Diluting Deterrence: Extra-Legal Actors and the Courts[*]

A. V. Chari[†]

October 2009
DRAFT VERSION

Abstract

The presence of extra-legal actors who take unilateral (or unsanctioned) preventive action against would-be criminals appears to have the potential to undermine the courts. I use game-theoretic reasoning to make three points: (1) The deterrent power of extra-legal actors depends on the cost of unilateral action (relative to multilateral/ officially sanctioned action) and the standards of proof of the courts, (2) When there exists a mechanism by which individuals can credibly signal their intention to not commit crimes, it may be optimal for the courts to keep the deterrent power of extra-legal actors intact, and (3) When there exists no such mechanism, the optimal strategy of the courts may be to control the extra-legal actor by changing the relative cost of unilateral action. But if the latter cannot be varied by the courts, the second-best solution would be to vary the standard of proof. In particular, lower standards of proof may be required in order to "deactivate" the extra-legal actor and thereby reduce the probability of wrongful punishments.

Introduction

Two recent developments, (1) the (near) unilateral invasion of Iraq in 2003 and (2) the controversial program of preventive detention of 'enemy combatants' in military custody, appear in their own ways to have undermined the rule of law and the role of multilateral organizations like the UN (in the former case).

The two cases mentioned above have certain similarities: in each case, an unconventional threat (the threat of nuclear proliferation and the threat of terrorism) was

[*] I am grateful to Talia Bar, Kaushik Basu, Ani Guerdjikova and Siddharth Sharma for comments on a previous version of this paper.

[†] Allyn Young Fellow, Department of Economics, Cornell University. Email: chari@cornell.edu.

deemed to call for an unconventional response. The essential characteristic of these threats is that they seem to call for pre-emptive action, because punishment after the fact is either moot (as in the case of suicide bombers) or simply not credible (as in the case of nuclear acquisition). In both cases a decision was made to bypass the established mechanisms (namely, the courts and the UN) for dealing with these threats, ostensibly because the standards of proof required by these institutions was inadequate for successful pre-emption.

This paper addresses the relation between courts and extra-legal actors. Using game-theoretic analysis, I show that the courts (or the UN) can (unwittingly, perhaps) moderate the influence of extra-legal/unilateral actors by means of two instruments: (1) the extent of 'cost-sharing', i.e. the extent to which unilateral action is more expensive for the extra-legal actor than multilateral or legally-sanctioned action, and (2) the standard of proof. Consider the case of a rogue nation attempting to acquire nuclear weapons. To the extent that the rogue's actions are imperfectly observable, the US may be willing to take action on the basis of relatively scanty evidence of acquisition activity. This threat of unilateral action may in turn be sufficient to deter any attempts at acquisition. However, in the presence of the UN, the US may be willing to stay its hand if it believes that further evidence may come to light in the future that could result in an intervention by the UN, which would be less costly for the US. The resulting delay in response may be sufficient to encourage the rogue to attempt acquisition. That is, the existence of a low-cost, albeit unreliable, alternative to unilateral action may be sufficient to undermine the ability of the US to deter rogues from acquiring nuclear weapons.

The relation between extra-legal actors and the courts has not (to my knowledge) been previously explored in the literature on deterrence. In game-theoretic terms, however, the situation modeled here has the essential flavor of models of entry deterrence in industrial organization: an incumbent monopolist tries to deter entry by threatening a price-war but the threat lacks credibility because it cannot be in the incumbent's interest to actually carry out the threat once entry has occurred. In our framework, the extra-legal actor deters crime by threatening action if there is any suspicious activity, but this threat becomes less credible when a low-cost alternative, i.e. the courts, comes into existence. What is interesting here is the form of a game of preventive action and the way in which the courts moderate the credibility of extra-legal action. When attempts to commit crimes can be thwarted by preventive action, there is an incentive for the would-be criminal to conceal his actions as well as the timing of his attempts.[1] In turn, the imperfect observability of actions

[1] As we will see in the next section, this gives rise to an imperfect information variant of a

creates a trade-off for the extra-legal actor: he can either take preventive action now, even though the evidence is lacking, and incur a large cost, or he can choose to wait, hoping that the crime is yet to be attempted and that when it is attempted it will be observed (with some probability) and be prevented by the courts. The relative cost of unilateral (unsanctioned) action and the relative willingness of the courts/multilateral organizations to take action on the basis of incomplete evidence will determine whether the extra-legal actor prefers to act immediately or prefers to wait instead. I show that it is possible that these two factors can be such as to make unilateral action non-credible, and thereby dilute the deterrent ability of the extra-legal actor.

When should the courts act to dilute deterrence? I show that in a setting where it is possible for potential criminals to credibly renounce any attempts at crime, it may be optimal for the courts to keep the deterrent power of extra-legal actors intact. A good example is the one of nuclear proliferation: there exists a credible mechanism by which countries can signal their intentions to not attempt nuclear acquisition (namely by allowing UN weapons inspectors into their facilities). In this case, the UN may want to keep both the level of multilateral cost-sharing, as well as its rate of intervention low, so as to enable deterrence by the US. But when such a credible signal does not exist, as for example in the case of individuals who are suspected of being terrorists, it may be optimal for the courts to undermine the extra-legal actor, so as to prevent the latter from administering wrongful punishments. This analysis implies that extra-legal actors do not regard the rule of law as a mere irrelevance; on the contrary, they may even want to take an active interest in dismantling it. A related implication is that the incidence of crime may paradoxically be lower in settings where the rule of law is very weak, because this is when the deterrent power of extra-legal enforcers is maximal.

The literature on the optimal standard of proof (see for example Kaplow and Shavell 1994 and Lando 2009) stresses the trade-off between achieving deterrence and minimizing the possibility of wrongful punishments. Increasing the probability of conviction by lowering the standard of proof may have a strong deterrent effect on criminal activity but comes at the cost of increasing the probability of wrongful convictions. When extra-legal actors are present, the determination of the optimal standard of proof is based on a very different reasoning: because the presence of extra-legal actors produces a situation of too many wrongful punishments, the optimal standard of proof will need to be set so as to deter the extral-legal actors.

so-called "timing game" (see for example Fudenberg and Tirole 1991).

