International Journal of Industrial Organization
21 (2003) 347–379

International Journal of Industrial Organization

www.elsevier.com/locate/econbase

Leniency programs and cartel prosecution

Massimo Motta[a], Michele Polo[b,*]

[a]*European University Institute, Florence Universitat Pompeu Fabra, Barcelona, Spain*
[b]*University of Sassari and IGIER, Via Salasco 5, 20136 Milan, Italy*

Received 28 September 2001; accepted 27 March 2002

Abstract

We study the enforcement of competition policy against collusion under leniency programs, which give reduced fines to firms that reveal information to the Antitrust Authority. Leniency programs make enforcement more effective but they may also induce collusion, since they decrease the expected cost of misbehaviour. We show that in the optimal policy the former effect dominates, calling for leniency programs when the Antitrust Authority has limited resources. We also show that these programs should apply to firms that reveal information even after an investigation is started.

JEL classification: K21; K42; L41

Keywords: Collusion; Antitrust; Self-reporting; Amnesty; Optimal law enforcement

1. Introduction

The enforcement of competition policy against collusion and price fixing agreements is one of the main fields of antitrust intervention. In the design of the policy we find today richer and more complex mechanisms than those based simply on an increase in fines. Since 1978 the US Antitrust Division of the Department of Justice has allowed for the possibility of avoiding criminal

*Corresponding author.
E-mail address: michele.polo@uni-bocconi.it (M. Polo).

0167-7187/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0167-7187(02)00057-7

sanctions if specific conditions occurred. In 1993 this policy was redesigned in the *Corporate Leniency Policy*, which establishes that criminal sanctions can be avoided in two cases: either if a colluding firm reveals information before an investigation is opened, as it was in the previous regime, or if the Division has not yet been able to prove collusion when a firm decides to cooperate.[1] The new leniency policy has shown in the first years of application a significant success in terms of the number of cases that the Division has been able to open and successfully conclude.

The European Union introduced in 1996 a new regulation in which more generous fine reductions can be given to firms which cooperate with the antitrust authority *before* an inquiry is opened, by providing evidence of a collusive agreement in which they have been involved, while limited reductions can be granted if cooperation occurs after the opening of a case.[2]

Although it is too early to evaluate the effects of this new policy, this paper argues that—as also the US experience suggests—this leniency program might lead to more effective enforcement against cartels. However, our analysis also indicates that the design of a leniency program might be improved by extending it to cover situations where firms reveal information even after an investigation has started.

More generally, the objective of this paper is to investigate the deterrence (abstain from collusion) and desistence (comply with the authority for a certain period if found guilty) properties of a leniency program[3] in antitrust cases.[4] The Antitrust Authority is motivated by the maximization of social welfare and aims at minimising the occurrence of collusion among firms by committing on a set of policy parameters which includes full and (possibly) reduced fines and on an allocation of internal resources that determines the probability that a cartel is reviewed and the probability that it is found guilty. After observing the Antitrust Authority's decisions, firms play an infinitely repeated game where they decide first whether they want to deviate or collude, and then whether they want to reveal information about the cartel to the Authority or not. In particular, we consider two

[1] Some additional restrictions on the firms entitled to benefit from this regime are introduced, as the fact that only the first can be given a fine reduction, and that it must be a junior partner in the cartel.

[2] See European Union (1996). To be more precise, a 75–100% reduction in fines can be given if firms reveal information before an inquiry is opened; a lower reduction (50–75%) can be granted if cooperation occurs after an investigation has started, *but that investigation has failed to provide sufficient grounds for initiating a procedure leading to a decision*; a 10–50% reduction in fines can be given for partial cooperation, such as providing additional evidence or not contesting the facts on which the Commission bases its allegations. Notice that while in the US the regime applies to criminal sanctions (which include both fines and incarceration), in the EU reductions are referred only to monetary fines. Criminal sanctions do not exist under EU competition law.

[3] In this paper we use the term leniency programs to refer to a reduction in monetary fines.

[4] Since our work, other papers on the use of leniency programs in antitrust have been written. See Rey (2000) for an excellent survey.

possible collusive strategies. One prescribes firms *never to reveal* (if one firm revealed, it would trigger Nash reversal forever) and to go back to collusion after a possible condemnation, the other requires firms *to reveal* to the Authority as soon as an investigation review is opened, and to go back to collusion after the procedure is closed.

In this setting we show that a leniency program might have two possible effects, depending on the policy parameters chosen. The positive effect is that it might lead firms to desist ex-post from colluding more frequently. If it convinces firms to reveal information whenever an investigation is opened, society will benefit not only because of a (temporary) cessation of collusive pricing, but also because investigations are shorter (information received by firms brings about the punishment with certainty and allows to avoid the costly prosecution stage of the investigation opening sooner new cases). But the leniency program might also give rise to a perverse effect. Since it allows colluding firms to pay reduced fines, it may have ex-ante a pro-collusive effect, given that it decreases the expected cost of anticompetitive behaviour. A priori, therefore, it would be difficult to conclude that a leniency program unambiguously increases welfare, without considering which policies are implementable and desirable.

