Corruption, Extortion, and the Boundaries of the Law[*]

Svetlana Andrianova[†] Nicolas Melissas[‡]

November 27, 2007

Abstract

We consider a set-up in which a principal must decide whether or not to legalise a socially undesirable activity. The law is enforced by a monitor who may be bribed to conceal evidence of the offence and who may also engage in extortionary practices. The principal may legalise the activity even if it is a very harmful one. The principal may also declare the activity illegal knowing that the monitor will abuse the law to extract bribes out of innocent people. Our model offers a novel rationale for legalising possession and consumption of drugs while continuing to prosecute drug dealers.

JEL-codes: D82, L22, K4.

Keywords: Moral Hazard, Collusion, Bribery, Non-contractible Output, Rewards and Punishments.

[*] We would like to thank without implicating Cecile Aubert, John Conley, Gianni De Fraja, Panicos Demetriades, Mathias Dewatripont, Paul Heidhues, Maurice Kugler, Claudio Mezzetti, Tim Worrall, two anonymous referees and Tracy Lewis, our co-editor. We also thank seminar participants at ITAM and at the University of Leicester for helpful comments and discussions.

[†] University of Leicester, University Road, Leicester LE1 7RH, UK. E-mail s.andrianova@le.ac.uk.

[‡] Corresponding author. Centro de Investigación Económica, Instituto Tecnológico Autónomo de México, Camino a Santa Teresa 930, México D.F. 10700, Mexico. E-mail nmelissas@itam.mx. Tel + 52.55.56.28.40.00 Fax + 52.55.56.28.40.58

1 Introduction

Corruption is—and has always been—a source of concern for policy makers. For example, Theodore Roosevelt in a 1903 speech said: "we must show our abhorrence of ... corruption, in public and in private life".[1] More recently, in 1999 the World Bank president James Wolfensohn asserted that: "... so far as [the World Bank] is concerned, there is nothing more important than the issue of corruption".[2]

Obviously, corruption can take many forms as it is prevalent in the political, corporate and criminal spheres of society. To illustrate the forms of corruption that we will study in this paper, consider the following examples.[3] In the 1960s and early 1970s the Hong Kong police was very corrupt and regularly received bribes in order not to report illegal gambling activities. Prior to 1975 the Philippines' Bureau of Internal Revenue was renowned for its corrupt tax inspectors. In one famous example two tax inspectors found evidence of a 2 million pesos tax fraud. Instead of reporting this to their superiors, they went to the contravening company and asked for one million pesos in exchange for their silence. What those examples share in common is that enforcers receive bribes in order to conceal evidence of illegal behaviour. We refer to this practice as *corruption.*

In a related vein, examples abound in which one party creates fake evidence to indict another. For example, in Lahore (Pakistan) police officers are notorious for putting drugs in the hotel rooms of unsuspecting tourists. The tourists then have to bribe the police officers to avoid going to jail. Similarly, in Guatemala the police force threatened to rip up the travel permits of Central American migrants unless the latter came up with a bribe of 150 quetzals (approximately 20 USD).[4] Practices of this kind illustrate what we model as *extortion* in this paper. If the police officer extorts, he creates fake evidence of

[1] Address at the Dedication Ceremonies of the Louisiana Purchase Exposition on April 30, 1903. http://www.jmu.edu/madison/center/main_pages/madison_archives/life/secretary/la_purchase/roosevelt.htm

[2] Plenary Address at the 9th International Anti-Corruption Conference in Durban, South Africa, 9–15 October, 1999. http://ww1.transparency.org/iacc/9th_iacc/papers/day1/plenary/d1pl_jwolfensohn.html

[3] All examples in this paragraph are taken from Klitgaard (1988).

[4] http://ciss.insp.mx.migracion/index.php?section=noticias&id_not=734&pagina=24.

an offence that would give him the opportunity to extract a bribe from someone. Note that we define extortion as planting fake evidence of an offence regardless of whether the "offender" pays a bribe or ends up in jail.[5]

In this paper we develop a principal-monitor-agent model to study law enforcement policy in the presence of corruption and extortion. The agent derives a private benefit from undertaking a socially undesirable activity.[6] The monitor chooses whether to monitor, to extort, or to do nothing. Monitoring involves a disutility for the monitor, but provides evidence of the offence, if committed. If the monitor finds evidence of wrongdoing, he can either accept a bribe from the agent, or he can send the evidence to the principal (i.e. the social planner). Extorting involves a lower disutility compared to monitoring, but allows the monitor to fabricate fake evidence with which to threaten to indict the agent. If the principal receives "evidence" of wrongdoing, she cannot know whether the "evidence" is genuine or fake. Therefore, she instructs a second monitor to evaluate the evidence. If the evidence is fake, the second monitor discovers it with some small probability and truthfully reports his findings to the principal. In this setup, the principal must decide (i) whether or not to declare the activity illegal and if so (iia) the jail sentence to be imposed on the agent and the reward to be given to the monitor in case the monitor reports evidence of the offence and in case the second monitor does not conclude that the evidence is fake, and (iib) the jail sentence to be imposed on the monitor in case the second monitor concludes that the evidence is fake.

Suppose the principal punishes a proven offence by a one-year jail sentence which is equivalent to a \$70 loss for the agent. Suppose the principal also gives \$50 to the monitor whenever he presents evidence of wrongdoing (and, of course, whenever the second monitor does not conclude that the evidence is fake). This penalty/reward scheme "allows" corruption to take place: The agent would prefer to pay, say, \$60 to the monitor

[5] All our examples clearly focus on "victimless crimes", all are from less developed countries where corruption is strife. In Section 4 we specify more carefully the environments in which our model and its conclusions apply.

[6] For example, the activity could be: smoking cannabis, drinking alcohol, gambling, etc.

and the monitor would prefer to receive \$60 from the agent rather than \$50 from the principal. Suppose now that the principal punishes a proven offence with a nine-month jail sentence which is equivalent to a \$60 loss for the agent. Suppose the principal also rewards \$61 to the monitor whenever he presents evidence of wrongdoing. In this example, corruption will not occur: The agent is not willing to pay a bribe higher than \$60 and the monitor does not accept a bribe inferior to \$61. Observe that incentives to monitor and incentives to break the law are almost identical in both penalty/reward schemes. Despite this, the principal strongly prefers the first penalty/reward scheme for two different reasons. First, in the latter case, she needs to raise taxes to pay \$61 to the monitor. (Taxes are assumed to be distortionary and therefore reduce welfare). Second, in the latter case she sends the agent to jail, which is an inefficient way of punishing her.[7] Therefore, in equilibrium the principal chooses a penalty/reward scheme such that the agent bribes the monitor both when she was caught breaking the law and when she was framed.

We also show that, in equilibrium, the monitor gets a bigger bribe if he catches the agent breaking the law than if he extorts. The reason for this is twofold. First, in the case when he extorts, the monitor is more eager to reach a settlement with the agent as he is afraid that, in case he sends the "evidence" to the principal, the second monitor will uncover his plot. Second, in the case when the monitor monitors, the agent knows that if she bribes the monitor she will retain her private benefit (e.g. in our drug example, the agent will keep her cannabis and will be able to enjoy smoking it). Hence, the monitor faces a trade-off: if he monitors, he incurs a disutility but it might allow him to extract a bigger bribe. The principal knows that, through the penalty/reward scheme, she can influence both bribes and with it the incentives to monitor, to extort and to obey the law.[8] Clearly, both bribes are increasing in the monitor's reward and in the

[7] The observation that it may sometimes be efficient to allow bribery for reasons indicated above is already noted in Becker and Stigler (1974).

[8] A detailed explanation of the interplay between direct and indirect channels through which a bribe affects the monitor's incentive to monitor and the agent's incentive to break the law awaits in Section 3.

agent's jail term. If the monitor extorts, he obtains a bribe with probability one, while if he monitors he obtains a bribe only if a genuine offence took place (which happens with a probability less than one). Therefore, the monitor's gain from extorting increases faster in the penalty/reward scheme than his gain from monitoring. As the amount of monitoring decreases, the gain from offending increases, and this reduces the amount of law abiding behaviour. Hence, the principal faces the following trade-off: if she increases the penalty/reward scheme, the monitor will monitor less (which is good for welfare as it reduces monitoring costs), but the agent will break the law more often.[9]

Which of these two opposite effects dominates depends on the severity of the offence (the more severe the offence, the lower the welfare when the agent broke the law). We find that when the offence is "very severe", law abiding behaviour is bound to increase welfare by a large amount and compensate for the requisite enforcement cost. In this case, the social planner sets the penalty/reward scheme such that both bribes are as low as possible. The monitor then always monitors and never extorts. If the offence is "not severe", the principal is much more concerned about reducing monitoring costs than inducing the agent into law abiding behaviour. She will therefore prefer to legalise the activity.[10] We also show that this "not severe"-region can — depending on the values of our exogenous parameters — be arbitrarily large. Suppose now that the offence is "moderately severe" (i.e. the offense lies between "not severe" and "very severe"). In this case the principal is also more concerned about reducing monitoring costs than in inducing the agent to obey the law. Therefore, if she declares the activity illegal, she chooses a penalty/reward scheme which induces the highest possible bribes.[11] These high

[9] A study of corruption in the medical sector in Buenos Aires in Di Tella and Schargrodsky (2003) provides empirical evidence for this trade-off.

[10] The U.S. interwar experience nicely illustrates that governments sometimes prefer to legalise undesirable activities. As is well known, alcohol prohibition led to widespread corruption and little law abiding behaviour. This induced some states to legalise the consumption of alcohol in the 1930's.

[11] The highest possible bribe is not an infinitely large number in our model as we realistically assume that the agent's punishment is bounded above. This upper bound on the agent's punishment also puts an upper bound on the monitor's reward as the principal must ensure that the agent successfully bribes

bribes induce the monitor to monitor "very little". This very low monitoring intensity induces the agent into a "small amount" of law abiding behaviour. As the offence is "moderately severe", the "small amount" of law abiding behaviour compensates for the "very little" monitoring cost and the principal is better off by declaring the activity illegal. Observe that in this case the principal declares the activity illegal even though she knows that the monitor will abuse the law to extort money out of innocent individuals. In Section 4 we discuss the policy implications of our results.

Our paper belongs to the literature on collusion (or corruption) when the principal cannot contract on output. No paper in that literature has found that the principal can gain by legalising a socially undesirable activity. The closest to ours is the paper by Polinsky and Shavell (2001).[12] To understand some of their results, suppose the economy is populated by 100 monitors. 70 of them are in the business of catching offenders. 30 of them are in the business of producing fake evidence. In contrast to our paper, the ratio 70/30 is *independent* of the penalty/reward scheme. The authors show that it may be optimal to set the reward at a high level. This high level discourages the 70 monitors from engaging in corruption (and induces more agents into law abiding behaviour). The high reward, however, induces the 30 monitors into producing fake evidence to extract a bribe from an agent. We show that the Polinsky and Shavell equilibrium extortion result is robust when the activity choice of the monitor is endogenous. Moreover, in their paper it is never optimal to legalise the undesirable activity because it is implicitly assumed that both corruption and extortion activities are costless. In contrast, we find that the

the monitor when she was framed.

[12] See also, a.o., Strausz (1997), Hindriks et al. (1999), Carrillo (2000a, 2000b), and Mishra (2002). Strausz (1997) analyses a set-up similar to ours. In Carrillo (2000a, 2000b) the monitor's wage is independent of his report. Mishra (2002) considers a two-monitor set-up and, using efficiency as a benchmark, compares a vertical hierarchy (i.e. the second monitor monitors the first one) with a horizontal one (i.e. both monitors monitor the agent). All four papers rule out extortion. Hindriks et al. (1999) analyse a set-up in which a tax inspector can extort and engage in corruption. However, they are primarily interested in studying the efficiency and equity consequences of different tax schemes which is not the purpose of this paper.

principal prefers to legalise the activity precisely because monitoring is more costly (in terms of effort) than extorting.