I show that this may involve choosing a low standard of proof, i.e. increasing the probability of convictions (including wrongful ones) for cases that are brought to court, in order to reduce the overall (unconditional) probability of Type I errors. This standard of proof may be sub-optimal relative to what the courts would like to implement in a setting without extra-legal actors. The basic issue is one that is familiar to economists: the standard of proof is being used to control the activities of the extra-legal actor as well as to rein in Type I errors. Instead, if the courts can also control the relative cost of unilateral action, they may be able to achieve the first-best solution.

The rest of the paper is devoted to formalizing the above intuition.

Theory

The Basic Setting

The model is set in discrete time, with infinitely many time periods. There are two risk-neutral players, the extra-legal actor S and his 'opponent' R. In each period R can choose whether or not to acquire a capability (or commit a crime) - acquisition incurs a cost c_B. In the same period, S observes a noisy signal of $R's$ action. Denoting the signal in period t by s_t, it is assumed that:

$$\begin{aligned}\Pr(s_t &= 0|Acq_t) = \theta_0; \quad \Pr(s_t = 1|Acq_t) = 1-\theta_0 \\ \Pr(s_t &= 0|NotAcq_t) = 1; \;\; \Pr(s_t = 1|NotAcq_t) = 0\end{aligned}$$

That is, there is a possibility of false negatives but there are no false positives.[2] θ_0 represents the maximal level of concealment that R can obtain. For simplicity, we will assume that concealment is costless (or equivalently, that it has only a fixed cost component that is subsumed in c_B) - this will guarantee that the maximal level of concealment will always be chosen. Allowing for a variable cost of concealment may provide an interesting extension of the model, but one that I have not explored in this paper. I will return to this point briefly after presenting the solution of the model.

Upon observing the signal, S can then choose to attack R in the current period and disrupt the acquisition or to wait. Attacking entails a cost C_G when the signal is negative, and a cost C_J (with $C_J \leq C_G$) when the signal is positive. This formulation allows for the possibility that when the signal is positive there is a

[2] This assumption is for convenience: allowing for false positives does not affect the basic intuition behind the model, but it does complicate the algebra.

possibility of intervention by the courts (or the UN) which entails a lower cost for S, so that $C_J < C_G$. In keeping with the interpretation of the legal authority as the UN or a similar multilateral organization, we will often refer to this possibility as 'cost-sharing', with the understanding that it simply refers to the more general assumption that $C_J < C_G$. Later, when discussing the implications of the model, we will draw a distinction between settings where the extent of cost-sharing can be controlled and settings where it makes more sense to think of varying the rate of intervention of the courts/judicial system.

The setup of the model assumes that there is no possibility of cost-sharing (or intervention by the formal legal system) when the signal is negative. On the other hand, the extra-legal actor, almost by definition, is less conservative when willing to consider taking action even in this state of the world. The assumption we are making is that the extra-legal actor is less conservative than the formal legal system and may consider taking action even when the signal is negative. This may be due to the high subjective cost of criminal activity as far as S is concerned, or may reflect the possibility that S is less concerned about Type I errors (i.e. wrongful punishments) than the courts (in our model, S is completely unconcerned about Type I errors).

The cost to R of being attacked is D. If R succeeds in acquiring the capability, he gets a benefit of B and S gets a negative payoff, $-F$. The game ends when either R successfully acquires or S attacks R. The game in any period t looks as depicted in Figure 1.

In the game tree above, V^R and V^S denote the continuation values for R and S (i.e. their expected payoffs in the subgame that begins in period $t+1$). Two assumptions are required to ensure that the game we have described is not trivial:

Assumption I: $F > C_G$
Assumption II: $\theta_0 B > c_B + (1-\theta_0)D$

Assumption I ensures that S will definitely attack if the signal is positive, even if there is no cost-sharing. Assumption II ensures that if S does not intend to attack when the signal is negative (i.e. $s_t = 0$), then R should strictly prefer to attempt acquisition in that period. I note for future reference that (given that Assumption I holds) the quantity $\theta_0 B - [c_B + (1-\theta_0)D]$ represents player R's maximum possible payoff in this game. Finally, I assume (for the moment, but I will relax this assumption subsequently) that there is no discounting of future payoffs, i.e. players are infinitely patient.

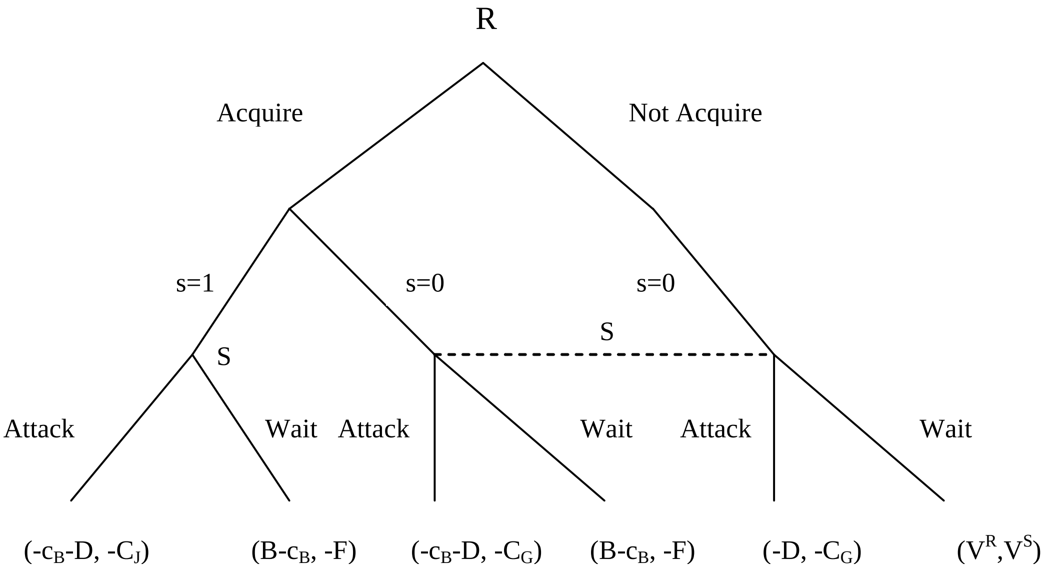

Figure 1: The game in any period

The appropriate solution concept in this game is that of Perfect Bayesian Equilibrium (PBE). Formally, a pure strategy for player R specifies an action at each time period t for every history of play up to time period t. A pure strategy for player S specifies an action at each time period t for every history of play up to period t and each of the two possible signals s_t. In a PBE, (1) Player S's strategy must be optimal for each subgame of the original game, given his beliefs at each point at which he may be asked to move; (2) Player R's strategy must be optimal in each subgame of the original game, and (3) Player S's beliefs at each point must be consistent with R's optimal strategy, where consistency implies that the beliefs are derived using Bayes' Rule.