Our analysis of the optimal enforcement policies shows that, if the Antitrust Authority has limited resources, and is therefore unable to prevent collusion ex-ante, the use of leniency programs improves welfare, by sharply increasing the probability of interrupting collusive practices and by shortening the investigations. Hence, in a second best perspective, fine reductions may be desirable because they allow to better implement ex-post desistence from collusion and put saved resources to new cases.[5]

The key mechanism of leniency programs is the rule that allows firms to receive fine reductions even after an investigation is opened. In this situation, the probability of paying the fine increases compared with the case where firms are not yet under scrutiny, and the exchange of reduced fines with cooperation becomes attractive. Conversely, we prove that limiting eligibility to the case where the inquiry has not yet been opened completely eliminates, at least in the setting of our model, the incentive to reveal and the effectiveness of the program.

The enforcement problem we study has several ingredients. We analyze the design of self-reporting incentives, having a group of (and not a single) defendants and considering ongoing (and not single episode) infringements and benefits. This paper is therefore related to several strands of literature that have often considered some of these features separately. The closest to our work are perhaps the studies on optimal enforcement under self-reporting schemes. Malik (1993) and Kaplow and Shavell (1994) are probably the first to have identified the potential benefits of

[5] But if resources are very scarce, the Authority is not able to credibly prove firms guilty with a probability sufficiently high to induce revelation, and leniency programs become ineffective.

schemes which elicit self-reporting by violators.[6] Self-reporting may reduce enforcement costs[7] and improve risk-sharing, as risk-averse self-reporting individuals face a certain penalty rather than the stochastic penalty faced by non-reporting violators (who pay only if caught).[8] There are two main differences between these papers and ours. First, they consider individual violators, rather than a group of violators like our colluding firms, which requires us to analyse in a game theoretic setting the conditions for self-reporting. Second, they deal with an illegal action which is taken and gives benefits only once, whereas we analyse ongoing infringements and benefits. These two elements explain why—unlike the earlier papers on self-reporting—an optimal programme might be one which gives generous penalty discounts in case of collaboration with the Antitrust Authority. A similar result can be found in Livernois and McKenna (1999), where a repeated game is played between the regulator and a polluting firm, and where by self-reporting a firm will return to compliance, which decreases future profits.

Another strand of literature related to our paper is that on plea-bargaining, where an individual is given the option to plead guilty in exchange for a less harsh penalty rather than waiting for a court decision. Landes (1971) has showed that this allows to save the prosecution costs (a motive which appears in our paper as well), while Grossman and Katz (1983) have identified the possible beneficial insurance and screening effects of settlements.[9] In the plea bargaining literature[10] the enforcer balances the goal of condemning the guilty agents and not condemning the innocent ones with the minimisation of resources devoted to enforcement. However, the issue of deterrence is generally not addressed: agents have (possibly) already committed a crime, and in most papers, whether the agent is innocent or guilty and how strong is the evidence against him (agent's type) is exogenous. The effects of the legal procedures on preventing the crime (collusion) or making it cease are instead at the centre of our analysis. In order to focus on deterrence, we make the simplifying assumption that there are no judicial type-I errors (innocents will never be found guilty) and that firms are symmetric (either they are all guilty

[6] A recent paper in this field is Innes (1999), who considers an extension of the environmental self-reporting schemes.

[7] In Malik (1993), who applies self-reporting to environmental violations, self-reporting decreases auditing costs but increases penalty costs. It is the relative importance of the auditing and punishment technologies which determine the desirability of the scheme. Kaplow and Shavell (1994) also note that if the imposition of penalties occurs more frequently under self-reporting, administration costs may increase.

[8] See also Arlen and Kraakman (1997) who analyse the effects of different corporate liability regimes, and the incentive schemes for corporations to monitor and report wrongdoings of their employees. At the other extreme, Tokar (2000) analyses whistleblowing of employees, who under the US False Claims Act are given monetary incentives to file cases against employers which defraud the US Government.

[9] But if innocent defendants are more risk-averse than guilty defendants, the former might plead guilty even if they are not.

[10] See also Reinganum (1998) for a plea bargaining model with asymmetric information.

or they all innocent). Consequently, we cannot address the insurance value or the possible self-selection effects of leniency programs.[11]

Our work also shares some features with studies on multi-defendant settlements, where a single plaintiff faces many defendants, a literature initiated by Easterbrook et al. (1980) and Polinsky and Shavell (1981). In particular, Kornhauser and Revesz (1994) analyse the case where there exists joint and several liability and the plaintiff's probabilities of success are highly correlated across defendants. Their setting presents game theoretic aspects similar to the case we analyse, as the decision of one of many defendants between settling or not is very similar to the decision of one of many colluding firms between revealing or not information to the Antitrust Authority.[12]

Finally, the issues studied here have some relationships with the literature on tax amnesties, even if the models used in that literature are very different from our own.[13] In particular, despite the different settings, we believe that our results might represent a contribution to the understanding of the effects of fully anticipated (or permanent) tax amnesties. While they might have the effect of reducing compliance, they could still be beneficial in a second-best perspective, when the tax authorities have not enough resources to avoid tax evasion.