Another strand in the literature on collusion (or corruption) in three-tier organizations considers a set-up with adverse selection and moral hazard in which the principal can contract on the agent's output.[13] In particular, in Khalil et al. (2007) extortion may occur when the agent worked hard but her output turned out to be low. Then the principal sends a monitor to check the agent's effort level. If the monitor finds that the agent worked hard, he has an incentive to extort the agent by threatening non-disclosure of the evidence favourable to the agent. If the principal chooses a reward scheme (a wage for the monitor and a wage for the agent) such that this kind of extortion takes place, then the agent is rewarded less when she provides effort. This reduces her incentives to exert effort in the first place. Anticipating this, the principal selects a reward scheme such that, in equilibrium, the monitor never extorts. Kofman and Lawarrée (1993) show that the principal may prefer not to hire the monitor[14] if her signal is sufficiently imprecise or if the agent's maximal punishment is low. We show that their negative result also holds if the principal cannot contract on output provided that the monitor can extort and engage in corruption. Moreover, in our model the principal may prefer to legalise the activity even if the monitoring technology is perfect and if the agent's maximal punishment is arbitrarily high. Khalil and Lawarrée (2003) assume that the principal cannot commit ex ante to hiring the monitor after the output realisation. They show that, due to conflicting non-commitment and corruption constraints, the principal may also prefer to legalise the undesirable activity. In contrast, our result does not rely on a commitment problem.

[13] See, a.o., Kofman and Lawarrée (1993), Kessler (2000), Khalil and Lawarrée (2003), Khalil et al. (2007), as well as the classic paper by Tirole (1986).

[14] In our model this corresponds to legalising the undesirable activity.

2 The benchmark model without corruption or extortion

An agent (A) must choose whether or not to undertake an activity (*activity* or *no activity*). If she does not undertake the activity, this results in welfare $\bar{w}$ and a payoff $\varepsilon \cdot (1-\alpha)$ to A, where ε is a positive constant and $1-\alpha$ is A's utility from leisure. α is a random variable drawn from a uniform distribution over the interval $[0,1]$. If she undertakes the activity, this leads to welfare $\underline{w}$ and, if undetected, to a private benefit of π. The activity is socially undesirable in the sense that $\Delta w \equiv \bar{w} - \underline{w} > \pi - \varepsilon \cdot (1-\alpha)$, or equivalently, that A's net benefit from undertaking the activity is smaller than the reduction in social welfare it generates. The activity, however, is beneficial from A's point of view as we assume that $\varepsilon < \pi$, i.e. that A's utility from leisure is always lower (even if $\alpha = 0$) than the private benefit she gets from undertaking the activity.

The principal (P) decides whether or not to legalise the activity. When the activity is declared illegal, then A's decision to undertake the activity is equivalent to her decision to break the law. If P declares the activity illegal, she must rely on a monitor (M) to enforce the law.[15] M has the choice between two actions: *monitor* and *do nothing*. If M chooses *do nothing*, he gets zero. If M chooses *monitor*, he suffers a monitoring disutility of $1 + \varepsilon \cdot \mu$, where ε is a positive constant and $\mu \sim U[0,1]$. If M monitors when A broke the law, then he finds evidence of A's offence with probability one.[16] If M chooses *do nothing*, then A's possible offence remains undetected. M's choice of action is not contractible. If M does not present any evidence of the offence to P, then we assume that P compensates M with a zero transfer.[17] Otherwise, if M presents evidence of the offence, then P transfers reward r to M and punishes A.

A priori P can punish A in many ways: P can use jail sentence, torture and, ulti-

[15] As a convention, we shall assume that P and A are female, while M is male.

[16] It may be more realistic to assume that M only finds evidence of the offence (when A broke the law) with probability $\lambda < 1$. Setting $\lambda = 1$, however, simplifies the exposition without affecting our results.

[17] This is w.l.o.g. because rewarding M for reporting no evidence only decreases his incentives to monitor. Hence, in the absence of any evidence, P should give the lowest possible reward to M which is determined by M's outside option.

mately, execution. We assume, however, that P cannot impose an infinite punishment upon A. Let $\bar{x}$ denote the maximum punishment A faces.[18] We think about $\bar{x}$ as something which is culturally determined and which restricts P's ability to punish A. This is realistic: in Texas the death penalty is regularly carried out, while this is not seen as an acceptable punishment in many European countries. Let $x \in [0, \bar{x}]$ denote A's punishment. Let f_A denote the monetary equivalent of punishment x: that is, A is indifferent between paying a fine f_A and receiving an amount x of negative utils. Similarly, $\bar{f}_A$ denotes the monetary equivalent of punishment $\bar{x}$.

Henceforth, we denote the realisations of α and μ by, respectively, $\tilde{\alpha}$ and $\tilde{\mu}$. In this section we consider the following timing of events:

$t = 0$ P decides whether or not to legalise the activity. If she declares the activity legal, A gets π and welfare equals $\underline{w}$, and the game ends. If P declares the activity illegal, she determines r and f_A.

$t = 1$ A's utility of leisure is realised; $\tilde{\alpha}$ is only known to A. M's disutility from monitoring is realised; $\tilde{\mu}$ is only known to M.

$t = 2$ A decides whether to obey the law or not. M decides whether to monitor or to do nothing.

$t = 3$ If M monitored when A broke the law, he reports this offence to P. M then gets $r - 1 - \varepsilon\tilde{\mu}$ while A gets $-f_A$. If M does not possess any evidence of wrongdoing, he gets (excluding possible monitoring costs) zero, and A's possible offence remains undetected leaving A with a payoff of $\varepsilon(1 - \tilde{\alpha})$ if A chose *no activity* or π if A chose *activity*.

Observe that as μ is realised at time one, P cannot design (r, f_A) such as to learn about M's disutility from monitoring. A strategy for A is a function which maps all possible realisations of her utility from leisure into $\{\textit{activity}, \textit{no activity}\}$. Similarly,

[18] $\bar{x}$ is measured in the amount of (negative) utils the agent gets from undergoing the punishment.

a strategy for M is a function mapping all possible realisations of his disutility from monitoring into $\{monitor, do\ nothing\}$. A is said to follow a monotone strategy if she chooses *activity* if and only if $1 - \tilde{\alpha}$ is less than some critical cut-off value denoted by $1 - \alpha^c$. Similarly, M is said to follow a monotone strategy if he chooses *monitor* if and only if $\tilde{\mu}$ lies below some cut-off value denoted by μ^c.[19] As A's strategy is monotone, it follows that $\Pr(A \text{ chooses } no\ activity) = \Pr(1 - \tilde{\alpha} > 1 - \alpha^c) = \Pr(\tilde{\alpha} < \alpha^c)$. As α is uniformly drawn from $[0, 1]$, $\Pr(\tilde{\alpha} < \alpha^c) = \alpha^c$. Similarly, $\mu^c = \Pr(M \text{ chooses } monitor)$. Each of these cut-off values can also be thought of as a fraction of all players in a given population taking a particular action. For example, in a population of agents that differ slightly in utility derived from leisure, the expected fraction of A's that would choose *no activity* is equal to α^c. (And similarly for μ^c.) It is worth noting that ε in our model captures the amount of heterogeneity among agents and monitors: the higher ε the greater the heterogeneity. In our analysis, however, we shall assume that this heterogeneity is small and we derive all our results for ε close to zero.[20] Nonetheless, we will explain the intuition behind some of our results by analysing the ε-not-close-to-zero case. We believe this will help the reader to understand better our results and also to assess better the robustness of our model. Henceforth, equilibrium values are denoted by an asterisk.

P chooses (r, f_A) to maximise

$$E(w) = \underline{w} + \alpha^* \Delta w - \mu^* + (1 - \alpha^*)[(1 - \mu^*)\pi - \mu^*(r + f_A)],$$

where $E(w)$ denotes expected welfare. Notice that the expected welfare is decreasing in r, which is justified by the observation that in order to pay r, P has to raise money, e.g. through taxation, which entails an inefficiency (due to costly collection of taxes, etc), and the inefficiency is increasing with the size of the transfer. Welfare is also decreasing

[19] It is innocuous to restrict attention to monotone strategies. It can easily be shown that even if A anticipates that M will not follow a monotone strategy, it is a best response for her to choose *activity* if and only if her utility from leisure lies below some critical level. The same remark also applies to M.

[20] We assumed that both α and μ are $\sim U[0, 1]$. As ε is close to zero, however, those assumptions are w.l.o.g. See Fudenberg and Tirole (1991, pp. 233–4) for more details.

in A's punishment. This captures our assumption that P punishes A by sending her to jail (i.e. the loss in A's utility is not compensated by the gain in someone else's).[21] Finally, observe that P takes into account how the penalty/reward structure influences α^* and μ^*.

Proposition 1 *In a corruption- and extortion-free world,*

(i) if $\Delta w < \pi + 2\frac{\pi}{\pi + \bar{f}_A}$, *$P$ declares the activity legal,*

(ii) if $\Delta w > \pi + 2\frac{\pi}{\pi + \bar{f}_A}$, *$P$ declares the activity illegal,* $f_A^* = \bar{f}_A$, *r^* is arbitrarily large, α^* is close to one, μ^* is close to* $\frac{\pi}{\pi + \bar{f}_A}$, *and $E(w)$ is close to* $\bar{w} - 2\mu^*$.

To gain some intuition, suppose ε is not close to zero and that the activity is declared illegal. It can then be easily shown that both μ^* and α^* are increasing in r. This is intuitive: if r increases, M has more incentives to monitor, and thus μ^* goes up. This increased monitoring intensity, however, reduces A's incentives to break the law. Therefore, α^* goes up. Hence, the advantage of increasing r (i.e. more law abiding behaviour) needs to be weighed against an increase in expected monitoring costs. For ε close to zero, however, the increase in μ^* is negligibly small. Hence, P wants to set r at an arbitrarily high level. This enables her to implement an α^* close to one.

It can also be shown (in the ε-not-close-to-zero case) that both μ^* and $1 - \alpha^*$ are decreasing in f_A. This is also intuitive: if f_A increases, it becomes more costly to break the law, and thus A breaks the law less. This increase in law abiding behaviour reduces M's gain from monitoring, and thus decreases μ^*.[22] As welfare is decreasing in μ^* and increasing in α^*, $f_A^* = \bar{f}_A$. As $\bar{f}_A < \infty$, P cannot implement a μ^* below a certain

[21] This assumption is sometimes realistic and is not entirely crucial for our results. It can be shown that, under some additional assumptions, all results presented in this paper remain valid if A were to be punished with a fine instead of a jail sentence. Observe also that as A's punishment represents a social waste, P has ex-post an incentive to forgive any of A's wrongdoings. Hence, we implicitly assume that P does not renegotiate her penalty/reward scheme, which can be justified on the basis that she cares about her reputation.

[22] If ε is close to zero, an increase in f_A hardly influences α^* (while it still reduces μ^*).

threshold level. More specifically, in the proof we show that μ^* is close to $\frac{\pi}{\pi+\bar{f}_A}$. This explains why welfare is strictly lower than $\bar{w}$. In the proof we also show that welfare is close to $\bar{w} - 2\frac{\pi}{\pi+\bar{f}_A}$. If P declares the activity legal, however, welfare is equal to $\underline{w} + \pi$ and the proposition follows.

In essence, Proposition 1 states the obvious: a social planner should declare an activity illegal if the cost to enforce the law is less than the benefit from law abiding behaviour. The non-trivial part of the proposition above stems from the fact that both the cost to enforce the law (i.e. the expected disutility from monitoring) and the amount of law abiding behaviour (i.e. α^*) are endogenous.

3 The model with corruption and extortion

In this section, M must choose one action from $\{$*monitor*, *extort*, *do nothing*$\}$. The possibility of corruption arises if M chooses *monitor* when A broke the law: then, having found the evidence of A's offence, M can choose to accept a bribe from A in return for hiding that evidence. The possibility of extortion arises if, instead of gathering the evidence of A's offence, M chooses to concoct fake evidence with which to incriminate A. Thus, extortion and monitoring are modeled as substitutes. To motivate this choice, consider our example of extortion in Lahore (see Introduction). If the police officer decides to extort, he must first wait until the tourist is out of her hotel, plant the evidence in the hotel room, wait for the tourist to return, bargain with the tourist over a possible bribe, bargain with the hotel room owner over how the tourist's bribe should be divided, etc. All these activities take time. The police officer could have used that time to find genuine proof of the tourist breaking the law.[23]

We also assume that monitoring is more costly (in terms of effort) than extorting.