What are the equilibria in this game? I start with the case in which there is no multilateral cost-sharing, i.e. $C_J = C_G$. In this case, it is possible to establish the following two propositions:

Proposition 1A: When $C_J = C_G$, there does not exist a PBE.

Proposition 1B: When $C_J = C_G$, all Nash Equilibria of the game yield a payoff of $-C_G$ to player S and $-D$ to player R and involve "strict" pre-emption, i.e. S attacks even before R attempts to acquire.

The proofs of these propositions may not be of interest to the casual reader, so I have relegated them to the Appendix.

Turning now to the case in which $C_J < C_G$, we find that a set of PBEs comes into existence when C_J is low enough. This is Proposition 2:

Proposition 2A: When $C_J > \frac{C_G - \theta_0 F}{1-\theta_0}$, there are no PBEs, and all Nash Equilibria yield a payoff of $-C_G$ to player S and $-D$ to player R, with "strict" pre-emption, as before.

Proposition 2B: When $C_J < \frac{C_G - \theta_0 F}{1-\theta_0}$, there exists a set of PBEs, in which R attains his maximum possible payoff, $\theta_0 B - [c_B + (1-\theta_0)D]$, and S attains the payoff $-[\theta_0 F + (1-\theta_0)C_J]$.

The proof of Proposition 2B is by construction. Again, I leave the details to the Appendix, but reproduce here a representative pair of strategies that form a PBE in this game:

R plays Acquire in each period with probability σ, where σ is "small" (the precise meaning of "small" is clarified in the Appendix).

S plays $Wait$ whenever the signal $s_t = 0$ and attacks when $s_t = 1$.

Essentially, this works because when S observes a negative signal ($s_t = 0$), he is tempted to wait rather than attack, in the hope that today's signal is not a false negative, and that when R does attempt to acquire the signal will be positive and S will not have to bear the full cost of unilateral action. In turn, this creates an opening for R to successfully acquire by keeping the acquisition probability low in each period and deterring a preventive strike by S. However, this kind of "creeping towards acquisition" strategy will not work when the extent of cost-sharing is limited, because in that case it becomes harder to convince S to stay his hand in the current period.

Propositions 1 and 2 taken together seem remarkable: starting from a situation where $C_J = (C_G - \theta_0 F)/(1-\theta_0)$, even a tiny increment in cost-sharing allows R to go from getting $-D$ to getting his maximum payoff. As for S, it is easily verified that his expected payoff in this new equilibrium is strictly greater than $-C_G$, so that he is also better off than before.

A Pre-Game Option

The analysis so far is interesting, but incomplete with respect to the intuition we originally began with. We expected to find that the absence of cost-sharing would allow for deterrence, and that the introduction of cost-sharing would produce a strictly inferior outcome for S. We have found instead that when cost-sharing is limited, the only possible Nash outcome of the game is for S to attack R for sure. Further, a more significant amount of cost-sharing allows R to acquire but still produces a better outcome (in terms of payoffs) for S. There is a missing element to the story: after all, the virtue of credible threats is that they need never be carried out, but the previous setting does not feature credible deterrence, even in the absence of cost-sharing. I now try to supply the missing element.

Suppose, then, that before the game is played, R can choose to credibly renounce the option of acquisition once and for all (perhaps by permanently allowing weapons inspectors into his country). Assume for simplicity that this renunciation is costless (actually, the condition I require is that the cost of renunciation is no greater than D). If R renounces, the game ends and each player receives a payoff of 0. If R refuses to renounce, the players then play the game described before.

It is now easy to see that in the absence of cost-sharing, R will strictly prefer to renounce his option of acquiring the capability because the only equilibrium outcome he can hope to get by playing the subsequent game is $-D$. When cost-sharing is introduced and exceeds a certain level, R will strictly prefer to retain the nuclear option, because he can attain the payoff $\theta_0 B - [c_B + (1-\theta_0)D]$ in the subsequent game.

The reader may wonder whether the same overall result may be obtained without 'cost-sharing' - that is, suppose instead that S finds it less costly to attack when the signal is positive (perhaps because there is less loss of international goodwill in this case), so that C_J is less than C_G, even though there is no 'cost-sharing'. The results above show that as long as C_J is not too different from C_G there is no effect on deterrence. If C_J were small enough without cost-sharing, however, then the analysis here would not be particularly interesting, since there would be no deterrence to begin with. Therefore, our assumption throughout will be that in the absence of cost-sharing, C_J is not small enough to destroy deterrence.

There is a small wrinkle in the above arguments, however: I showed that with cost-sharing it is possible for R to achieve his maximum payoff in a PBE, but I did not characterize all the possible equilibria in the game. In particular, suppose that there also exist equilibria in which R obtains a negative payoff - is it possible that R may choose to stay out because of uncertainty about which particular equilibrium will be played in the subsequent game? To answer this question, we need a theory of equilibrium selection. In this particular instance, one can appeal to an argument in the spirit of forward induction (see for example, Ben-Porath and Dekel 1988 and Van Damme 1989): if R chooses to stay in the game, then S can only conclude that R has chosen to stay in because he expects to play the PBE with the high payoff. By choosing to stay in, therefore, R signals which equilibrium will be played in the subsequent game, and this equilibrium must indeed be played by rational players. This resolves the wrinkle of multiple equilibria.

One last point: In light of the fact that R will definitely choose to acquire if he retains the option, it may appear suboptimal for the courts to refrain from intervening unless the signal is positive. In fact, one could even argue that the courts should intervene at once, since R has signaled his commitment to acquiring by staying in. However, doing so may result in the punishment of someone for a crime he is yet to commit. As long as the courts are committed to the principle that this is unacceptable (enough), they will only respond to positive signals.