The paper continues as follows. In Section 2 we set up the basic model, in which every firm which decides to cooperate with the Antitrust Authority is given a fine reduction. In Section 3 the firms' decisions given the policy parameters are studied, while in Section 4 we analyze the optimal policies. Section 5 deals with the case where leniency applies only if information is disclosed before an inquiry is open; concluding remarks follow in Section 6.

2. The model

We consider a group of perfectly symmetric firms (an industry) which consider colluding taking into account the enforcement activity of the Antitrust Authority (AA from now on). Moreover, in the equilibrium analysis we shall consider symmetric industries: hence, all firms and industries will take the same (collusive or deviating) strategy in the economy. The AA is able to commit to a certain enforcement policy, which might entail the use of leniency programs (LP

[11] For the same reason, we do not have multiple equilibria with different crime rates, as identified by Schrag and Scotchmer (1997).

[12] See also Kobayashi (1992), where by offering a plea discount to one defendant the prosecutor obtains information which raises the probability of conviction of other defendants. Kobayashi shows that if the culpability of an individual is positively correlated with the amount of incriminating evidence about the other defendants, the prosecutor will offer the highest plea discount to the most culpable defendant, to maximise deterrence. While our symmetric setting does not allow us to analyse this case, this result would carry over to a properly modified extension of our model.

[13] See Andreoni (1991), Malik and Schwab (1991) and Das-Gupta and Mookherjee (1996).

hereafter). LP grant reduced fines to those firms which cooperate in the investigation by revealing information which proves the existence of a collusive agreement. The content of the collusive agreement, therefore, has to prescribe both the market conduct and the behaviour towards the AA. A cartel, for example, may prescribe to its members to replicate the monopoly configuration and to refuse any cooperation with the AA during the inquiries, or, conversely, it may allow the members to reveal information if the AA opens a review of the industry.

We now describe the policy choices of the AA, moving then to the firms' strategies and to the timing of the game.

2.1. Enforcement choices

The AA goal is the maximization of a utilitarian welfare function. Four parameters summarize the enforcement policy.

– The full fines $F \in [0,\overline{F}]$ for firms that are proved guilty and that have not cooperated with the AA, where $\overline{F}$ is exogenously given by the law.

– The reduced fines $R \in [0, F]$ specified by a LP together with the eligibility conditions.[14] We shall consider initially the benchmark case in which all[15] the firms that cooperate even after an investigation is opened can be granted reduced fines R.[16]

– The probability $\alpha \in [0, 1]$ that the firms are reviewed by the AA. (This review—or monitoring—stage is the first stage of an investigation.)

– The probability $p \in [0, 1]$ that the AA successfully concludes the investigation when firms do not cooperate. (This prosecution stage is the second and last stage of the investigation.)

When the AA is running an investigation, it is able to collect and use evidence up to the current period. Once the investigation is opened, the AA has to conclude it with a decision. We assume that the AA does not make (type I) judicial errors: if an industry where firms are not colluding is reviewed, the investigation does not

[14] Spagnolo (2000), who builds on the previous version of this paper (Motta and Polo, 1999), considers the case of negative fines (i.e., rewards) for firms which provide information to the AA. However, offering rewards for firms that have colluded is generally politically unfeasible.

[15] Throughout the paper, we assume that information given by a single firm is enough to prove that all the firms which have taken part in the collusion are guilty. This might be interpreted as the case where each firm has access to the minutes of the meetings, or has copies of letters, faxes or e-mail messages which all the firms have used to coordinate on the collusive outcome. Since an important component in the working of cartels is the coordination of moves among participants, the access of each partner to some information regarding the others seems realistic.

[16] In Motta and Polo (1999) we also consider the case where only the first comer is eligible for leniency. We find there that restricting the LP to the first firm is sub-optimal, but otherwise results are qualitatively similar.

enter the prosecution stage. A review on colluding firms can be ended in two ways: either some cartel member reveals information to the AA, in which case the participants are found guilty with probability one (and there is no need to enter the prosecution stage), or nobody reveals information. In this case the AA has to go on with the investigation, trying to prove the firms guilty, which occurs with probability p (type II errors might occur) and takes more time.

We assume that the AA has an exogenous budget that can be used to promote enforcement: hence, we have fixed rather than variable enforcement costs. The important decision regarding the policy parameters will be the allocation of internal resources, which determines the trade-off between the monitoring and prosecution rates α and p. In Section 4 we discuss in detail the enforcement technology and costs.

Moreover, when the AA proves firms guilty, it is able to impose compliance in the current period, for instance by imposing restrictions and remedies on firms' behaviour, e.g. non cooperative pricing. This temporary desistence effect of an adverse decision wants to capture the common fact that a guilty firm is often required to produce reports to the AA for a certain period on its market strategies and is subject to a light monitoring regime in that phase.