[23] Actually, our results rest on the idea that if M makes "a lot" of money through extortion, this reduces his incentives to monitor. We model this by assuming that *exort* and *monitor* are actions which take the same time to complete (which, as argued above, is sometimes realistic). Alternatively, one could come up with a model in which M, having already earned "a lot" of money by framing A, prefers to spend the rest of his time with his family instead of performing any monitoring activity.

We defend this assumption on the grounds that finding proof of misconduct can be a lengthy, difficult and even dangerous job. For example, it may involve laboratory tests (such as investigating finger prints, blood samples, etc), interrogations, tapping phone conversations, shadowing people, and such like. Obviously, extortionary activities also require effort, but one would not expect them to be as costly as the activities required to uncover proof of misconduct. Let $d_{monitor}$ and d_{extort} denote the disutility of, respectively, monitoring and extorting. In this paper, we assume that $d_{monitor} = 1+\varepsilon\tilde{\mu}$ and $d_{extort} = 0$; however, we expect our main result to hold as long as $d_{monitor} > d_{extort} \geq 0$.

If M chooses *extort*, then A gets framed with probability one.[24] Suppose M sends the fake evidence to P. P does not know whether the evidence is genuine or fake. Therefore, whenever she receives a report (or a piece of "evidence") incriminating A, she asks the opinion of a second monitor. The second monitor rapidly evaluates M's report and incurs no disutility of effort. We assume that

$$\Pr(\text{second monitor finds that evidence is fake}) = \begin{cases} p & \text{if evidence is fake,} \\ 0 & \text{if evidence is genuine.} \end{cases} \tag{1}$$

The second monitor truthfully reports his findings to P.[25] If P finds out that the evidence is fake, she imposes punishment y on M. Similar to the discussion above for the feasible range of the punishment imposed on A, we realistically assume that $y \in [0, \bar{y}]$, with f_M and $\bar{f}_M$ representing the monetary equivalents of y and $\bar{y}$. If the second monitor did not find out that the evidence is fake, P imposes penalty f_A on A and gives r to M.

We now know enough to explain expected payoffs when M chooses *extort*. To fix

[24] This assumption implicitly relies on our "one monitor–one agent" set-up. Alternatively, we could have assumed that our economy consists of N agents and one monitor. A would then be framed with probability $\frac{1}{N}$ whenever M chooses *extort*. This alternative modeling strategy, however, should not qualitatively affect our results.

[25] This can best be justified on the grounds that the second monitor is a "bureaucrat" who never had any contact with A or M in the past, nor will he have any during the course of his "investigation". E.g., in the context of our Lahore example, the second monitor works in an office in Islamabad.

ideas, consider our leading example. Suppose the tourist leaves her hotel room at 1 p.m. At 1.05 p.m. the police officer puts the drugs in the tourist's hotel room and waits for the tourist to return. Suppose the tourist returns at 4 p.m. and that, between 1 p.m. and 4 p.m., she did not break the law. The tourist and the police officer then bargain over a possible bribe. Let B_f denote the bribe M gets when M extorts A. If the tourist successfully bribes the police officer, they get (at 4.05 p.m.) $-B_f + \varepsilon(1-\tilde{\alpha})$ and B_f, respectively. Their joint payoffs then add up to $\varepsilon(1-\tilde{\alpha})$. If the tourist does not succeed to bribe the police officer, M sends the fake evidence to P, the police officer gets $(1-p)r - pf_M$ while the tourist gets $-(1-p)f_A + \varepsilon(1-\tilde{\alpha})$. Their expected joint payoffs then add up to $(1-p)(r-f_A) - pf_M + \varepsilon(1-\tilde{\alpha})$. Let $\beta \in [0,1]$ (respectively $1-\beta$) denote the bargaining power of A (respectively M). A and M will agree to conceal the fake evidence whenever there is a non-negative excess surplus from concealing it as opposed to declaring it,[26] namely $\varepsilon(1-\tilde{\alpha}) \geq (1-p)(r-f_A) - pf_M + \varepsilon(1-\tilde{\alpha})$ $\Leftrightarrow$ $(1-p)(f_A - r) + pf_M \geq 0$, with A paying $1-\beta$ of this surplus to M (on top of $(1-p)r - pf_M$) as the bribe. Suppose now that the tourist bought cannabis at 2 p.m. We then assume that the tourist smokes her cannabis immediately.[27] Hence, at 4 p.m. π has already been consumed and does not affect the size of the bribe. This implies that M's expected gain from extorting is independent of A's action and, as above, A and M will agree to conceal the fake evidence if $(1-p)(f_A - r) + pf_M \geq 0$. Summing up this discussion, the extortion surplus (ES) is

$$ES \equiv \max\{0,\ (1-p)(f_A - r) + pf_M\},$$

$$\text{if} \quad (1-p)(f_A - r) + pf_M \geq 0, \quad \text{then} \quad B_f = (1-p)r - pf_M + (1-\beta)ES.$$

If M chooses *monitor*, then detection of A's offence (if committed) is certain, as before. If A abides by the law, she gets $\varepsilon(1-\tilde{\alpha})$ while M gets $-1-\varepsilon\tilde{\mu}$. More interestingly,

[26] We assume w.l.o.g. that if M is indifferent between accepting and not accepting the bribe, he accepts the bribe (and A does not go to jail).

[27] A knows that a police officer might be waiting for her in her hotel room at 4 p.m. If there is a small probability that she will be subject to a body search, she will prefer to smoke her cannabis immediately.

suppose M monitors while A breaks the law. A priori, M can prove A's guilt in two different ways. First, M can "catch" A while she is still in possession of the drug (i.e. before she had the time to enjoy π). Second, M can find out that A smoked cannabis via a blood test (i.e. after she enjoyed π). In this paper we assume that M, with probability 1, catches A before she had the time to enjoy π.[28] Let B_b denote the bribe M could extract from A if he catches A breaking the law. Observe that if M catches A before she had the time to enjoy π, the police officer can say: "If I report this offence to P, you get $-f_A$ *and* you cannot enjoy your cannabis". Thus, if A and M agree to conceal the evidence, A gets $\pi - B_b$ while M gets $B_b - 1 - \varepsilon\tilde{\mu}$, and their joint payoffs add up to $\pi - 1 - \varepsilon\tilde{\mu}$. Otherwise, if M gives the evidence to P, A gets $-f_A$ while M gets $r - 1 - \varepsilon\tilde{\mu}$ and their joint payoffs add up to $r - f_A - 1 - \varepsilon\tilde{\mu}$. As above, A and M will agree to conceal the evidence whenever there is a non-negative excess surplus from concealing the evidence as opposed to declaring it, namely $\pi - 1 - \varepsilon\tilde{\mu} \geq r - f_A - 1 - \varepsilon\tilde{\mu} \Leftrightarrow \pi + f_A - r \geq 0$, with A paying $1 - \beta$ of this surplus to M (on top of r) as the bribe. Analogously with the above, the corruption surplus (CS) is

$$CS \equiv \max\{0,\ \pi + f_A - r\},$$

$$\text{if} \quad \pi + f_A - r \geq 0, \quad \text{then} \quad B_b = r + (1-\beta)(\pi + f_A - r). \tag{2}$$

The discussion above allows us to state expected payoffs for M and A.

$$M \text{ gets} \begin{cases} (1-\alpha)[r + (1-\beta)CS] - 1 - \varepsilon\tilde{\mu} & \text{if} \quad \textit{monitor}, \\ (1-p)r - pf_M + (1-\beta)ES & \text{if} \quad \textit{extort}, \\ 0 & \text{if} \quad \textit{do nothing}. \end{cases} \tag{3}$$

Note from (3) that when

$$(1-p)r - pf_M + (1-\beta)ES > 0 \tag{4}$$

[28] It would be more reasonable to assume that with probability q M catches A before she had the time to enjoy π, while with probability $1-q$ M catches A after A enjoyed her private benefit. This generalization, however, complicates the exposition and should not qualitatively affect our results.

then *do nothing* is the least attractive action in terms of the expected gain, and hence it will not be chosen in equilibrium. Consequently, when (4) holds, then $1-\mu$ captures the probability of extortion. Alternatively, when (4) is violated, then *extort* results in the least attractive gain, which implies that in this case $1-\mu$ captures the probability of no monitoring. This condition also determines the payoffs to A, as follows:

If (4) holds, then

$$A \text{ gets} \begin{cases} \mu(-f_A + \beta CS) + (1-\mu)[\pi - (1-p)f_A + \beta ES] & \text{if } \textit{activity}, \\ (1-\mu)[-(1-p)f_A + \beta ES] + \varepsilon(1-\tilde{\alpha}) & \text{if } \textit{no activity}. \end{cases} \tag{5}$$

If (4) does not hold, then

$$A \text{ gets} \begin{cases} \mu(-f_A + \beta CS) + (1-\mu)\pi & \text{if } \textit{activity}, \\ \varepsilon(1-\tilde{\alpha}) & \text{if } \textit{no activity}. \end{cases}$$

In the remainder of this section we maintain the following assumptions.

Assumption 1 $$p\bar{f}_M < (1-p)\pi \Leftrightarrow p < \tfrac{\pi}{\pi+\bar{f}_M} \tag{A1}$$

(A1) states that if M sends fake evidence to P, he knows that it is very unlikely he will end up in jail for that. We will shortly see that, in equilibrium, M extorts. Perhaps this result would also hold under a weaker version of (A1).[29] However, to explain extortion p cannot be too high; if it were, M would never dare to send fake evidence to P and this would eliminate his incentives to extort in the first place. Recall that we assumed that the second monitor is honest (i.e. does not collude with either M or A). We justified that assumption on the basis that the second monitor never meets (and has never met) A or M. In that sense, the second monitor's investigation is a very superficial one. Hence, one would expect p to be low.

Assumption 2 $$\bar{f}_A > \beta\Big(\pi - \tfrac{p}{1-p}\bar{f}_M\Big) \tag{A2}$$

[29] The results of Bolton (1987) indicate that if M were risk-averse and if Pr(second monitor finds that evidence is fake| evidence is genuine) > 0, it might be optimal to set $f_M^* < \bar{f}_M$. We conjecture that our results will then hold under the following (weaker) assumption: $p < \frac{\pi}{\pi+f_M^*}$.

(A2) ensures that the maximal penalty imposed on A for breaking the law cannot be too small, which is clearly an innocuous assumption.

ASSUMPTION 3 $$\beta p \bar{f}_M + (1-\beta)(1-p)\pi > 1 \tag{A3}$$

Observe that (A3) is trivially satisfied if $\beta p \bar{f}_M > 1$. If $\beta p \bar{f}_M < 1$, the assumption states that π cannot be "low". The purpose of this assumption will be explained below.

P maximizes welfare taking into account how (r, f_A, f_M) influences (i) incentives for A and M to conceal any "evidence" of wrongdoing (and, thus, whether or not she will send someone to jail and whether or not she will have to raise taxes to reward M), (ii) incentives for M to either extort, or monitor, or do nothing, and (iii) incentives for A to either break or obey the law. Our main result is summarised below (in the proposition below $1 < \kappa < \kappa'$).

Proposition 2 *There exists a unique equilibrium outcome in which:*

(i) if $\Delta w \in (\pi, \kappa\pi)$, then P legalises the activity;

(ii) if $\Delta w \in (\kappa\pi, \kappa'\pi)$, then P declares the activity illegal, $f_A^* = \bar{f}_A$, $f_M^* = \bar{f}_M$ and $r^* = \frac{p}{1-p}\bar{f}_M + \bar{f}_A$. $\alpha^* \in (p, 1)$, while M monitors with probability $\mu^* \in (0,1)$ and extorts with probability $1-\mu^*$. A bribes M whenever she is caught breaking the law and when she is framed;

(iii) if $\Delta w > \kappa'\pi$, then P declares the activity illegal, (f_A^*, r^*) are set such that $\beta r^* + (1-\beta)f_A^* = \beta\pi$ and $f_M^* = \bar{f}_M$. $\alpha^* \in (p, 1)$, while $\mu^* = 1$. A bribes M whenever she is caught breaking the law.