Optimal Cost-Sharing

The model developed here suggests that the benefits of cost-sharing are discontinuous. From the perspective of S, he is indifferent to the level of cost-sharing as long as it is below the threshold level $C_J = (C_G - \theta_0 F)/(1 - \theta_0)$. If cost-sharing above this level is unavoidable, S should prefer more to less, because his payoff in this case is $-[\theta_0 F + (1 - \theta_0)C_J]$. Maximal cost-sharing is therefore the second-best solution for S. This conclusion needs to be modified once we allow for discounting. I sketch the main argument here.

Suppose, to keep the algebra simple, that S is infinitely patient, but that R is impatient, and discounts future payoffs by the factor δ (< 1). Denote by x the maximal payoff $\theta_0 B - [c_B + (1 - \theta_0)D]$ and suppose that R belongs to a continuum of types, with different values of x.

When R discounts future payoffs, the PBE suggested earlier still stands, and R does attain the payoff x, but *when* he expects to attain the payoff (and therefore the magnitude of the discounted payoff) depends on how slowly (or quickly) he creeps towards acquisition. In turn, the 'rate of creep', represented by the per-

period probability σ is constrained by the extent of cost-sharing. If cost-sharing is minimal, σ must be very small to avoid inducing S to attack now, which implies a low (but still positive) discounted payoff. Denote by $\overline{\sigma}(C_J)$ the maximum possible rate of creep, the notation making it clear that this maximum value is a function of C_J. As argued, it must be that $\overline{\sigma}'(\cdot) < 0$. It is easily verified that R's discounted payoff is given by:

$$\pi^R(x, C_J) = \frac{\overline{\sigma}(C_J)}{1 - (1 - \overline{\sigma}(C_J)\delta)} x$$

Suppose now that retaining the nuclear option is not costless for R, perhaps because it invites economic sanctions - let this cost be denoted by c_R. It follows that only those types of R for whom $\pi^R(x, C_J) > c_R$ will choose not to renounce their nuclear options. This defines a threshold value of x, which we will denote by $x^*(C_J)$, below which R will choose to stay out.

The overall payoff to S can now be written as a function of C_J:

$$\pi^S(C_J, x) = \Pr(x < x^*(C_J)).0 + [1 - \Pr(x < x^*(C_J))].[-\theta_0 F - (1 - \theta_0)C_J]$$

The ex-ante profit function confirms that limiting the extent of cost-sharing is beneficial in terms of reducing the probability of "entry" (i.e. non-renunciation) of R, but that conditional on entry, more cost-sharing is preferable to less. Depending on the distribution of types, this ex-ante profit may be maximized at an interior value of C_J in the interval $[0, \frac{C_G - \theta_0 F}{1 - \theta_0}]$, i.e. maximal cost-sharing may not be optimal for S (as a second-best solution).

On a different note, I remarked earlier that allowing for concealment to have a variable cost may change the results in an interesting way. Although I have not worked it out algebraically in this paper, the essential intuition is that increasing the level of concealment (i.e. varying θ_0) presents a trade-off for R: on the one hand, it will reduce the probability of detection, but on the other hand (in addition to raising the overall cost of acquisition) it will also make S more unwilling to wait, thereby lowering the rate of acquisition, $\overline{\sigma}$. Depending on the exact specification of the cost of concealment, the optimal level of concealment may be less than maximal.

Finally, what is the optimal level of cost-sharing from a social perspective? On the one hand, keeping cost-sharing low deters entry by R and this is certainly beneficial; on the other hand, conditional on entry, cost-sharing can reduce the probability of a certain kind of Type I error, namely the punishment of R for a crime he is yet to attempt (even though the fact of his entry may signal his intention to attempt the crime). The optimal policy in this setting hinges on whether R indeed has a credible way of signaling his intentions before the game. If there exists no such mechanism, it may be necessary to protect him from wrongful punishment by setting the level of cost-sharing so as to disable the extra-legal actor. In the next subsection, I discuss what this may entail, starting with a situation in which it is possible to vary the rate of intervention of the courts. I show that changing the rate of intervention is a second-best solution - controlling the extent of cost-sharing is a better solution.

Optimal Standard of Proof In the context of multilateral versus unilateral interventions, it makes sense to think of the "optimal" extent of cost-sharing. In terms of our example of the courts versus extra-legal detention programs, however, the variable of interest is not the level of cost-sharing, but rather the willingness of the courts to intervene. It should seem intuitive that effectively undermining the extra-legal actor may require that the courts themselves adopt a greater willingness to prosecute suspected offenders, which may actually entail an increase in the number of wrongful convictions handed down by the courts. However, formalizing this argument is somewhat difficult within the framework of the model we have been analyzing. Instead, consider a one-shot version of the original game, in which we will assume that a fraction $(1-\phi)$ of the 'opponents' of S are law-abiding and will never attempt to acquire. The remaining fraction play the following game with S as drawn in Figure 2.

In this one-shot game, R has one opportunity only to acquire.[3] This game has only one PBE:

R plays Acquire with probability $\sigma^R = \frac{1}{\phi}\frac{C_G}{C_G+\theta(F-C_G)}$
S plays Attack if $s=1$
S plays Attack with probability $\sigma^S = 1-\frac{c_B}{\theta_0 B-(1-\theta_0)D}$ *if* $s=0$
S has posterior beliefs $\Pr(Acquire|s=1)=1$, $\Pr(Acquire|s=0)=\frac{C_G}{F}$

In this PBE, the probabilities of Type I and Type II errors are given by:

$$\Pr(\text{Type I error}) = \Pr(\textit{Attack}|\textit{Not Acquire}).\,\Pr(\textit{Not Acquire}) = \sigma^s[\phi(1-\sigma^R)+(1-\phi)] \tag{1}$$

$$\Pr(\text{Type II error}) = \Pr(\textit{Wait}|\textit{Acquire}).\Pr(\textit{Acquire}) = (1-\sigma^s)\sigma^R\theta_0\phi \tag{2}$$

[3] As noted earlier, preventive action naturally creates incentives for the would-be criminal to (a) conceal his actions and (b) conceal the timing of his attempt. The current formulation sacrifices the second element.