We initially treat the policy parameters as exogenous, focusing on the game played by the firms once the policy is set. When moving to the analysis of the optimal policies we will describe the constraints of the AA and explain how the policy parameters are determined.

2.2. Firms' collusive strategies

We analyse two different collusive strategy profiles of firms.

– In the first one, CR (Collude and Reveal), firms collude from $t = 1$ on, as long as no deviation occurs. If in period t no inquiry is opened, they realize a profit Π_M at the end of the period. If in period t the AA opens a review, firms reveal information to the AA, pay the (reduced) fine R and are forced to non cooperative pricing for the current period, with profits $\Pi_N < \Pi_M$. In $t + 1$, since no deviation from the equilibrium strategy occurred, they go back to the collusive strategy. If a deviation either in the marketplace or in the revelation policy occurs, they use Nash punishment forever with profits Π_N in every period.

– In the second collusive strategy, CNR (Collude and Not Reveal), firms collude from $t = 1$ on, as long as no deviation occurs. If in period t no inquiry is opened, they realize a profit Π_M at the end of the period. If in period t a review is opened, they do not reveal any information to the AA (which needs therefore another period to conclude the investigation) and obtain profits Π_M. In $t + 1$, if they are proved guilty, they pay the fine F and are forced to non-cooperative prices, with profit Π_N; in $t + 2$ they revert to the collusive behaviour. If in $t + 1$ they are not proved guilty, they obtain profits Π_M and will go on colluding. If, however, a firm deviates in the marketplace or reveals information to the AA, Nash punishment starts and goes on forever with profits Π_N in every period.

Hence the firms combine the usual grim strategies of the supergame literature[17] with a revelation policy which is agreed upon, and they interpret any deviation from either the market strategy or the revelation policy as a break-down of the cartel. Moreover, firms collude until they are proved guilty, and restart collusion after an inquiry is concluded, as long as no deviation from the prescribed strategy occurred. These strategies are consistent with the idea that if the conditions for collusive behaviour are satisfied, firms tend to coordinate their actions as long as they are not forced to take non-cooperative actions by a sentence of the AA, and they restart collusion once the AA moves its attention to other industries.

2.3. Timing of the game

To summarise, the game starts at $t=0$ with the AA setting the policy parameters; in $t=1$ firms select a collusive strategy (market allocation and revelation policy) and decide whether to collude or to deviate from the proposed agreement.[18] If collusion is chosen, every period has the following structure: firms collude; immediately afterwards, the AA opens an investigation with probability α, which ends in the same period if the firms cooperate, or continues in the following period if no firm reveals. If firms are condemned, they are forced to play non cooperatively for the current period and to pay the (reduced or full) fine. After an investigation is concluded, the game restarts with the collusive strategy if no deviation from the equilibrium strategy occurred, while it goes on with the punishment phase if some firm deviated either from the market or from the revelation strategy.

We can now proceed to analyse the equilibrium of the game. We first consider (Section 3) the subgame perfect equilibria in the repeated game among firms once the policy parameters F, R, α and p have been set by the AA. We shall identify in the (α, p) space the regions corresponding to the different equilibria, which identify the incentive compatibility constraints when the AA designs the optimal policy. In Section 4 we shall consider the optimal policy choices of the AA, thus finding the solutions of the whole game.

3. The firms’ decisions

From the description of the strategies, and given that firms and industries are symmetric three possible outcomes are relevant. In a NC (No Collusion)

[17] See Friedman (1971) and, for a textbook presentation, Tirole (1988, chapter 6).

[18] As will be clear from the equilibrium analysis, the game is stationary once the policy parameters are set and therefore we can restrict the attention to deviations at $t=1$.

equilibrium collusion does not arise in any industry, because each participant would prefer to deviate rather than join a collusive agreement. In this case full ex-ante deterrence is reached. Alternatively, in each industry firms collude and reveal if monitored (CR) or they collude but refuse to reveal any information if an investigation is opened (CNR): in both cases a cartel starts, and the AA obtains ex-post desistence (for one period) when it is able to condemn the firms. When examining the conditions for the existence of a collusive equilibrium, we have to consider two possible deviations: a deviation from the market strategy, and a deviation from the revelation policy agreed upon by the firms. In what follows, we study the incentive compatibility constraints of the firms and determine the equilibrium outcomes.

3.1. CR: collude and reveal

We consider first the conditions for a CR equilibrium to exist, in which firms collude in the market and reveal information to the AA if a review is opened. From the timing of the game, we know that in each period $t \geqslant 1$ the firms are reviewed with probability α and, if monitored, they reveal, are forced to compete non-cooperatively for the current period and pay the reduced fine R; then, the game restarts. Hence, we can easily obtain the value of the CR strategy, V_{CR}:

$$V_{\mathrm{CR}} = \alpha(\Pi_N - R) + (1-\alpha)(\Pi_M) + \delta V_{\mathrm{CR}} = \frac{\Pi_M}{1-\delta} - \alpha\frac{\Pi_M - \Pi_N + R}{1-\delta} \qquad (1)$$

where Π_M are the profits from collusion, $\Pi_N < \Pi_M$ the non-cooperative profits obtained during the compliance phase and $\delta \in (0, 1)$ is the discount factor. Notice that the first term corresponds to the value of collusion in the standard case where no antitrust intervention is considered. The value of collusion becomes smaller when antitrust enforcement takes place for two reasons (occurring with probability α): the firms pay the fine R if found guilty, and they have a profit loss $\Pi_M - \Pi_N$ when the AA forces them to interrupt the collusive behaviour for the current period.