P has three good reasons for not wanting to send anyone to jail. First, note that A can be punished in two ways: either she pays a bribe to M, or she spends a certain amount of time in jail. Punishment by means of a bribe merely represents a transfer of money between two risk-neutral agents and is therefore welfare neutral. Punishment by imprisonment, however, represents a loss in A's utility which is not compensated by any increase in M's utility. Second, P prefers A to reward M (via a bribe) whenever M

uncovers proof of misconduct rather than to raise taxes to pay r to M. Third, if P sets (r, f_A, f_M) such that A always bribes M, welfare increases by π each time A breaks the law, while if she sets (r, f_A, f_M) such that A never bribes M, welfare increases by π only if A breaks the law *and* if M did not monitor.

It follows from our previous paragraph that we can, without loss of generality, continue to explain the intuition behind our results under the assumption that A always bribes M. In our model, M does not care whether $B_b = 10$ because r is "high" and f_A "low", or because f_A is "high" and r "low". Instead, M is only interested in the total size of the bribe. An identical remark applies to B_f. Therefore, the intuition behind Prop. 2 is best understood using B_b and B_f. Note also that to reduce M's incentives to extort, P sets $f_M^* = \bar{f}_M$. It is then straightforward to check that, in equilibrium,

$$B_f = (1-p)B_b - K, \tag{6}$$

where $K = (1-\beta)(1-p)\pi + \beta p \bar{f}_M$. In particular, (6) shows that P cannot increase M's gain from monitoring (i.e. increasing B_b), without increasing his gain from extorting. Furthermore, A only chooses *no activity* if she is punished whenever M caught her breaking the law. This implies that B_b may not be less than π. Hence,

$$B_f \geq (1-p)\pi - (1-\beta)(1-p)\pi - \beta p \bar{f}_M = \beta\left((1-p)\pi - p\bar{f}_M\right) > 0,$$

where the strict inequality follows from (A1). Therefore, for all his possible monitoring costs, M prefers *extort* over *do nothing*.

There does not exist an equilibrium in which A always[30] obeys the law. By contradiction, suppose such an equilibrium exists. Then, M's gain from monitoring $= -1 - \varepsilon\tilde{\mu} < 0 < B_f =$ his gain from extorting. Anticipating that M always extorts, it is, however, not a best reply for A to obey the law. Similarly, there does not exist an equilibrium in which A always strictly prefers to break the law. By contradiction, suppose such an equilibrium exists. Then, M's gain from monitoring $= B_b - 1 - \tilde{\mu}\varepsilon$. It follows from (6) that he prefers to monitor if and only if $B_b - 1 - \tilde{\mu}\varepsilon > (1-p)B_b - K \Leftrightarrow K - 1 - \tilde{\mu}\varepsilon > -pB_b$, which, for

[30] With "always" we mean "for all his or her possible types".

ε close to zero and under (A3), is satisfied. Anticipating that M always monitors, it is easy to check that A cannot strictly prefer to break the law. Hence, in equilibrium there must exist a type of agent who is indifferent between *activity* and *no activity*. Using a similar reasoning, one can check that in equilibrium there must also exist a type of monitor who is indifferent between *monitor* and *extort*.

It follows from above that monitor μ^* is indifferent between *monitor* and *extort* if

$$(1-\alpha)B_b - 1 - \varepsilon\mu^* = (1-p)B_b - K. \tag{7}$$

The left-hand side of the above equation represents monitor μ^*'s gain from monitoring, while the right-hand side represents his gain from extorting. Observe that (7) implicitly defines the identity of the indifferent monitor (i.e. μ^*) as a function of (the exogenous) $1-\alpha$. Agent $1-\alpha^*$ is indifferent between breaking and obeying the law if

$$\mu(\pi - B_b) + (1-\mu)(\pi - B_f) = -(1-\mu)B_f + (1-\alpha^*)\varepsilon. \tag{8}$$

The left-hand side of (8) represents agent $1-\alpha^*$'s gain from breaking the law, while the right-hand side represents her gain from obeying the law. As above, (8) implicitly defines the identity of the indifferent agent (i.e. $1-\alpha^*$) as a function of (the exogenous) μ. Equations (7) and (8) can respectively be rewritten as,

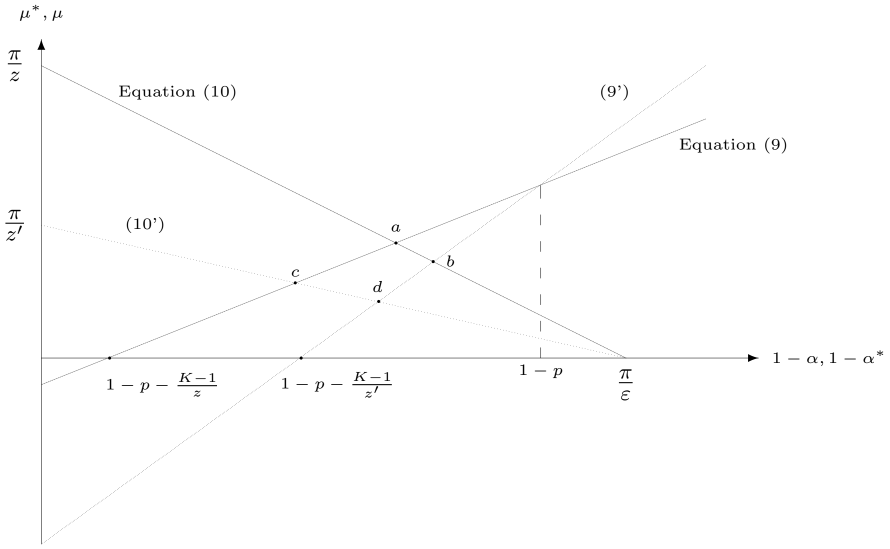
$$\mu^* = \frac{1}{\varepsilon}[K - 1 - (1-p)B_b + (1-\alpha)B_b] \qquad \text{and} \tag{9}$$

$$\mu = \frac{1}{B_b}[\pi - (1-\alpha^*)\varepsilon] \tag{10}$$

It follows from (9) that μ^* is increasing in $1-\alpha$. This is intuitive: If A is more likely to break the law, it becomes more profitable to monitor and the indifferent monitor will possess a higher cost of monitoring. Similarly, (10) reveals that $1-\alpha^*$ is decreasing in μ. This is also intuitive: if M is more likely to monitor, breaking the law becomes more costly and the indifferent agent will possess a lower utility of leisure. $(\mu^*, 1-\alpha^*)$ is determined where (9) and (10) meet. It can be checked that $1-\alpha^*$ and μ^* are, respectively, decreasing and increasing in p: the higher p, the lower M's gain from extorting and the higher the monitoring cost of the indifferent M. This increase in μ^* induces A to break the law less often. In the Appendix we even show that $1-\alpha^* < 1-p$.

To understand how $(\mu^*, 1-\alpha^*)$ is affected by a change in B_b, suppose that ε is not close to zero, and consider Fig. 1.

Figure 1: An increase in B_b and its effect on $1-\alpha^*$ and μ^*.



Suppose P changes the penalty/reward structure such that B_b increases from z to z'. In the figure the lines called "Equation (9)" and "Equation (10)" represent, respectively, (9) and (10) when $B_b = z$. Similarly, "(9')" and "(10')" represent, respectively, (9) and (10) when $B_b = z'$. It transpires from Fig. 1 that an increase in B_b influences μ^* through both a direct and an indirect channel. For a graphical representation of this direct channel, suppose that the change in B_b leads to a rotation of "Equation (9)" but leaves "Equation (10)" unaffected. The equilibrium would then move from a to b which entails a lower μ^*. To understand this, consider (7), which embodies our assumption that if M monitors he gets B_b only if A breaks the law. If M extorts, he gets B_b with probability $(1-p)$ (minus some constant). We have noted above that $1-\alpha^* < 1-p$. As A is not "very likely" to break the law, M's gain from extorting increases faster (in B_b) than his gain from monitoring. Hence, the indifferent monitor now possesses a lower monitoring cost. To understand the intuition behind the indirect channel, suppose the increase in B_b leads to a rotation of "Equation (10)" but leaves "Equation (9)" unaffected. The equilibrium would then move from a to c which also entails a lower monitoring intensity. In c the agent is also more likely to obey the law. This is intuitive: As B_b increased, A has to pay a bigger bribe to avoid prison. This reduces her gain from breaking the law, and thus decreases $1-\alpha^*$. The decrease in $1-\alpha^*$, however, reduces M's gain from monitoring. Hence, both the direct and the indirect channel lead to a negative relation between μ^* and B_b. This results also holds for ε close to zero.

Any increase in B_b also influences $1-\alpha^*$ via a direct and an indirect channel. The direct channel of A is identical to the indirect one of M: as A has to pay a bigger bribe to avoid prison, $1-\alpha^*$ decreases (from a to c). The indirect channel of A is identical to the direct one of M: as M has less incentives to monitor, this makes it more profitable for her to break the law, and thus $1-\alpha^*$ increases (from a to b). Hence, an increase in B_b has an ambiguous effect on $1-\alpha^*$ (which explains why point d lies almost vertically under point a). Observe that the strength of A's direct and indirect channels depends on the slopes of (9) and (10). In particular, if ε is close to zero, (9) is almost a vertical line while (10) is almost a horizontal one. Point c then lies almost vertically under point a. Economically speaking this means that, when ε is close to zero, A's direct channel vanishes and thus $1-\alpha^*$ is *increasing* in B_b.[31]

Hence, if P changes (r, f_A, f_M) such as to reduce B_b, this influences welfare in two ways. On the one hand, it reduces the probability that A will break the law, and thus that welfare will increase by $\bar{w}-(\underline{w}+\pi)$. On the other hand, it increases the probability that welfare will decrease by one (i.e. by the disutility of monitoring). The first effect dominates the second one if Δw is "high", namely if $\Delta w > \kappa'\pi$. This is intuitive: if Δw is high, law abiding behaviour is set to increase welfare by a large amount and compensate for any disutility incurred by M. If $\Delta w > \kappa'\pi$, welfare is thus maximised by setting B_b at its lowest possible level. P cannot set $B_b = 0$, as she must ensure that A does not

[31] This discussion also highlights that our positive relation between $1-\alpha^*$ and B_b does *not* rest on our assumption that ε is close to zero. As long as ε is below a critical value (and that "critical value" can potentially be "very high"), A's indirect channel continues to dominate her direct one and the positive relationship between $1-\alpha^*$ and B_b would still hold.

get a positive payoff whenever she is caught breaking the law. Therefore, in case (*iii*) of Prop. 2 P chooses a penalty/reward scheme such that $B_b = \pi \Leftrightarrow \beta r^* + (1-\beta)f_A^* = \beta\pi$. The second effect dominates the first one if $\Delta w \in (\pi, \kappa'\pi)$. When the activity is illegal, P will thus select (r, f_A, f_M) which implements the lowest possible monitoring intensity. From our explanations following Fig. 1 this means that B_b must be set at its highest possible level. As B_b is increasing in f_A, $f_A^* = \bar{f}_A$. P cannot set r at an arbitrarily high level as A will then not be able to bribe M after having been framed. Therefore, P sets r such that, in equilibrium, $ES = 0$, which is equivalent to $r^* = \frac{p}{1-p}\bar{f}_M + \bar{f}_A$. This penalty/reward scheme implements a "very low" monitoring intensity. We have argued above that $1-\alpha^*$ is bounded above by $1-p$. If there were no second monitor, $1-\alpha^*$ would be bounded above by one and P would prefer to legalise the activity whenever $\Delta w \in (\pi, \kappa'\pi)$. The second monitor, however, makes it easier for P to implement a higher α for any given μ.[32] This implies that even though μ^* is "very low", α^* is only "low" (and not "very low"). Therefore, if Δw is "intermediate" (i.e. if $\Delta w \in (\kappa\pi, \kappa'\pi)$), P is better off by declaring the activity illegal, i.e. the "low" amount of law abiding behaviour compensates for M's "very low" monitoring intensity. If $\Delta w \in (\pi, \kappa\pi)$, the contrary situation prevails and P prefers to declare the activity legal.