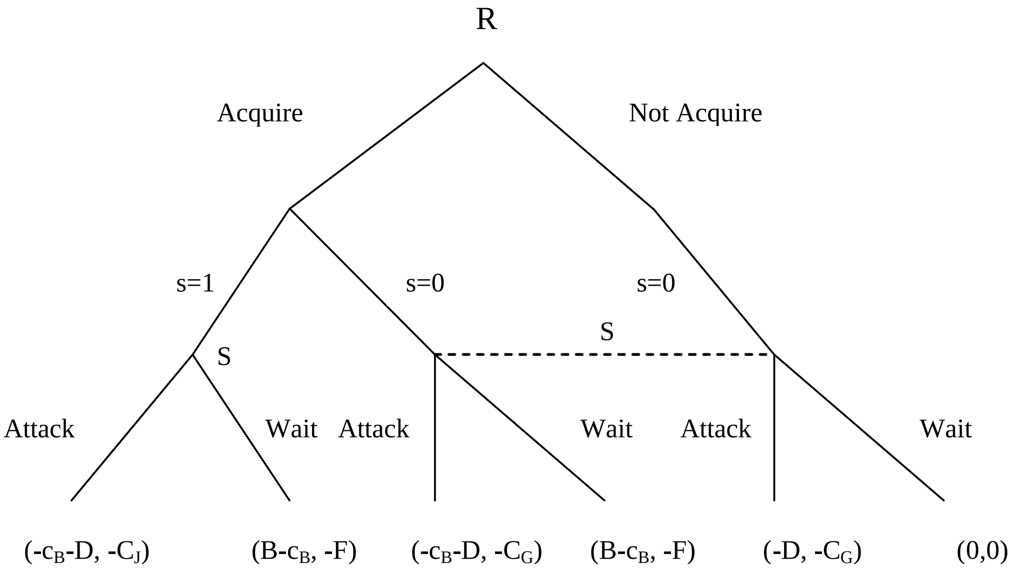

Figure 2: One shot-version of the game

I should emphasize that these are the *unconditional* probabilities of Type I and II errors, as distinct from the probabilities conditional on the actions of R, which are simply given by σ^S and $1-\sigma^S$.

I now introduce the courts. Suppose that, as before, the court will prosecute with probability 1 if $s=1$. In addition, suppose also that even if $s=0$ the courts will be ready to try R and will punish him with probability λ_0 if he is actually innocent and with probability λ_1 if he is guilty. However, once the courts have found R innocent, S can no longer take any action against him. I leave the process of determination of λ_0 and λ_1 unmodeled - one possibility is that when a case comes to court, the defendant is allowed to present some specific additional evidence (over and above the already observed signal) in his defence, and that (i) a fraction λ_0 of the innocent individuals are unable to present such evidence and (ii) a fraction $(1-\lambda_1)$ of the guilty individuals are able to present such evidence. Finally, I assume also that $C_J = 0,$ so that if the court intervenes, S will not have to incur any cost (i.e. maximal cost-sharing). This allows us to focus on the probability of the court's intervention.

The game is now as follows: After observing the signal, S can either choose to attack on his own, incurring the cost C_G, or can leave the matter to the court and incur no cost of his own. It is possible to show that in this case, the PBE changes to:

R plays Acquire with probability $\sigma^R = \frac{1}{\phi}\frac{C_G}{C_G+\theta_0[(1-\lambda_1)F-C_G)]}$
S plays Attack if $s = 1$
S plays Attack with probability $\sigma^S = 1 - \frac{c_B}{(1-\lambda_1)[\theta_0 B-(\frac{1-\lambda_0}{1-\lambda_1}-\theta_0)D]}$ *if* $s = 0$
S has posterior beliefs $\Pr(Acquire|s = 1) = 1$, $\Pr(Acquire|s = 0) = \frac{C_G}{(1-\lambda_1)F}$

It is easy to verify that σ^S is smaller and σ^R is now higher than before. Again, this is familiar - the option of leaving matters to the courts reduces the level of deterrence. This result is worth contrasting with that of a more 'standard' case: In a setting with no extra-legal actors, increasing the probability of indictment conditional on being guilty (λ_1) while holding fixed the probability of indictment conditional on being innocent (λ_0) should reduce the rate of criminality. In contrast, we have shown that when extra-legal actors are present, increasing λ_1 while holding λ_0 fixed actually *encourages* criminality by reducing the deterrent ability of the extra-legal actor. What does this imply for unconditional Type I and II errors in our model? The probabilities of Type I and II errors are given by:

$$\begin{aligned}\Pr(\text{Type I error}) &= [\sigma^s + (1-\sigma^S)\lambda_0][\phi(1-\sigma^R) + (1-\phi)] \qquad (3)\\ &= [1 - \frac{c_B}{\theta_0 B - (\frac{1-\lambda_0}{1-\lambda_1} - \theta_0)D}.\frac{1-\lambda_0}{1-\lambda_1}][\phi(1-\sigma^R) + (1-\phi)]\end{aligned}$$

$$\begin{aligned}\Pr(\text{Type II error}) &= (1-\sigma^s)(1-\lambda_1)\sigma^R\theta_0\phi \qquad (4)\\ &= \frac{c_B}{\theta_0 B - (\frac{1-\lambda_0}{1-\lambda_1} - \theta_0)D}.\sigma^R\theta_0\phi\end{aligned}$$

If we assume $\lambda_0 < \lambda_1$, the net effect on Type I error is ambiguous - on the one hand, the propensity to acquire is higher, but on the other hand, the probability of punishment conditional on acquiring is lower because S now attacks with lower probability. Marginal effects are however unambiguous. Holding λ_0 fixed, an increase in λ_1 unambiguously reduces (increases) both the conditional as well as the unconditional probability of Type I (Type II) error.

That increasing λ_1 actually *increases* the probability of Type II error is due to its effect on (i) the propensity of S to attack R and thereby its effect on (ii) the rate of criminality, a result that we have noted as being opposite to what we would expect in a setting without extra-legal actors. The effect on probabilities of Type I error deserves a little more explanation. This result is partially intuitive and partially an artifact of our particular setup: An increase in λ_1 reduces the propensity of the extra-legal actor to punish innocent people and this accounts for the reduction in the probability of punishment conditional on being innocent - this is the sense in which the presence of the courts can moderate the extent of Type I errors committed by S. The reduction in the unconditional probability of Type I error is however magnified by the assumption that everyone is brought to court, because a greater rate of criminality implies a lower rate of innocence, and therefore a smaller number of cases of Type I error.