If an investigation is opened, there is no incentive to deviate from the revelation policy agreed upon in a CR equilibrium: by not revealing when the other firms are expected to cooperate with the AA, the (deviating) firm would get the full fine F instead of the reduced fine R, and would break the cartel, with further future losses. Hence, to establish the conditions for a CR equilibrium the relevant constraint is that the firm cannot be better off by deviating (in the market) from the beginning. In this case the value of the deviating strategy is

$$V_D = \Pi_D + \delta\frac{\Pi_N}{1-\delta} \qquad (2)$$

where $\Pi_D > \Pi_M$ are the profits in the deviation phase, which is followed by Nash punishment forever. Moreover, notice that, since the AA does not make type I error, even if the industry is reviewed the deviating firm is not found guilty of collusive behaviour and is immediately acquitted with no fine or restriction on current behaviour. The following Lemma states the conditions for a CR subgame perfect equilibrium to exist.

Lemma 1. *For given policy parameters (F, R, α, p), a subgame perfect equilibrium exists in which firms collude and reveal when monitored if*

$$\alpha < \alpha_{\mathrm{CR}}(R) = \frac{\Pi_M - (1-\delta)\Pi_D - \delta\Pi_N}{\Pi_M - \Pi_N + R} \tag{3}$$

Proof. The condition immediately follows from the inequality $V_{CR} > V_D$. □

Notice that $\alpha_{\mathrm{CR}}(R) \geqslant 0$ for $\delta \geqslant (\Pi_D - \Pi_M)/(\Pi_D - \Pi_N)$ which is the usual critical discount factor when firms collude with no threat of prosecution. For the remaining of the paper we focus on the interesting case where this minimal condition holds. Note that $\alpha_{\mathrm{CR}}(R)$ is decreasing in R and $\alpha_{\mathrm{CR}}(0) < 1$ (see also Fig. 1). Hence, granting more generous discounts increases the threshold value $\alpha_{\mathrm{CR}}(R)$ relaxing the constraint for a CR equilibrium and making collusion more attractive. In this sense, LP have a pro-collusive effect since they decrease the expected cost of anticompetitive behaviour. Notice that the probability of independent prosecution p plays no role in a CR equilibrium, since the evidence to prove the existence of the cartel is provided by the colluding firms themselves. Since firms stop collusion for one period and pay a reduced fine every time they are reviewed, we need a sufficiently low α in order to make firms better off colluding (and revealing) rather than deviating.

3.2. CNR: collude and not reveal

We proceed as before, deriving first the value of the game in a CNR equilibrium. From the timing of the game we know that, in each period $t \geqslant 1$, the AA opens a review with probability α; if monitored, firms do not reveal and continue colluding in the current period, while in the next period they are condemned with probability p; in this case, they pay the full fine F and behave non-cooperatively for the current period, while if not proved guilty collusion continues; after two periods the game restarts. If firms are not monitored, in a CNR equilibrium some other industry will be reviewed and the AA will not open new reviews for two periods, having to conclude the cases opened; hence, firms will have two periods of safe collusive profits before the game restarts. The value of the game in a CNR equilibrium is therefore:

$$V_{\mathrm{CNR}} = \alpha\{\Pi_M + \delta[p(\Pi_N - F) + (1-p)\Pi_M]\} + (1-\alpha)(1+\delta)\Pi_M + \delta^2 V_{\mathrm{CNR}} \tag{4}$$

After rearranging, we obtain:

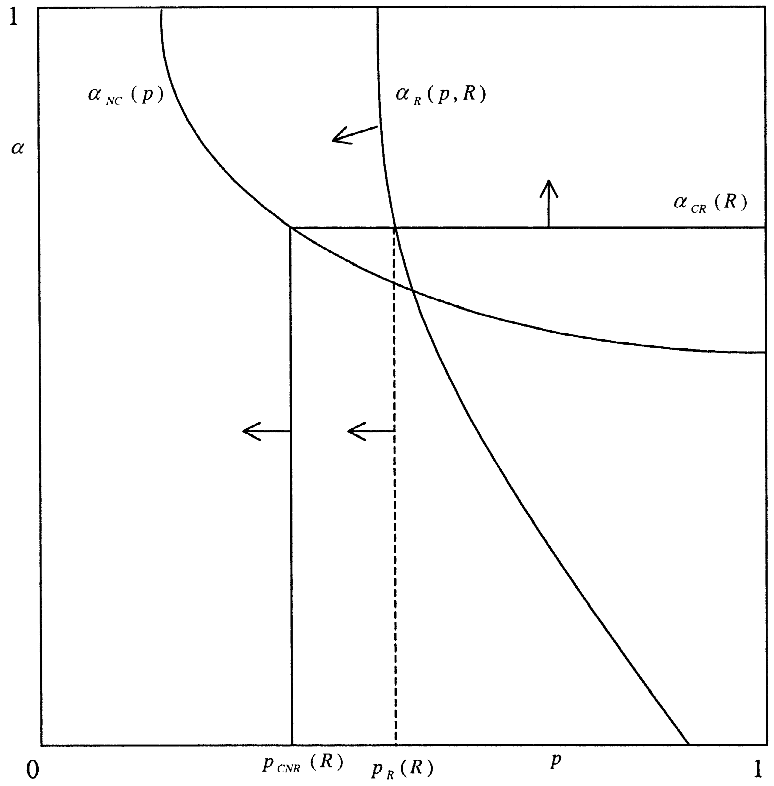

Fig. 1. Incentive compatibility constraints.

$$V_{\mathrm{CNR}} = \frac{\Pi_M}{1-\delta} - \alpha p \frac{\delta(\Pi_M - \Pi_N + F)}{1-\delta^2} \tag{5}$$

where, as before, the standard cartel profits are reduced by the expected losses from antitrust enforcement, where now the ex-ante probability of being fined is αp. In order to establish if CNR is a subgame perfect equilibrium we have to consider two constraints: (i) the firms prefer to collude and not reveal rather than deviate (in the market) from the beginning; (ii) when reviewed by the AA, they prefer not to reveal rather than cooperate in the investigation.

The conditions for the first constraint to hold can be easily obtained by requiring that $V_{\mathrm{CNR}} \geqslant V_D$. Substituting the expressions and rearranging we get:

$$\alpha < \alpha_{\mathrm{NC}}(p) = \frac{(1+\delta)(\Pi_M - (1-\delta)\Pi_D - \delta\Pi_N)}{\delta p(\Pi_M - \Pi_N + F)} \tag{6}$$

The curve $\alpha_{\mathrm{NC}}(p)$, shown in Fig. 1, does not depend on R and is decreasing in p: a higher probability p of being condemned must be balanced by a lower probability α of being reviewed, in order to maintain the firms indifferent between CNR and deviate.

Consider now the second constraint for a CNR subgame perfect equilibrium, which becomes relevant when LP are introduced: for a CNR equilibrium to exist we want that in the subgame starting when the AA opens an investigation, the firms prefer not to reveal. The value of the game if the firm reveals once monitored, deviating from the prescriptions of a CNR equilibrium, is:

$$V_R|\alpha = \frac{\Pi_N}{1-\delta} - R \tag{7}$$

If instead the firm does not reveal, according to the equilibrium strategy, the value of the game from that point on is:

$$V_{\mathrm{NR}}|\alpha = \Pi_M + \delta[p(\Pi_N - F) + (1-p)\Pi_M] + \delta^2 V_{\mathrm{CNR}} \tag{8}$$

Substituting the expression for V_{CNR} and rearranging we obtain:

$$V_{\mathrm{NR}}|\alpha = \frac{\Pi_M}{1-\delta} - \frac{\delta p(1-\delta^2(1-\alpha))(\Pi_M - \Pi_N + F)}{1-\delta^2} \tag{9}$$

Not revealing is optimal once monitored if $V_{\mathrm{NR}}|\alpha \geqslant V_R|\alpha$. Substituting and rearranging we obtain the second relevant locus to establish a CNR equilibrium:

$$\begin{aligned}\alpha &< \alpha_R(p,R)\\ &= \frac{(1+\delta)\{\Pi_M - \Pi_N + R(1-\delta) - \delta p(1-\delta)(\Pi_M - \Pi_N + F)\}}{\delta^3 p(\Pi_M - \Pi_N + F)}\end{aligned} \tag{10}$$

Hence, the first constraint (no deviation in market strategy) leads to the $\alpha_{\mathrm{NC}}(p)$ curve while the second constraint (no deviation in the revelation policy) gives the $\alpha_R(p, R)$ curve. This latter shifts down when the reduced fine R is lowered. When R is close to the full fine F, the $\alpha_R(p, R)$ curve is always above the $\alpha_{\mathrm{NC}}(p)$ curve for α and p in [0, 1] and the no revelation constraint never binds: since revealing once monitored entails the breakdown of the cartel and a fine R for sure, when R is close to F, there is no incentive to reveal. When instead R becomes small, the $\alpha_R(p, R)$ curve at some point becomes lower than the $\alpha_{\mathrm{NC}}(p)$ curve for some α and p in [0, 1], meaning that the no revelation constraint becomes binding for the existence of the cartel when p is sufficiently high (Fig. 1 gives an example where for p high enough $\alpha_R(p, R) < \alpha_{\mathrm{NC}}(p)$, i.e. the no revelation constraint binds in a CNR equilibrium). We can now state the conditions for a CNR equilibrium.

