa + \sqrt{\kappa(\kappa - \Pi)}} \\
&= \frac{\frac{\kappa^2 - \kappa(\kappa-\Pi)}{\Pi}}{\kappa + \sqrt{\kappa(\kappa - \Pi)}} \\
&= \frac{\kappa(\kappa - \kappa(\kappa - \Pi))}{\Pi} \\
&= \frac{\kappa}{\kappa + \sqrt{\kappa(\kappa - \Pi)}} \xrightarrow{\kappa \to \infty} \frac{1}{2},
\end{aligned}$$

where the first inequality comes from the previous calculations.

References

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

Bahoo, S., I. Alon, and A. Paltrinieri (2020). Corruption in international business: A review and research agenda. *International Business Review 29*(4), 101660.

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.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy 76*(2), 169–217.

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

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

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2013, 09). Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India*. *The Quarterly Journal of Economics 128*(4), 1499–1545.

Forge, F. and R. Serrano (2013). Cooperative games with incomplete information: Some open problems. *International Game Theory Review 15*(02), 1340009.

Garrett, D., G. Georgiadis, A. Smolin, and B. Szentes (2021). Optimal technology design.

Gründler, K. and N. Potrafke (2019). Corruption and economic growth: New empirical evidence. *European Journal of Political Economy 60*, 101810.

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

Isaksson, A.-S. and A. Kotsadam (2018). Chinese aid and local corruption. *Journal of Public Economics 159*, 146–159.

Kajii, A. and S. Morris (1997). The robustness of equilibria to incomplete information. *Econometrica 65*(6), 1283–1309.

Kofman, F. and J. Lawarrée (1993). Collusion in hierarchical agency. *Econometrica 61*(3), 629–656.

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

Laffont, J.-J. and D. Martimort (2002). *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press.

Mathevet, L., J. Perego, and I. Taneva (2020). On information design in games. *Journal of Political Economy 128*(4), 1370–1404.

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

Morris, S. and T. Ui (2005). Generalized potentials and robust sets of equilibria. *Journal of Economic Theory 124*(1), 45–78.

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

Myerson, R. B. (2007). Virtual utility and the core for games with incomplete information. *Journal of Economic Theory 136*(1), 260–285.

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.

Stapenhurst, C. (2019). *How Many Corruptible Monitors does it take to Implement an Action?* Ph. D. thesis, University of Edinburgh.

Strausz, R. (1997). Delegation of monitoring in a principal-agent relationship. *Review of Economic Studies 64*(3), 337–357.

Tacconi, L. and D. A. Williams (2020). Corruption and anti-corruption in environmental and resource management. *Annual Review of Environment and Resources 45*(1), 305–329.

Taneva, I. (2019, November). Information design. *American Economic Journal: Microeconomics 11*(4), 151–85.

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

Topsøe, F. (2004). Some bounds for the logarithmic function. *RGMIA Res. Rep. Collection 7*(2), 1–20.

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