More realistically, though, it may not be feasible to increase λ_1 without at the same time increasing λ_0. For concreteness we will focus on the algebraically simple case where $\lambda_1 = \lambda_0 = \lambda$. We will not have much to say about the optimal level of λ from the perspective of S, because this is relatively trivial - because S is not concerned about Type I errors, he will prefer a λ that is high enough to completely deter acquisition. More interesting is the optimal choice of λ from the perspective of the courts. To highlight the basic point, we will assume that the courts strictly prefer to have zero Type I errors, so that they will always choose $\lambda = 0$ in the absence of S. [4]

It will be important to distinguish a couple of different cases in the subsequent analysis. To this end, we define λ^S and λ^R :

$$\lambda^S = 1 - C_G/F$$

$$\lambda^R = 1 - \frac{c_B}{\theta_0 B - (1-\theta_0)D}$$

λ^S is the threshold probability of intervention by the courts such that S is just indifferent between attacking and waiting; λ^R is the threshold probability of being attacked that will make R just indifferent about attempting to acquire. Notice that because the PBEs that we derived for this game involve mixed strategies, the probability of being attacked (by either S or by the courts) is exactly this threshold probability, λ^R. As the probability of intervention by the court, λ, changes, the

[4] To put it more precisely, one could say that the courts have lexicographic preferences over Type I and Type II errors, with the former being accorded greater priority.

probability of attack by S changes by just enough to keep the combined probability of attack at λ^R. In what follows, we will assume that $\lambda^R > \lambda^S$.

To derive the optimal λ for the courts, we will graph the probability of Type I error as a function of λ. When S is not present, this probability is:

$$\begin{aligned} \Pr(\text{Type I error}) &= \lambda(1-\phi) \text{ if } \lambda < \lambda^R \\ &= \lambda \text{ if } \lambda \geq \lambda^R \end{aligned} \qquad (\mathbf{A})$$

When S is present, this probability becomes:

$$\begin{aligned} \Pr(\text{Type I error}) &= [1 - \frac{c_B}{\theta_0 B - (1-\theta_0)D}.][\phi(1-\sigma^R) + (1-\phi)] \text{ if } \lambda < \lambda^S (\mathbf{B}) \\ &= \lambda(1-\phi) \text{ if } \lambda^S \leq \lambda < \lambda^R \\ &= \lambda \text{ if } \lambda \geq \lambda^R \end{aligned}$$

The functions **A** and **B** are graphed in Figure 3.[5] In the absence of S, the courts can reduce the probability of Type I error to 0 by setting $\lambda = 0$. When S is present, the best the courts can do is to set $\lambda = \lambda^S$, and thereby reduce the probability of error to $\lambda^S(1-\phi)$.

To put the matter in terms familiar to economists, the problem facing the courts is one of too few instruments: λ is doing double-duty, being used to deter the would-be deterrer as well as to directly reduce Type I errors. Putting it in this way clarifies that setting λ so as to undermine the extra-legal actor can only be a second-best solution. A first-best solution would involve devising an altogether different instrument to control the extra-legal actor. A natural candidate for such an instrument is the level of cost-sharing, represented by C_J. In this section, we have set C_J at zero, and it is difficult to envisage the possibility of improving on this by making it negative (i.e. effectively rewarding S each time the courts take action against R). A more sensible alternative may be to devise sanctions such that C_G, the cost of unilateral action, is increased. For instance, imposing sanctions on S such that C_G becomes greater than F would guarantee that S never acts on his own, at which point the courts can then set $\lambda = 0$, and thus obtain the first-best solution. In the two examples we have been discussing, namely (a) unilateral action against rogue nations and (b) extra-legal detention, the extra-legal actor was difficult or infeasible to punish. The foregoing analysis suggests that in these cases the first-best may not be attainable by the courts.

[5] Although neither function is continuous, we have drawn them as such in order to make the figure more readable.

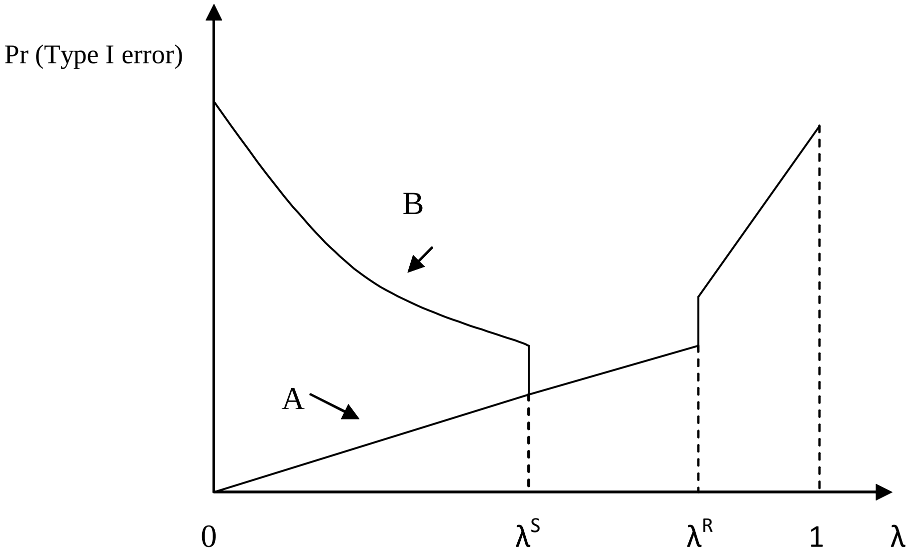

Figure 3: Probability of Type I error in the two regimes

Finally, as before, if we allow R to credibly signal his intentions before the game begins, the implications for the optimal standard of proof are different. In this case, the optimal policy for the courts may be to keep the deterrent power of S intact by keeping cost-sharing and λ low. We have already made this point in the context of cost-sharing, so I do not formalize it here.

Concluding Remarks

The presence of extra-legal actors who take unilateral (or unsanctioned) preventive action against would-be criminals has focused attention on the role of courts and multilateral institutions. This paper has used game-theoretic reasoning to make three points: (1) The deterrent power of extra-legal actors depends on the cost of unilateral action (relative to multilateral/ officially sanctioned action) and the standards of proof of the courts, (2) When there exists a mechanism by which individuals can credibly signal their intention to not commit crimes, it may be optimal for the courts to keep the deterrent power of extra-legal actors intact, and

(3) When there exists no such mechanism, the optimal strategy of the courts may be to control the extra-legal actor by changing the relative cost of unilateral action. But if the latter cannot be varied by the courts, the second-best solution would be to vary the standard of proof. In particular, lower standards of proof may be required in order to "deactivate" the extra-legal actor and thereby reduce Type I errors.