Lemma 2. *For given policy parameters (F, R, α, p), a CNR equilibrium exists if* $\alpha < \min\{\alpha_{NC}(p), \alpha_R(p, R)\}$.

Proof. It immediately follows from the discussion above. □

3.3. CNR vs. CR

So far we have identified the conditions such that a CR or a CNR strategy is preferred to a deviating strategy. Comparing the conditions for the existence of a CR ($\alpha < \alpha_{\mathrm{CR}}(R)$) and a CNR ($\alpha < \min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p,R)\}$) subgame perfect equilibrium when $R < F$, we can notice that there are regions of parameters that admit both types of equilibria. Then, the condition $V_{\mathrm{CNR}} > V_{\mathrm{CR}}$ allows to identify when CNR dominates CR: similar to what is standard in the analysis of collusion, we assume that firms are able to select the most profitable allocation to be implemented not only concerning the market strategy but also with reference to the revelation policy. Substituting and rearranging we get:

$$p < p_{\mathrm{CNR}}(R) = \frac{(1+\delta)(\Pi_M - \Pi_N + R)}{\delta(\Pi_M - \Pi_N + F)} \tag{11}$$

That is, when the respective equilibrium conditions hold, not revealing is preferred when the probability p of being condemned is sufficiently low. Notice that $p = p_{\mathrm{CNR}}(R)$ can be rewritten as:

$$p\frac{\delta(\Pi_M - \Pi_N + F)}{1+\delta} = \Pi_M - \Pi_N + R \tag{12}$$

Then, for $p = p_{\mathrm{CNR}}(R)$, the expected average losses suffered with probability p in a CNR equilibrium (the left hand side term) equal the average losses suffered with certainty in the CR case (the right hand side expression).

Before moving to a full statement of the subgame perfect equilibria in the game, we need further to identify when the flat locus $\alpha_{\mathrm{CR}}(R)$, which identifies the region where a CR equilibrium exists, cuts the downward sloping locus $\min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p, R)\}$, below which a CNR equilibrium exists. It is easy to show that setting $\alpha_{\mathrm{CR}}(R) = \alpha_{\mathrm{NC}}(p)$ and solving for p gives $p = p_{\mathrm{CNR}}(R)$, i.e. the three curves intersect in the same point as shown in Fig. 1. Setting $\alpha_{\mathrm{CR}}(R) = \alpha_R(p)$ and solving for p we get:

$$p_R(R) = \frac{(1+\delta)(\Pi_M - \Pi_N + R)(\Pi_M - \Pi_N + R(1-\delta))}{\delta(\Pi_M - \Pi_N + F)\left[\Pi_M - \Pi_N + R(1-\delta^2) - \delta^2(1-\delta)(\Pi_D - \Pi_N)\right]} \tag{13}$$

Since $p_{\mathrm{CNR}}(R)$ and $p_R(R)$ correspond to the intersection of the horizontal line $\alpha_{\mathrm{CR}}(R)$ with the two downward sloping curves $\alpha_{\mathrm{NC}}(p)$ and $\alpha_R(p, R)$, $p_{\mathrm{CNR}}(R) < p_R(R)$ implies that at $\alpha = \alpha_{\mathrm{CR}}(R)$ the former is to the left of the latter, as in Fig. 1, and vice versa. Simple calculations show that $p_{\mathrm{CNR}}(R) < p_R(R)$ when $R < \delta(\Pi_D -$

Π_N). Hence, when the reduced fine is sufficiently low, the relevant curves correspond to Fig. 1.

Since $\alpha_{\mathrm{CR}}(R)$, $\alpha_R(p, R)$, $p_{\mathrm{CNR}}(R)$ and $p_R(R)$ all depend on the reduced fine R, the relations among these curves and $\alpha_{\mathrm{NC}}(p)$, as well as the regions corresponding to the different equilibria, change with R. In particular, when R becomes lower and lower than F, the $\alpha_R(p, R)$ curve moves down and to the left while $\alpha_{\mathrm{CR}}(R)$ increases, both $p_{\mathrm{CNR}}(R)$ and $p_R(R)$, which are initially greater than one, decrease, and when $R < \delta(\Pi_D - \Pi_N)$, $p_{\mathrm{CNR}}(R)$ becomes lower than $p_R(R)$. In Fig. 1 the different curves are drawn, with the arrows showing how the curves shift when R is lowered. In the following proposition we give the conditions for a CNR, a CR and a NC equilibrium to exist and, in case of multiple equilibria, to be Pareto dominant, as a function of the policy parameters.