We are left to explain the usefulness of (A3). In our model, extorting has one advantage over monitoring in the sense that it is less costly in terms of effort. (6), however, reveals that monitoring has one advantage over extorting, namely $B_b > B_f$. There are two reasons which explain this finding. First, if M catches an offending agent, he can—by reporting this to P—prevent A from enjoying π. A knows this and is therefore willing to pay a higher bribe than if she had been framed. Second, if M extorts he knows that he might end up in jail if he sends the "evidence" to P. In contrast, if M catches A while she broke the law, M knows that he will never end up in jail due to (1).[33] Observe

[32] Equation (26) in the Appendix states the equilibrium relationship between α and μ.

[33] One can come up with other reasons why, in practice, B_b should be higher than B_f. For example, if M chooses *monitor*, he could catch real criminals who, on average, should be richer (and thus more able to pay large bribes) than innocent people. Alternatively, it is much easier to fabricate "evidence" which

also that $B_b - B_f$ is increasing in π. If π is "low" (i.e. if (A3) is not satisfied when $\beta p \bar{f}_M < 1$), the difference between the two bribes is not enough to compensate M for his disutility from monitoring. This implies that M, irrespective of $\tilde{\mu}$ and of $1 - \alpha$, prefers to extort. As M never monitors, A always breaks the law.[34] As M never monitors and as A never obeys the law, P is indifferent between legalising the activity and declaring it illegal. Therefore, our model only yields "rich" insights once (A3) is satisfied.

So far we assumed that agents differ in their utility from leisure. Suppose now that agents were to differ solely in their attitudes towards risk.[35] Suppose also that, after M reported evidence of wrongdoing to P, A faces a trial with a random outcome (i.e. A might spend little time in jail or she may face a lengthy jail sentence). In such a model one would expect the more risk-averse agents to obey the law and the less risk-averse ones to break it. Hence, a framed person is more likely to be more risk-averse than a truly guilty one. This also implies that a framed agent is (on average) very keen to avoid the court room, and, thus, that B_f will be relatively high. Mirroring this argument, B_b will be relatively low. Hence, comparing the outcome we get when agents have different risk preferences with the one we get when all agents are risk-neutral, we expect $B_b - B_f$ to be lower in the former case. Hence, M has less incentives to monitor in the case with different risk preferences. As M monitors less, we expect A to obey the law less in that

"proves" that someone possessed drugs and attempted to sell them rather than to fabricate "evidence" which "proves" that someone actually sold them. Hence, if M chooses *extort*, he can "only" charge A with possession and attempted sale of illegal drugs. If M chooses *monitor*, however, he might uncover evidence that A actually sold drugs. Selling illegal drugs, however, is a serious crime which carries a much heavier penalty than attempted sale and possession. Hence, in this example, on average B_b will be higher than B_f.

[34] This reasoning is similar to the one in Polinsky and Shavell (1989). In that paper the authors show that court errors may increase the probability of litigation. In that case the agent realises that even if she obeys the law, she will end up in court (and this dulls her incentives to obey the law). In our model, due to extortion, the agent knows that even if she obeys the law, she will be framed (and this removes her incentives to obey the law).

[35] The idea that guilty defendants and innocent defendants have different risk preferences is well studied in the plea bargaining literature: see for example Grossman and Katz (1983).

case. As it becomes more difficult to induce A into law abiding behaviour, P should be more willing to legalise the activity when agents have different risk preferences.

4 Conclusions and policy implications

In this paper we focused on the class of crimes for which the police force can easily (and with little risk of being caught) frame innocent individuals. We believe that victimless crimes taking place in less developed countries are more likely to fit this definition. To understand this, consider the following three examples. Suppose a police officer in Lahore reports that he caught Mrs. Smith in possession of cannabis. How is Mrs. Smith going to prove that she is innocent? She can say that it was the police officer who put the drugs in her hotel room, but does she not have an incentive to say that anyway? Similarly, Manuela (a central American immigrant in Guatemala) can also claim that she did not lose her travel permit (or that she did not enter the country illegally) and that it was the police officer who stole her travel permit, but doesn't she have an incentive to say that anyway? Suppose now that a police officer indicts Mr. Jones on the (false) accusation that he raped Miss White. To fabricate this type of evidence, the police officer needs Miss White's participation. Suppose the judge, after hearing Mr. Jones's claim that he was framed, orders for an independent investigation into the facts. In principle, Miss White could then be interrogated for a potentially long time. Moreover, the independent investigator can start his interrogation by telling Miss White: "If you now confess that your original testimony is wrong, we will "forgive" the fact that you lied to us and we will only prosecute the police officer..." The first two examples above are taken from less developed countries in which it is very unlikely that the police officer's actions will ever be scrutinized by an independent investigator. Moreover, even if an independent investigation takes place, at the end of the day it will be the police officer's word against Mrs. Smith's (or against Manuela's). Our third example highlights the difficulties of fabricating fake evidence of a crime which involves a victim.

Call the "laissez-faire region", the range of Δw for which it is optimal to legalise the activity. We know from Propositions 1 and 2 that the laissez-faire region in the benchmark case is equal to $\left(\pi, \pi + 2\frac{\pi}{\pi+\bar{f}_A}\right)$, while it is equal to $(\pi, \kappa\pi)$ in Section 3.

Proposition 3 *If $\pi - 2 \leq \frac{\bar{f}_M - \pi}{\bar{f}_M + \pi}\bar{f}_A$ the laissez-faire region with a corrupt and extorting M is larger than the one if M were honest. $(\pi, \kappa\pi)$ is decreasing in π and can be arbitrarily large.*

One would expect the maximal punishments on A and M to be quite high in comparison to π. Hence, the condition stated in the first sentence of the Proposition is a very mild one.[36] Suppose P declares the activity illegal and that $\Delta w \in (\pi, \kappa'\pi)$. We know from our previous section that $B_b - B_f$ is increasing in π. Hence, the higher π, the higher M's incentives to monitor (and the higher μ^*). This increase in monitoring activities induces A to obey the law more often, which, in turn, increases P's incentives to declare the activity illegal. It is also straightforward to check that if p is close to zero and if π is close to its lower bound (stated in (A3)), then the laissez-faire region in Section 3 becomes arbitrarily large. In that sense, our model highlights the importance of π: if π is not "high", and if the police force can easily (and with little risk of being caught) frame innocent individuals, then policy makers should seriously consider legalising the undesirable activity.

Perhaps more surprisingly, our model also shows that a social planner should sometimes allow extortion to take place. In essence, if $\Delta w \in (\kappa\pi, \kappa'\pi)$ a social planner knows that if she declares the activity illegal, the police force will abuse the law to extort money from innocent people. She also knows, however, that the police force will not always extort and will sometimes punish real offenders, thereby increasing law abiding behaviour (and welfare).

Proposition 4 *If $\beta\bar{f}_M > (1-\beta)\pi$, welfare is non-decreasing in p.*

[36] It can be shown that if $\bar{f}_A$ is close to zero and p close to $\frac{\pi}{\pi+\bar{f}_M}$ and $\pi > 2$, the laissez-faire region with an honest monitor is larger than with a corrupt and extorting one. This result, however, relies on the fact that some of our exogenous variables take unrealistic values.

As above, one would expect M's maximal punishment to be quite high in comparison with A's private benefit from breaking the law. Hence, one would expect the (sufficient) condition stated in the proposition to hold, unless A's bargaining power were very low. The result stated in the proposition is intuitive and is also present in a.o. Kofman and Lawarrée (1993) and Carrillo (2000a, 2000b). Similarly, Klitgaard (1983) also stresses the importance of establishing an effective anti-corruption unit.

Another observation relates to the size of rewards and punishments in the presence of corruption and extortion. As follows from our analysis in Section 3, the gain from extorting increases faster in f_A and r than the gain from monitoring.[37] Hence, in the presence of corruption and extortion it may be optimal to give low rewards to M and to impose low punishments on A.[38]

It is sometimes argued that governments should legalise possession and consumption of recreational drugs such as cannabis while continuing to prosecute those who sell it. This policy recommendation is typically defended on the grounds that the cost of catching one consumer of cannabis outweighs any benefit to society. Our model also supports this policy recommendation albeit for a different reason.[39] As mentioned in footnote 33, it is much easier to fabricate “evidence” which “proves” that someone possessed drugs rather than to fabricate “evidence” which “proves” that someone is a drugs dealer. Hence, our model does not apply to the latter type of crime. Furthermore, in comparison to the cost of monitoring, the private benefit associated with consuming cannabis is not very high.

[37] More specifically, the right-hand side of (7) increases faster in B_b than its left-hand side.

[38] Observe, however, that if P found out that M created fake evidence of wrongdoing, she should punish him as hard as she can. The optimality of “low” rewards and punishments has also been noted by a.o. Kofman and Lawarrée (1993), Polinsky and Shavell (2001) and Kugler et al. (2005).

[39] We are grateful to one of our referees for pointing this out to us.

Appendix

Proof of Proposition 1

Suppose ε is close to zero. There can not exist an equilibrium in which $\alpha = 1$. For if $\alpha = 1$, it is a best reply for M to set $\mu = 0$. Anticipating that $\mu = 0$, it is, however, a best reply for A to set $\alpha = 0$. Hence, if there exists an equilibrium in which $\alpha^* > 0$, $\alpha^* < 1$. Agent $1 - \alpha^*$ is indifferent between undertaking the activity or not if $(1-\mu)\pi - \mu f_A = 0$ $\Leftrightarrow \mu = \frac{\pi}{\pi + f_A}$.

There cannot exist an equilibrium in which M strictly prefers to monitor (i.e. $(1-\alpha)r - 1 > 0$), as P can gain by reducing r. There cannot exist an equilibrium in which $\alpha^* > 0$ and in which $\mu^* = 0$ as A can profitably deviate by setting $\alpha = 0$. Hence, if there exists an equilibrium in which $\alpha^* > 0$, M must be indifferent between *monitor* and *no monitor*, i.e. $(1-\alpha)r - 1 = 0 \Leftrightarrow 1 - \alpha = \frac{1}{r}$.

P's problem is then stated as follows:

$$\max_{r, f_A} \qquad \underline{w} + \alpha^* \Delta w - \mu^* + (1-\alpha^*)[(1-\mu^*)\pi - \mu^*(r + f_A)] \tag{11}$$

$$\text{s.t. } \mu^* = \frac{\pi}{\pi + f_A}; \quad 1 - \alpha^* = \frac{1}{r}; \quad r \geq 0,\ f_A \geq 0; \quad \alpha^*,\ \mu^* \in [0,1]$$

Observe that M always has the possibility to do nothing which yields him the same payoff as the one he gets if he does not work for P. Hence, M's participation constraint is trivially satisfied. Substitution of μ^* and $1-\alpha^*$ into (11) yields $\max_{\{f_A, r\}} \bar{w} - \frac{\Delta w}{r} - \frac{2\pi}{\pi + f_A}$. Hence, for $\varepsilon \to 0$, the objective function is maximised when $f_A = \bar{f}_A$ and when r is arbitrarily large. Then, welfare is close to $\bar{w} - 2\frac{\pi}{\pi + \bar{f}_A}$. If the latter is greater than $\underline{w} + \pi$, then P optimally declares the activity illegal. Otherwise, P legalises the activity. ■

Proofs of results in section 3

As seen from the discussion in Section 3, it is convenient to distinguish between the following two cases, depending on whether (3) holds or not. Case I: (3) holds; if $\tilde{\mu} > \mu^*$, then M strictly prefers *extort*, while if $\tilde{\mu} < \mu^*$, M strictly prefers *monitor*. Case II: (3)

does not hold; if $\tilde{\mu} > \mu^*$, then M strictly prefers *do nothing*, while if $\tilde{\mu} < \mu^*$, then M strictly prefers *monitor*.