The ability of courts to moderate the influence of extra-legal actors implies that the latter may have an interest in dismantling or undermining the former. Relatedly, one may observe the incidence of crime to be lowest when the rule of law is weak, because this is when the deterrent ability of extra-legal actors is maximal.

References

Ben-Porath, E. and Dekel, E., 1992. Coordination and the potential for self-sacrifice. *Journal of Economic Theory* 57, pp. 36–51.

van Damme, E., 1989. Stable Equilibria and Forward Induction. *Journal of Economic Theory* 48, pp. 476-496.

Kaplow, L. and Steven M. Shavell, 1994. Accuracy in the Determination of Liability. *Journal of Law and Economics* 37:1, pp. 1-15.

Lando, H., 2009. Prevention of Crime and the Optimal Standard of Proof in Criminal Law. *Review of Law and Economics* 5:1.

Proofs

Proofs of Propositions 1 and 2:

Denote by σ_t^R the probability that R attempts to acquire in period t, and by σ_t^S the probability in period t that when S observes s_t= 0, he attacks. We can then write the strategies for S and R as $\{\sigma_j^S\}_1^\infty$ and $\{\sigma_j^R\}_1^\infty$ respectively. The value of the subgame that begins at time t can now be written for each of the two players in terms of these strategies:

$$\begin{aligned} V_t^R &= \sigma_t^R[\theta_0(\sigma_t^S(-D) + (1-\sigma_t^S).B) + (1-\theta_0)(-D) - c_B] + (1-\sigma_t^R)[(1-\sigma_t^S)V_{t+1}^R + \sigma_t^R(-D)] \\ &= \sigma_t^R[-\sigma_t^S(\theta_0 B + D) + MV] + (1-\sigma_t^R)[-\sigma^S D + (1-\sigma_t^S)V_{t+1}^R] \end{aligned}$$

where $MV = \theta_0 B - [c_B + (1-\theta_0)D]$, R's highest possible expected payoff in the game, and

$$\begin{aligned} V_t^S &= \sigma_t^R[\theta_0(\sigma_t^S(-C_G) + (1-\sigma_t^S).(-F)) + (1-\theta_0)(-C_J)] + (1-\sigma_t^R)[(1-\sigma_t^S)V_{t+1}^S + \sigma_t^R(-C_G)] \\ &= \sigma_t^S[\theta_0\sigma_t^R(-C_G) + (1-\sigma_t^R)(-C_G)] + (1-\sigma_t^S)[\theta_0\sigma_t^R(-F) + (1-\sigma_t^R)V_{t+1}^S] + \sigma_t^R(1-\theta_0)(-C_J) \end{aligned}$$

These expressions will be key to proving the results. To simplify them further, we first note that irrespective of the level of C_J, it can never happen in a PBE that $\sigma_t^R = 1$, for this would in turn invite $\sigma_t^S = 1$, but then it would not be optimal for R to set σ_t^R to 1. This implies that at time t, R is either indifferent between playing $Acquire$ and $Not\ Acquire$ or strictly prefers to not acquire. In either case, he will be getting the payoff from playing $Not\ Acquire$ (because if he is indifferent, the expected payoffs from the two actions must be equal). The upshot of this is that we can always write:

$$V_t^R = -\sigma^S D + (1-\sigma_t^S)V_{t+1}^R$$

Iterating this forward, we have:

$$V_t^R = (1-\varphi_{kt})(-D) + \varphi_{kt}V_{t+k}^R$$

where $\varphi_{kt} = \prod_{i=0}^{i=k-1}(1-\sigma_{t+i}^S)$.

Taking limits as $k \to \infty$, it follows that if the sequence $\{\sigma_j^S\}_1^\infty$ is bounded away from zero, then the value of every subgame is $-D$ for R, since V_{t+k}^R is bounded. The nature of the sequence $\{\sigma_j^S\}_1^\infty$ in turn depends on the level of C_J.

Case 1: Suppose that C_J is such that $-[\theta_0 F + (1-\theta_0)C_J] < -C_G$ (this includes the special case $C_J = C_G$). In this case, we first note that S can never set $\sigma_t^S = 0$ in a PBE, because that would invite R to set $\sigma_t^R = 1$ and S would get a payoff of $-[\theta_0 F + (1-\theta_0)C_J]$ which is worse than his payoff from attacking, $-C_G$. By a similar argument as before, it follows that the payoff of S in period t will always be the payoff from choosing to play $Attack$. Thus:

$$V_t^S = -[\sigma_t^R(\theta_0 C_G + 1 - \theta_0)C_J) + (1-\sigma_t^R)C_G] \tag{5}$$

Suppose now that at time t, S prefers playing $Attack$ to $Wait$, i.e. $\sigma_t^S = 1$.Then it must be that:

$$\frac{\theta_0\sigma_t^R}{\theta_0\sigma_t^R+(1-\sigma_t^R)}(-C_G)+\frac{(1-\sigma_t^R)}{\theta_0\sigma_t^R+(1-\sigma_t^R)}(-C_G) > \frac{\theta_0\sigma_t^R}{\theta_0\sigma_t^R+(1-\sigma_t^R)}(-F)+\frac{(1-\sigma_t^R)}{\theta_0\sigma_t^R+(1-\sigma_t^R)}V_{t+1}^S$$

But if $\sigma_t^S = 1$ then we must also have $\sigma_t^R = 0$, which implies that $V_{t+1}^S < -C_G$. But this can never happen in equilibrium, because S can always guarantee himself at least $-C_G$ in every subgame by playing $Attack$ at the first opportunity. This establishes that in a PBE, we cannot have $\sigma_t^S = 1$ (although, I note for later reference that this can happen in a Nash Equilibrium, because playing $Attack$ for sure today renders tomorrow's payoff irrelevant).

Therefore S must be indifferent between playing $Attack$ and $Wait$ at each point in time. Then we must have at time t:

$$\theta_0\sigma_t^R(-C_G) + (1-\sigma_t^R)(-C_G) = \theta_0\sigma_t^R(-F) + (1-\sigma_t^R)V_{t+1}^S$$

Plugging this into Eqn (5) and iterating forward, we have:

$$V_t^S = (1-\psi_{tk})[-\theta_0 F - (1-\theta_0)C_J] + \psi_{tk}V_{t+k}^S$$

where $\psi_{tk} = \prod_{i=0}^{i=k-1}(1-\sigma_{t+i}^R)$.

Recalling that $-[\theta_0 F + (1-\theta_0)C_J] < -C_G$, it follows that necessary for V_t^S to be at least as good as $-C_G$ is that in the limit as $k \to \infty$, V_{t+k}^S should be strictly better than $-C_G$ and that σ_{t+k}^R should go to 0. However, an examination of Eqn (5) shows that if the latter condition holds, then V_{t+k}^S converges to $-C_G$, so that $V_t^S \leq -C_G$. $V_t^S = -C_G$ can only happen if R will never attempt acquisition, in which case S should never attack, and the equilibrium unravels. As noted earlier, we cannot have $V_t^S < -C_G$, because S can always guarantee himself at least $-C_G$ in every subgame (by playing $Attack$ at the first opportunity). It follows that S cannot possibly be indifferent between playing $Attack$ and $Wait$ in a PBE as long as $-[\theta_0 F + (1-\theta_0)C_J] < -C_G$. This contradiction establishes that there are no PBEs when $-[\theta_0 F + (1-\theta_0)C_J] < -C_G$.

The difference between Nash Equilibria and PBEs in this game is that if either σ_t^S or σ_t^R is equal to 1, implying that the game ends for sure after period t, the strategies for the two players in the rest of the game are completely unrestricted in a Nash Equilibrium, but continue to be restricted to be mutually consistent in a PBE.

Consider the subgame starting at $t = 1$, i.e. the original game. We will for the moment restrict ourselves to considering only those subgames that are reached with positive probability. At any t which may be reached with positive probability, it must still be the case that $\sigma_t^S > 0$ and $\sigma_t^R < 1$. Furthermore, the argument we used before still goes through - there cannot be a Nash Equilibrium in which $\sigma_t^S < 1$ for all t. It follows that a Nash Equilibrium must have $\sigma_t^S = 1$ at some t. Suppose t^* is the first such point. Because $\sigma_t^R < 1$ up to this point, we can be sure that subgames prior to this one will be reached with positive probability and those that follow after this will never be reached. Because t^* will be reached with positive probability and $\sigma_t^S = 1$, we must have $\sigma_{t^*}^R = 0$. Now consider S's decision at t^*-1. At this point, the payoff to playing $Wait$ is $\frac{\theta_0\sigma_{t^*-1}^R}{\theta_0\sigma_{t^*-1}^R+(1-\sigma_{t^*-1}^R)}(-F) + \frac{(1-\sigma_{t_*-1}^R)}{\theta_0\sigma_{t^*-1}^R+(1-\sigma_{t^*-1}^R)}(-C_G)$ which can only be equal to the payoff from attacking, $-C_G$, if $\sigma_{t^*-1}^R = 0$. This argument applies all the way back to $t = 1$. We have shown therefore that a generic Nash Equilibrium has strategies such that S attacks for sure at some point, and that R does not attempt to acquire while the game is in progress (to be sure, S's randomization until t^* should place enough weight on $Attack$ to discourage R from trying to acquire). This implies that the Nash Equilibrium payoffs are $-C_G$ and $-D$ for S and R respectively. For example, it is readily confirmed that the following strategies form a Nash Equilibrium:

$$\begin{aligned} \sigma_t^R &= 0 \text{ if } t \leq t^* \\ &= 1 \text{ otherwise} \end{aligned}$$

$$\begin{aligned} \sigma_t^S &= 1-\varepsilon \text{ if } t < t^* \\ &= 1 \text{ if } t = t^* \\ &= 0 \text{ otherwise} \end{aligned}$$

where ε is arbitrarily small. We have thus established Propositions 1A, 1B and 2A.

CASE 2: When $-[\theta_0 F + (1-\theta_0)C_J] \leq -C_G$, there exists a set of PBEs, in which R attains his maximum possible payoff, $\theta_0 B - [c_B + (1-\theta_0)D]$, and S attains the payoff $-[\theta_0 F + (1-\theta_0)C_J]$. This part of the proof is by construction. We will show that the following is a generic PBE:

R plays $Acquire$ in each period t with probability $\sigma_t^R < \frac{C_G - [\theta_0 F + (1-\theta_0)C_J]}{(1-\theta_0)(F-C_J)}$ such that σ_t^R is bounded away from zero.

S plays $Wait$ whenever $s_t = 0$ and $Attack$ whenever $s_t = 1$.

S has the posterior $\Pr(Acquire|s_t = 0) = \frac{\theta_0 \sigma_t^R}{\theta_0 \sigma_t^R + (1-\sigma_t^R)}$

Given S's strategy, R is indifferent between playing $Acquire$ today and playing $Acquire$ later (because he is infinitely patient). Therefore he is happy to randomize at each point in time. Given R's strategy, when S sees $s_t = 0$, he forms his belief using Bayes' Rule:

$$\begin{aligned} \Pr(Acquire|s_t &= 0) = \frac{\Pr(s_t = 0|Acquire)\Pr(Acquire)}{\Pr(s_t = 0|Acquire)\Pr(Acquire) + \Pr(s_t = 0|Not\ Acquire)\Pr(Not\ Acquire)} \\ &= \frac{\theta_0 \sigma_t^R}{\theta_0 \sigma_t^R + (1-\sigma_t^R)} \end{aligned}$$

Given this posterior belief, the value of waiting (instead of attacking) is $\Pr(Acquire|s_t = 0).(-F)$ $+(1-\Pr(Acquire|s_t = 0))V_t^S$ where V_t^S is the continuation payoff. Given the strategies of

R and S it is clear that R will eventually attempt acquisition and that $V_t^S = -[\theta_0 F + (1-\theta_0)C_J]$. The value of attacking is given by $-C_G$. Some algebra shows that S strictly prefers to wait as long as $\sigma_t^R < \frac{C_G - [\theta_0 F + (1-\theta_0)C_J]}{(1-\theta_0)(F - C_J)}$. Thus, we have verified that the strategies and beliefs above constitute a PBE. This completes the proof of Proposition 2B.