DEFINITION 1 *A is said to be indifferent if there exists a type of A who is indifferent between activity and no activity. M is said to be indifferent if there exists a type of M who is indifferent between any two of his three actions. Similarly, A (M) is said to strictly prefer one of her (his) actions, if all her (his) possible types have a strict preference for that action.*

We first assume that A and M are indifferent between two actions and compute equilibrium behaviour. Next (in Lemma 8) we show that this assumption is without loss of generality. Let $\mathcal{I}_C = 1$ when $\pi + f_A - r \geq 0$ and let $\mathcal{I}_C = 0$ otherwise. Similarly, let $\mathcal{I}_E = 1$ when $(1-p)(f_A - r) + pf_M \geq 0$ and let $\mathcal{I}_E = 0$ otherwise. P's maximization problem in Case I boils down to

$$\begin{aligned}
\max_{r,f_A,f_M} \quad & \underline{w} + \alpha^*\Delta w - \mu^* + (1-\alpha^*)[(1-\mu^*) + \mu^*\mathcal{I}_C]\pi - pf_M(1-\mu^*)(1-\mathcal{I}_E) \\
& -(f_A + r)[(1-\alpha^*)\mu^*(1-\mathcal{I}_C) + (1-\mu^*)(1-p)(1-\mathcal{I}_E)] \\
s.t. \quad & (1-p)r - pf_M + (1-\beta)ES > 0 \qquad (12) \\
& 1-\alpha^* = \tfrac{(1-p)r - pf_M + (1-\beta)ES + 1}{r + (1-\beta)CS} \qquad (13) \\
& \mu^* = \tfrac{\pi}{\pi + f_A - \beta CS} \qquad (14) \\
& (1-p)r - pf_M + (1-\beta)ES + 1 \leq r + (1-\beta)CS \qquad (15) \\
& \beta CS \leq f_A \qquad (16) \\
& r \geq 0; \quad f_A \in [0, \bar{f}_A]; \quad f_M \in [0, \bar{f}_M] \qquad (17)
\end{aligned}$$

(12) re-states (3). Constraint (13) represents, for ε close to zero, the condition which ensures that monitor μ^* is indifferent between *monitor* and *extort*. Constraint (14) represents, for ε close to zero, the condition which ensures that agent $1-\alpha^*$ is indifferent between *activity* and *no activity*. Constraints (15) and (16) ensure that α^* and μ^*, as stipulated by (13) and (14), fall into $[0, 1]$. (17) ensure that the reward is non-negative and the punishments are bounded. Finally, as M always has the possibility to do nothing, his IR constraint is not included in the above maximisation problem.

For ease of reference, let v_e denote the LHS of (3). Using this notation, we can identify the following parameter ranges generated by the two programmes above: $\Omega_1 = \{(r, f_A, f_M)|\ \pi + f_A - r \geq 0$ and $v_e \leq 0\}$, $\Omega_2 = \{(r, f_A, f_M)|\ \pi + f_A - r \geq 0$ and $v_e > 0$ and $(1-p)(f_A - r) + pf_M \geq 0\}$, $\Omega_3 = \{(r, f_A, f_M)|\ \pi + f_A - r < 0$ and $v_e \leq 0\}$, $\Omega_4 = \{(r, f_A, f_M)|\ \pi + f_A - r < 0$ and $v_e > 0$ and $(1-p)(f_A - r) + pf_M \geq 0\}$, $\Omega_5 = \{(r, f_A, f_M)|\ \pi + f_A - r \geq 0$ and $v_e > 0$ and $(1-p)(f_A - r) + pf_M < 0\}$, and $\Omega_6 = \{(r, f_A, f_M)|\ \pi + f_A - r < 0$ and $v_e > 0$ and $(1-p)(f_A - r) + pf_M < 0\}$.

Lemma 1 *If $v_e \leq 0$, then $(1-p)(f_A - r) + pf_M \geq 0$.*

Proof: Suppose instead that $v_e = (1-p)r - pf_M + (1-\beta)ES \leq 0$ and $(1-p)(f_A - r) + pf_M < 0$. Both inequalities can only be satisfied if $f_A + \frac{p}{1-p}f_M < r \leq \frac{p}{1-p}f_M$ which is impossible since $f_A \geq 0$. ■

Lemma 2 *If $(1-p)(f_A - r) + pf_M < 0$, then $v_e > 0$.*

Proof: Suppose $(1-p)(f_A - r) + pf_M < 0$ but that $(1-p)r - pf_M \leq 0$. Both inequalities can only be satisfied if $f_A + \frac{p}{1-p}f_M < r \leq \frac{p}{1-p}f_M$, which is impossible.■

Lemma 3 *If $p\bar{f}_M < (1-p)\pi$ and if $\pi + f_A - r < 0$, then $(1-p)(f_A - r) + pf_M < 0$.*

Proof: Let $pf_M < (1-p)\pi$ and $\pi + f_A - r < 0$. The second inequality is equivalent to $\pi < r - f_A$, and together with the first inequality implies that $pf_M < (1-p)\pi < (1-p)(r - f_A)$. Hence: $(1-p)(f_A - r) + pf_M < 0$. ■

Lemma 4 *If $p\bar{f}_M < (1-p)\pi$, then $\Omega_3 = \Omega_4 = \{\emptyset\}$.*

Proof: $(r, f_A, f_M) \in \Omega_4$ only if $\pi + f_A < r$ and $(1-p)(f_A - r) + pf_M \geq 0$. By Lemma 3, both inequalities cannot hold simultaneously when $p\bar{f}_M < (1-p)\pi$. $(r, f_A, f_M) \in \Omega_3$ only if $v_e \leq 0$ and $\pi + f_A < r$. By Lemma 1, the first inequality implies $(1-p)(f_A - r) + pf_M \geq 0$, which, by Lemma 3, cannot hold together with $\pi + f_A < r$ when $p\bar{f}_M < (1-p)\pi$. ■

Lemma 5 *If $p\bar{f}_M < (1-p)\pi$, then there does not exist an equilibrium in which $(r, f_A, f_M) \in \Omega_1$ and in which A is indifferent between activity and no activity.*

Proof: By Lemma 1, if $(r, f_A, f_M) \in \Omega_1$, $(1-p)(f_A-r)+pf_M \geq 0$. Hence, $(r, f_A, f_M) \in \Omega_1$ only if $(1-p)r - pf_M + (1-\beta)[(1-p)(f_A - r) + pf_M] \leq 0$, or equivalently:

$$r \leq \frac{p}{1-p} f_M - \frac{1-\beta}{\beta} f_A. \tag{18}$$

If $(r, f_A, f_M) \in \Omega_1$, $CS = \pi + f_A - r \geq 0$. As A must be indifferent between her two actions, (16) can be written as $r \geq \pi - \frac{1-\beta}{\beta} f_A$. But it cannot hold simultaneously with (18) if $p\bar{f}_M < (1-p)\pi$. ■

Proof of Proposition 2

We first tackle P's maximisation problem when $(r, f_A, f_M) \in \Omega_2$. We next show (in Lemmas 6, 7 and 8) that P cannot gain by choosing (r, f_A, f_M) either from Ω_1, Ω_5, or Ω_6. Let RB denote the reward to M if he catches A breaking the law. Let RF denote the reward to M if he extorts A. Let PB denote A's punishment after she was caught breaking the law. Let PF denote A's punishment when she is the victim of extortion. Observe that if $(r, f_A, f_M) \in \Omega_2$, $RB = B_b$ and $RF = B_f$, where B_b and B_f are defined in Section 3. A is indifferent between *activity* and *no activity* if $\mu^*(-PB) + (1-\mu^*)(\pi - PF) = (1-\mu^*)(-PF)$ (when ε is close to 0), or equivalently:

$$\mu^* = \frac{\pi}{\pi + PB} \tag{19}$$

M is indifferent between *monitor* and *extort* if $(1-\alpha^*)RB - 1 = RF$ (when ε is close to 0), or:

$$1 - \alpha^* = \frac{RF + 1}{RB} \tag{20}$$

If $(r, f_A, f_M) \in \Omega_2$, then, as seen from (2) and (5), $RB = r + (1-\beta)(\pi + f_A - r)$ and $PB = f_A - \beta(\pi + f_A - r)$. Hence, if $(r, f_A, f_M) \in \Omega_2$, then $PB = RB - \pi$. Using this insight, re-write (19) as: $\mu^* = \frac{\pi}{RB}$. Observe that $p\bar{f}_M < (1-p)\pi$ and $(r, f_A, f_M) \in \Omega_2$ imply $r \in \Big(\max\{0, \frac{p}{1-p} f_M - \frac{1-\beta}{\beta} f_A\},\ f_A + \frac{p}{1-p} f_M\Big]$. Observe also that if $r = \max\{0, \frac{p}{1-p} f_M - \frac{1-\beta}{\beta} f_A\}$, then $RB = \max\{(1-\beta)(\pi + f_A), (1-\beta)\pi + \beta \frac{p}{1-p} f_M\}$. Similarly, if $r = f_A + \frac{p}{1-p} f_M$, then $RB = f_A + (1-\beta)\pi + \beta \frac{p}{1-p} f_M$. Depending on

(f_A, f_M), RB can thus take any value in $[(1-\beta)\pi,\ \bar{f}_A + (1-\beta)\pi + \beta\frac{p}{1-p}\bar{f}_M]$. As $RB = \frac{\pi}{\mu} \leq \bar{f}_A + (1-\beta)\pi + \beta\frac{p}{1-p}\bar{f}_M$, this implies that

$$\mu \geq \frac{\pi}{\bar{f}_A + (1-\beta)\pi + \beta\frac{p}{1-p}\bar{f}_M} \tag{21}$$

As $RB = \frac{\pi}{\mu} \geq (1-\beta)\pi$, this implies that $\mu \leq \frac{\pi}{(1-\beta)\pi}$, which is always satisfied. Observe that the RHS of (21) is less than 1 if $\beta\pi < \bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M$, which, by assumption, is satisfied. Hence, any $\mu \in \Big[\pi/[\bar{f}_A + (1-\beta)\pi + \beta\frac{p}{1-p}\bar{f}_M],\ 1\Big]$ can be implemented by a $(r, f_A, f_M) \in \Omega_2$.

Substituting $RB = \frac{\pi}{\mu}$ into (20) and re-arranging, one has:

$$\alpha(\mu) = 1 - \mu\frac{RF+1}{\pi} \tag{22}$$

If $(r, f_A, f_M) \in \Omega_2$, then, as seen from (3),

$$\begin{aligned} RF = v_e &= (1-p)r - pf_M + (1-\beta)[(1-p)(f_A - r) + pf_M] \\ &= \beta(1-p)r - \beta pf_M + (1-\beta)(1-p)f_A \end{aligned} \tag{23}$$

Since $RB = r + (1-\beta)(\pi + f_A - r) = \frac{\pi}{\mu}$, then

$$r = \frac{1-(1-\beta)\mu}{\beta\mu} \cdot \pi - \frac{1-\beta}{\beta}f_A \tag{24}$$

Substituting r in (23) with (24), one has

$$RF(\mu) = \frac{1-(1-\beta)\mu}{\mu}(1-p)\pi - \beta pf_M \tag{25}$$

Observe that, if $p\bar{f}_M < (1-p)\pi$, then $RF(\mu) > 0$ for any μ. Inserting (25) into (22) and re-arranging yields

$$\alpha(\mu) = p + \mu\Big[(1-\beta)(1-p) + \frac{\beta pf_M - 1}{\pi}\Big] \tag{26}$$

Observe that, under Assumption 1, $(1-\beta)(1-p) + \frac{\beta p\bar{f}_M - 1}{\pi} < 1-p$, and thus that $\alpha(\mu) \in (p, 1)$. Observe also that, under assumption 3, $(1-\beta)(1-p) + \frac{\beta p\bar{f}_M - 1}{\pi} > 0$. If $(r, f_A, f_M) \in \Omega_2$, then welfare equals

$$Wel_{\Omega_2}(\mu) = \underline{w} + \pi + \alpha(\mu)[\Delta w - \pi] - \mu, \tag{27}$$

where $\alpha(\mu)$ is given by (26). It can be checked that $\partial Wel_{\Omega_2}(\mu)/\partial\mu > 0$ iff $\Delta w > \kappa'\pi$, where $\kappa' \equiv \frac{\beta p\bar{f}_M+(1-\beta)(1-p)\pi}{\beta p\bar{f}_M+(1-\beta)(1-p)\pi-1} > 1$. Note that $\kappa' > 0$ if $(1-\beta)(1-p)\pi + \beta p\bar{f}_M > 1$, which, by assumption, is satisfied.

Thus, if $\Delta w \in (\pi,\ \kappa'\pi)$, and if the activity is declared illegal, P chooses a $(r, \bar{f}_A, \bar{f}_M) \in \Omega_2$ which implements the lowest possible μ (namely, $\mu^* = \pi/[\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi]$), while if $\Delta w > \kappa'\pi$, then P chooses a $(r, f_A, \bar{f}_M) \in \Omega_2$ which implements $\mu^* = 1$.

Suppose $\Delta w \in (\pi, \kappa'\pi)$. If P legalises the activity, then welfare equals $\underline{w} + \pi$. It follows from the above paragraph that if P declares the activity illegal then

$$Wel_{\Omega_2} = \underline{w} + \pi + \frac{p\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi - 1}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}\Big(\Delta w - \pi\Big) - \frac{\pi}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}.$$

It can be checked that P legalises the activity if and only if $\Delta w < \kappa\pi$, where $\kappa = \frac{p\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}{p\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi-1}$. Observe that $\kappa < \kappa'$ if and only if $p\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi > \beta p\bar{f}_M + (1-\beta)(1-p)\pi$, which is obviously satisfied. ■

Lemma 6 *Suppose P maximises welfare subject to the constraint that both A and M must be indifferent between their two actions. Then, P cannot gain by choosing (r, f_A, f_M) from Ω_5 instead of Ω_2.*

Proof: Observe that if $(r, f_A, f_M) \in \Omega_2$, then welfare is given by (27), while if $(r, f_A, f_M) \in \Omega_5$, then

$$Wel_{\Omega_5} = \underline{w} + \pi + \alpha^*(\Delta w - \pi) - \mu^* - (1-\mu^*)\Big[pf_M + (1-p)(r + f_A)\Big]. \qquad (28)$$

Hence, P prefers to choose (r, f_A, f_M) from Ω_5 instead of from Ω_2 only if it implements a higher α (for a given μ). We will prove that this necessary condition is not satisfied.

It follows from Lemma 2 that $(r, f_A, f_M) \in \Omega_5$ if $CS > 0$ and $ES = 0$, which can be summarised as $r \in \Big(f_A + \frac{p}{1-p}f_M,\ f_A + \pi\Big]$. Using an identical reasoning as that used in Prop. 2, if $(r, f_A, f_M) \in \Omega_5$, then $RB = \frac{\pi}{\mu}$. We will shortly see that if $(r, f_A, f_M) \in \Omega_5$, it is optimal to set $r = f_A + \frac{p}{1-p}f_M$. Using an identical procedure as in Prop. 2, it can be checked that if $r = f_A + \frac{p}{1-p}f_M$, then RB can take any value in $\Big[(1-\beta)\pi,\ \bar{f}_A+(1-\beta)\pi+\beta\frac{p}{1-p}\bar{f}_M\Big]$, which is the same range as the one we obtained when

$(r, f_A, f_M) \in \Omega_2$. Observe that if $(r, f_A, f_M) \in \Omega_5$, then $(v_e =)RF = (1-p)r - pf_M$. Since (24) also holds when $(r, f_A, f_M) \in \Omega_5$, substituting it into the expression for $RF(\mu)$ here, one obtains:

$$RF(\mu) = (1-p)\Big[\frac{\pi}{\beta\mu} - \frac{1-\beta}{\beta}(\pi + f_A)\Big] - pf_M \tag{29}$$

Observe that $RF(\mu) > 0$ if and only if

$$f_A < \frac{1-(1-\beta)\mu}{(1-\beta)\mu} \cdot \pi - \frac{\beta}{1-\beta} \cdot \frac{p}{1-p} f_M \tag{30}$$

Plot (30) together with $r \geq f_A + \frac{p}{1-p} f_M$. Both inequalities are satisfied only if $f_A \leq \frac{1-(1-\beta)\mu}{\mu} \cdot \pi - \beta\frac{p}{1-p} f_M$ It follows from (22) that, for any given μ, $\alpha(\mu)$ is decreasing in RF. It follows from (29) that $RF(\mu)$ is decreasing in f_A. Hence, it is optimal to set $f_A = \frac{1-(1-\beta)\mu}{\mu} \cdot \pi - \beta\frac{p}{1-p} f_M \equiv f_A(\mu)$. Observe that if f_A is fixed in this way, this implies that $r = f_A + \frac{p}{1-p} f_M$. Inserting $f_A(\mu)$ into (29), one has, after re-arranging, $RF(\mu) = (1-p)\pi\frac{1-(1-\beta)\mu}{\mu} - \beta p f_M$. Inserting the latter into (22), one has $\alpha(\mu) = p + \mu\Big[(1-\beta)(1-p) + \frac{\beta p f_M - 1}{\pi}\Big]$, which is equivalent to (26). Hence, any implementable μ if $(r, f_A, f_M) \in \Omega_5$ does not result in more law abiding behaviour than if μ were implemented by $(r, f_A, f_M) \in \Omega_2$. ■

Lemma 7 *Suppose P maximises welfare subject to the constraint that both A and M must be indifferent between their two actions. Then, P cannot gain by choosing (r, f_A, f_M) from Ω_6 instead of Ω_2.*

Proof: If $(r, f_A, f_M) \in \Omega_6$, then P's maximization problem stated on p. 28 simplifies to:

$$\begin{aligned}
\max_{r, f_A, f_M} Wel_{\Omega_6} = \;& \underline{w} + \pi + \alpha^*(\Delta w - \pi) - \mu^* - (1-\alpha^*)\mu^*\pi - (1-\mu^*)pf_M - \\
& -(f_A + r)[(1-\alpha^*)\mu^* + (1-\mu^*)(1-p)] \qquad (31)\\
s.t. \qquad & 1 - \alpha^* = \tfrac{(1-p)r - pf_M + 1}{r} \qquad (32)\\
& \mu^* = \tfrac{\pi}{\pi + f_A} \qquad (33)\\
& f_A \in [0, \bar{f}_A], \quad f_M \in [0, \bar{f}_M] \qquad (34)\\
& r > \pi + f_A \quad \text{and} \quad r > f_A + \tfrac{p}{1-p} f_M \qquad (35)\\
& r \geq \tfrac{1 - pf_M}{p} \quad \text{and} \quad r \geq \tfrac{pf_M - 1}{1-p} \qquad (36)
\end{aligned}$$

where (35) define $(r, f_A, f_M) \in \Omega_6$, while (36) ensure that $\alpha^* \geq 0$ and $\alpha^* \leq 1$.

Note that, in this case $\mu^* \in \left[\frac{\pi}{\pi+\bar{f}_A}, 1\right]$. Note also that the lowest monitoring intensity that can be implemented by any $(r, f_A, f_M) \in \Omega_2$ (as shown in the proof of Prop. 2) equals $\frac{\pi}{\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}$, which is greater than $\frac{\pi}{\pi+\bar{f}_A}$ under (A1). We will show, however, that despite the fact that P can implement a lower monitoring intensity if she chooses $(r, f_A, f_M) \in \Omega_6$, she still prefers to select (r, f_A, f_M) from Ω_2. We now break the proof in two different cases.

Case 1 ($p\bar{f}_M > 1$): Observe that welfare obtained from the maximization problem above is weakly less than the one that would be obtained if we were to replace Wel_{Ω_6} by

$$Wel'_{\Omega_6} = \underline{w} + \pi + \alpha^*(\Delta w - \pi) - \mu^* - (1-\alpha^*)\mu^*\pi - (f_A + r)[(1-\alpha^*)\mu^* + (1-\mu^*)(1-p)].$$

Observe also that Wel'_{Ω_6} is maximized when $f_M = \bar{f}_M$. Moreover, note that in this case $r > \pi + f_A$ implies the other three constraints in (35)–(36) provided that additionally $p\bar{f}_M < (1-p)\pi$. Furthermore, it follows from (32) that, in this case, $\frac{\partial \alpha^*}{\partial r} < 0$. Hence, Wel'_{Ω_6} is maximized only if r is close to its lower bound, $\pi + f_A$. Replacing r by its lower bound in (35), and rewriting yields

$$\alpha^* = p + \frac{pf_M - 1}{\pi + f_A}. \tag{37}$$

Rewriting (33) as $f_A(\mu) = \frac{\pi}{\mu} - \pi$ and inserting into our last expression of α^* yields

$$\alpha^*(\mu) = p + \frac{p\bar{f}_M - 1}{\pi}\mu. \tag{38}$$

Observe that the RHS of (38) is less than the RHS of (26) if and only if (A1) holds. This proves that, in this case, if P wants to implement a $\mu \in \left[\frac{\pi}{\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}, 1\right]$ using a $(r, f_A, f_M) \in \Omega_6$, this does not result in more law abiding behaviour by A than if the same μ were implemented by a $(r, f_A, f_M) \in \Omega_2$.

Replacing α^*, f_A and r in Wel'_{Ω_6} by, respectively, $\alpha^*(\mu)$ in (38), $\frac{\pi}{\mu} - \pi$ and $\frac{\pi}{\mu}$, and then differentiating the resulting $Wel'_{\Omega_6}(\mu)$ with respect to μ, one obtains:

$$\frac{\partial Wel'_{\Omega_6}}{\partial \mu} = \frac{p\bar{f}_M - 1}{\pi}(\Delta w - \pi) + 2\left(p\bar{f}_M - 1\right) + (1-p)\pi\left(\frac{2}{\mu^2} - 1\right) - 1,$$

which is strictly positive as we assume that $1 < p\bar{f}_M < (1-p)\pi$. Hence, Wel'_{Ω_6} is maximal when $f_M = \bar{f}_M$, $f_A = 0$ (as this implements $\mu = 1$) and $r = \pi$. Hence we have established that the maximal welfare that can be obtained when $(r, f_A, f_M) \in \Omega_6$ is weakly less than Wel'_{Ω_6} if $(r, f_A, f_M) = (\pi, 0, \bar{f}_M)$. Moreover, it follows from above that Wel'_{Ω_6} if $(r, f_A, f_M) = (\pi, 0, \bar{f}_M) \leq Wel_{\Omega_2}$ if $(r, f_A, f_M) \in \Omega_2$ and if it implements $\mu = 1$, which is weakly less than the maximal welfare that can be obtained when $(r, f_A, f_M) \in \Omega_2$.

Case 2 ($p\bar{f}_M < 1$): In this case $\frac{\partial \alpha^*}{\partial r} > 0$, and it follows from 32 that α^* is bounded above by p, which, if $(1-\beta)(1-p)\pi + \beta p \bar{f}_M > 1$, is less than the RHS of (26). As above, this proves that, in this case, if P wants to implement a $\mu \in \left[\frac{\pi}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}, 1\right]$ using a $(r, f_A, f_M) \in \Omega_6$, this does not result in more law abiding behaviour by A than if the same μ were implemented by a $(r, f_A, f_M) \in \Omega_2$.

Setting $f_A = f_M = r = 0$ and $\alpha = p$, (31) becomes $\underline{w} + p\Delta w + (1-p)(1-\mu)\pi - \mu$. Since, straightforwardly, this is a decreasing function of μ, we would maximise welfare if μ is set at the lowest feasible level, namely, $\mu = \frac{\pi}{\pi + \bar{f}_A}$. With this substitution, we conclude that 31 is bounded above by $\underline{w} + p\Delta w + \frac{\pi}{\pi + \bar{f}_A}\Big((1-p)\bar{f}_A - 1\Big) \equiv Wel_{\Omega_6}$. From the proof of Prop. 2 we know that welfare if $(r, f_A, f_M) \in \Omega_2$ is bounded below by

$$\underline{w} + \pi + \frac{p\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi - 1}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}\Big(\Delta w - \pi\Big) - \frac{\pi}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi} \equiv Wel_{\Omega 2}$$

or, after simplifying:

$$Wel_{\Omega_2} = \underline{w} + \frac{p\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi - 1}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}\Delta w + \pi\frac{(1-p)\bar{f}_A}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}$$

To verify that P cannot gain by choosing a reward/punishment scheme from Ω_6 instead of Ω_2, we calculate:

$$\begin{aligned} Wel_{\Omega_2} - Wel_{\Omega_6} &= \Delta w\Big(\frac{p\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi - 1}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi} - p\Big) \\ &\quad + (1-p)\bar{f}_A\pi\Big[\frac{1}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi} - \frac{1}{\pi + \bar{f}_A}\Big] + \frac{\pi}{\pi + \bar{f}_A}. \end{aligned}$$

It can then be checked that the term in round brackets is equal to $\frac{\beta p\bar{f}_M + (1-\beta)(1-p)\pi - 1}{\bar{f}_A + \beta\frac{p}{1-p}\bar{f}_M + (1-\beta)\pi}$ which is positive because of our assumption that $\beta p\bar{f}_M + (1-\beta)(1-p)\pi > 1$; while the term in square brackets is equal to $\frac{\frac{\beta}{1-p}((1-p)\pi - p\bar{f}_M)}{(\pi+\bar{f}_A)(\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}$, which is also positive due to our assumption that $p\bar{f}_M < (1-p)\pi$.■

Lemma 8 *If $p\bar{f}_M < (1-p)\pi$, P cannot gain by declaring the activity illegal and by setting (r, f_A, f_M) such that either A or M strictly prefer one of their actions.*

Proof: Recall that Lemmas 4 and 5 imply that if $p\bar{f}_M < (1-p)\pi$, the reward/punishment scheme must be chosen from Ω_1, Ω_2, Ω_5 or Ω_6.

We can discard the case in which P declares the activity illegal and in which A strictly prefers to break the law as this yields a welfare not greater than the one obtained by choosing $(r, f_A, f_M) \in \Omega_2$. There does not exist an equilibrium in which P declares the activity illegal and in which A strictly prefers to obey the law. To check, note that A only strictly prefers to obey the law if $\mu^* > 0$. As A obeys the law, this implies that if M monitors, he gets -1. If he extorts, he gets RF, which is strictly positive when (r, f_A, f_M) is chosen from Ω_2, Ω_5 or Ω_6. If he does nothing, he gets zero. Thus, if M anticipates that $\alpha = 1$, it is a best reply for him to set $\mu = 0$, a contradiction.

Hence, if there exists an equilibrium in which either A or M strictly prefer one of their actions it must be that either (a) M strictly prefers to monitor and A is indifferent, or (b) M strictly prefers to extort (or to do nothing) and A is indifferent. We can discard (b) because if $\mu = 0$, then A strictly prefers to break the law. Consider (a). It follows from our proof of Lemma 2 that in such a candidate equilibrium $RB = \pi$ (as otherwise A would not be indifferent given that $\mu = 1$) and that $\alpha < 1 - \frac{RF+1}{\pi}$ (as otherwise M would not strictly prefer to monitor). Suppose P deviates from this candidate equilibrium by choosing a $(r, f_A, f_M) \in \Omega_2$ which increases RB by an ε (we know from our proof of Lemma 2 that Ω_2 includes (r, f_A, f_M)'s which implement a $\mu = 1$ and a μ close to one.). This uniquely implements $\mu = \frac{\pi}{\pi+\varepsilon} < 1$ (as otherwise A would not be indifferent). As $\mu \in (0,1)$, A must obey the law with a probability close to $1 - \frac{RF+1}{\pi}$ to make M indifferent. As A abides by the law more often and as μ is decreased by an ε and as P never sends anyone to jail, this deviation increases welfare. ■

Proof of Proposition 3

It follows from Proposition 1 and from the proof of Proposition 2 that the laissez-faire region with an honest M is smaller than the one with a corrupt and extorting M if and only if $\pi + 2\frac{\pi}{\pi+\bar{f}_A} < \kappa\pi \Leftrightarrow$

$$2p\bar{f}_A + 2\beta\frac{p}{1-p}\bar{f}_M + 2(1-\beta)\pi < 2 + \pi + \bar{f}_A. \tag{39}$$

Replacing p by $\frac{\pi}{\pi+\bar{f}_M}$ in the inequality above we conclude that (39) is satisfied if $\pi - 2 \leq \frac{\bar{f}_M-\pi}{\bar{f}_M+\pi}\bar{f}_A$ which is the inequality stated in the proposition. ■

Proof of Proposition 4

It follows from Proposition 2 that welfare is continuous in Δw and that if $\Delta w \in (\pi, \kappa\pi)$, welfare $= \underline{w} + \pi$, if $\Delta w \in (\kappa\pi, \kappa'\pi)$, welfare $= \underline{w} + \pi + \alpha_2^*(\Delta w - \pi) - \mu_2^*$ and if $\Delta w \geq \kappa'\pi$, welfare $= \underline{w} + \pi + \alpha_3^*(\Delta w - \pi) - 1$, where $\alpha_2^* = \frac{p\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi-1}{\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}, \mu_2^* = \frac{\pi}{\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}$, $\alpha_3^* = p + (1-\beta)(1-p) + \frac{\beta p\bar{f}_M-1}{\pi}, \kappa = \frac{p\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi}{p\bar{f}_A+\beta\frac{p}{1-p}\bar{f}_M+(1-\beta)\pi-1}$ and $\kappa' = \frac{\beta p\bar{f}_M+(1-\beta)(1-p)\pi}{\beta p\bar{f}_M+(1-\beta)(1-p)\pi-1}$. It is routine to check that $\frac{\partial\alpha_2^*}{\partial p} > 0$, $\frac{\partial\mu_2^*}{\partial p} < 0$, $\frac{\partial\alpha_3^*}{\partial p} > 0$ and $\frac{\partial\kappa}{\partial p} < 0$. Furthermore, it is also straightforward to check that, if $\beta\bar{f}_M > (1-\beta)\pi$, $\frac{\partial\kappa'}{\partial p} < 0$. These insights, combined with the fact that $\alpha_2^* < \alpha_3^*$, imply the proposition. ■

References

Becker G., Stigler G. Law enforcement, malfeasance, and compensation of enforcers, The Journal of Legal Studies 1974; 3(1); 1–18.

Bolton P. The principle of maximum deterrence revisited, University of Berkeley Working Paper 8749, August 1987

Carrillo J. Corruption in hierarchies, Annales d'Economie et de Statistique 2000a; 59; 37–61.

Carrillo J. Grafts, bribes and the practice of corruption, Journal of Economics & Management Strategy 2000b; 9(2); 257–286.

Di Tella R. and Schargrodsky E. The role of wages and auditing during a crackdown on corruption in the city of Buenos Aires, Journal of Law and Economics 2003, 46; 269–92.

Fudenberg D., Tirole J. Game Theory. The MIT Press: Cambridge, Massachusetts; 1991.

Grossman G. M. and Katz M. L. Plea bargaining and social welfare, American Economic Review 1983, 73(4); 749–57.

Hindriks J, Keen M, Muthoo A. Corruption, extortion and evasion, Journal of Public Economics 1999, 74; 395-430.

Kessler A. On monitoring and collusion in hierarchies, Journal of Economic Theory 2000, 91(2); 280–91.

Khalil F, Lawarrée J. Incentives for corruptible auditors in the absence of commitment, Journal of Industrial Economics 2006, 54(2); 269–91.

Khalil, F., J. Lawarrée, and S. Yun (2007) "Bribery vs. Extortion: Allowing the Lesser of Two Evils", *mimeo*, University of Washington, Seattle.

Klitgaard R. Controlling corruption. University of California Press: Berkeley and Los Angeles, California; 1988.

Kofman F, Lawarrée J. Collusion in hierarchical agency, Econometrica 1993, 61(3); 629–56.

Kugler M, Verdier T, Zenou Y. Organized crime, corruption and punishment, Journal of Public Economics 2005, 89; 1639–1663.

Mishra A. Hierarchies, incentives and collusion in a model of enforcement, Journal of Economic Behavior and Organization 2002, 47; 165–178.

Polinsky A. M., Shavell S. Corruption and optimal law enforcement, Journal of Public Economics 2001, 81(1); 1–24.

Polinsky A. M., Shavell S. Legal error, litigation, and the incentive to obey the law, Journal of Law, Economics, and Organization 1989, 5(1); 99–108.

Strausz R. Delegation of monitoring in a principal-agent relationship, Review of Economic Studies 1997, 64(3); 337–57.

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

Appendix B (not for publication)

So far we assumed that P can only punish A or M by putting her/him in jail. In this appendix we show that, under an additional (sufficient) condition, our results do not change if P could levy hefty fines on them.

Proposition 5 *Suppose f_A and f_M are pure transfers. Suppose, additionally, that either $p\bar{f}_M > 2$ or that $p\bar{f}_M < 1$. Then, Proposition 2 remains unchanged.*

Proof: If f_A and f_M are pure transfers, then the objective function will no longer include any terms with f_A or f_M. That means that the parts of our proofs that potentially might change and require a check are those involving the objective function.

Observe that all our results when $(r, f_A, f_M) \in \Omega_2$ remain unchanged as P's objective function does not include any f_A or f_M then.

In the proof of Lemma 6, the objective function stated in (28) now becomes

$$Wel'_{\Omega_5} = \underline{w} + \pi + \alpha^*(\Delta w - \pi) - \mu^* - (1-\mu^*)(1-p)r. \tag{40}$$

However, the rest of the proof remains the same: all the reasoning there applies when (40) is used in place of (28).

In the proof of Lemma 7, the maximand in (31) should now read:

$$Wel_{\Omega_6} = \underline{w} + \pi + \alpha^*(\Delta w - \pi) - \mu^* - (1-\alpha^*)\mu^*\pi - r[(1-\alpha^*)\mu^* + (1-\mu^*)(1-p)] \tag{41}$$

This amendment does affect the reasoning in Case 1 slightly, since we cannot use Wel'_{Ω_6} as the (upper) estimate of Wel_{Ω_6}. We have to use Wel_{Ω_6} in (41) itself. Simplify (41) slightly to get

$$Wel_{\Omega_6} = \underline{w} + \alpha\Delta w + (1-\alpha)(1-\mu)\pi - \mu - r[(1-\alpha)\mu + (1-\mu)(1-p)]. \tag{42}$$

Then following the steps of the proof, substitute for α and r with $p + \frac{p\bar{f}_M - 1}{\pi}\mu$ and $\frac{\pi}{\mu}$, respectively, to obtain after re-arranging:

$$Wel_{\Omega_6} = \underline{w} + (p + \frac{p\bar{f}_M - 1}{\pi}\mu)\Delta w - \mu(1-p)\pi + (p\bar{f}_M - 1)\mu^2 - \mu - \frac{\pi}{\mu}(1-\mu)(1-p) \tag{43}$$

Differentiating with respect to μ and manipulating:

$$\frac{\partial Wel_{\Omega_6}}{\partial \mu} = \frac{p\bar{f}_M - 1}{\pi}\Delta w + 2(p\bar{f}_M - 1)\mu + \pi(1-p)(\frac{1}{\mu^2} - 1) - 1. \tag{44}$$

This partial derivative is positive when $\pi < (p\bar{f}_M - 1)\Delta w$. Replacing Δw by its lower bound (i.e. by π), we conclude that the partial derivative is positive when $p\bar{f}_M > 2$. In that case, P maximizes welfare by choosing a (r, f_A, f_M) which implements a $\mu^* = 1$. As argued in Lemma 7, however, this does not result in more welfare than if $\mu^* = 1$ were implemented by a $(r, f_A, f_M) \in \Omega_2$. The reasoning in Case 2 of Lemma 7 (i.e. when $p\bar{f}_M < 1$) does not rely on the fact that P can only punish A or M by sending them to jail. Therefore, all the results mentioned in Case 2 remain valid. ■