Proposition 1. *In the repeated game played by the firms from* $t = 1$ *on, once the policy parameters* (F, R, α, p) *are set, we can describe the Subgame Perfect Equilibrium (SPE) in the* (α, p) *space as follows:*

– When R is close to F, such that $p_{\mathrm{CNR}}(R) > p_R(R) > 1$*, the unique SPE is NC above the locus* $\min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p, R)\}$ *while the Pareto dominant SPE is CNR below the locus.*

– When R is lower, such that $p_{\mathrm{CNC}}(R) > 1 > p_R(R)$*, the Pareto dominant SPE is CR above* $\alpha_R(p, R)$ *and below* $\alpha_{\mathrm{CR}}(R)$*; the Pareto dominant SPE is CNR below the locus* $\min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p,R)\}$ *while the unique SPE is NC otherwise.*

– When R is such that $1 > p_{\mathrm{CNR}}(R) > p_R(R)$*, the Pareto dominant SPE is CR for* $p \in [p_R(R), p_{\mathrm{CNR}}(R)]$ *and* $\alpha \in [\alpha_R(p,R), \alpha_{\mathrm{CR}}(R))$ *and* $p \in [p_{\mathrm{CNR}}(R), 1]$ *and* $\alpha \in [0, \alpha_{\mathrm{CR}}(R))$*, it is CNR below the locus* $\min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p,R)\}$ *for* $p \in [0, p_{\mathrm{CNR}}(R)]$ *while the unique SPE is NC otherwise.*

– Finally, for $0 < R < \delta(\Pi_D - \Pi_N)$ *we have* $1 > p_R(R) > p_{\mathrm{CNR}}(R)$ *and the Pareto dominant SPE is CR for* $p \in [p_{\mathrm{CNR}}(R), 1]$ *and* $\alpha \in [0, \alpha_{\mathrm{CR}}(R))$*, it is CNR for* $p \in [0, p_{\mathrm{CNR}}(R))$ *and* $\alpha < \alpha_{\mathrm{NC}}(p)$ *while the unique SPE is NC otherwise.*

Proof. See Appendix B. □

The proposition identifies the regions where the CNR, CR and NC equilibria exist. Recall that different conditions must be satisfied for these equilibria to exist. To summarise, CR exists if $\alpha < \alpha_{\mathrm{CR}}(R)$; CNR exists when $\alpha < \min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p,R)\}$; NC exists if $\alpha \geqslant \alpha_{\mathrm{CR}}(R)$ and when $\alpha \geqslant \min\{\alpha_{\mathrm{NC}}(p), \alpha_R(p, R)\}$. Finally, the condition $p < p_{\mathrm{CNR}}(R)$ determines whether firms prefer the CNR over the CR collusive strategy (and vice versa when the inequality does not hold). The level of the reduced fine R affects the above conditions, and this explains why Proposition 1 is stated for different levels of R. In particular, only NC and CNR equilibria exist when R is too close to F: the reduction in profits from Π_M to Π_N would occur with certainty in the current period in a CR equilibrium and with probability p in the following period in a CNR equilibrium. Hence, when the saving in fines is negligible, if the firms find it convenient to collude, they prefer not to reveal. Consequently, LP require a sufficient discount in fines to become effective.

Rather than illustrating all the equilibria stated in Proposition 1 for different values of R, we just show in the following figures the equilibrium solutions for a very low value of R to save space. However, we should stress that, as Proposition 2 will show, whenever an agency decides to introduce a leniency program, it will find it optimal to choose $R = 0$. Therefore, intermediate values of R will never arise at the equilibrium of the whole game.

Fig. 2 describes the three regions of parameters in the space (α, p) for a very low value of $R < F$. When both α and p are high (the north-east region) deterrence is very effective and the value of colluding is decreased, making deviation more attractive: as a result, no cartel arises (NC). When α is low but p is high (the south-east region) ex-ante deterrence is not very effective and firms start colluding; but refusing to cooperate with the AA once the review is opened is not rewarding since firms will likely receive the full fine (p is high), and firms prefer to collude but reveal (CR) if monitored. Finally, for low p deterrence is not effective as well and firms prefer to collude; but now they prefer not to reveal information if monitored (CNR), since the ability of the AA to condemn them to the full fine is very low.

The NC, CR and CNR regions in Fig. 2 are obtained under the assumption that the firms, if found guilty, will behave non-cooperatively for one period, then restart collusion in the following period.[19]

We can now compare the outcomes of the game played by the firms with and without leniency programs. There are two effects at work: under LP, there is a region of parameters (labelled 1 in Fig. 2) which induces CR under leniency programs that would rather prevent collusion when no reduced fines are granted: since the expected cost of misbehaviour is lower, LP have a pro-collusive effect in this case. However, when collusion arises, the use of LP allows to obtain ex-post desistence more easily, by inducing revelation and by shortening the investigations, for certain values of the policy parameters labelled region 2. In order to establish which effect prevails, we have now to move to the implementation of the optimal policies.

Before studying the optimal policies, however, we would like to stress that fine discounts in our setting must be more generous than in Kaplow and Shavell

[19] We do not regard the assumption that firms stop colluding for just one period as crucial. If the AA monitored the firms for T periods after a sentence, the qualitative features of Fig. 2 would remain, with the three regions identified by similar boundaries. In Motta and Polo (1999) we consider the case where firms proved guilty do not collude any more (while if the AA is unable to condemn the firms, the cartel is not reviewed again in the future). That case implies that if found guilty, a firm is constantly monitored by the AA (while a second inquiry is never opened if the firms were not proved guilty). The results do not change in that alternative scenario.

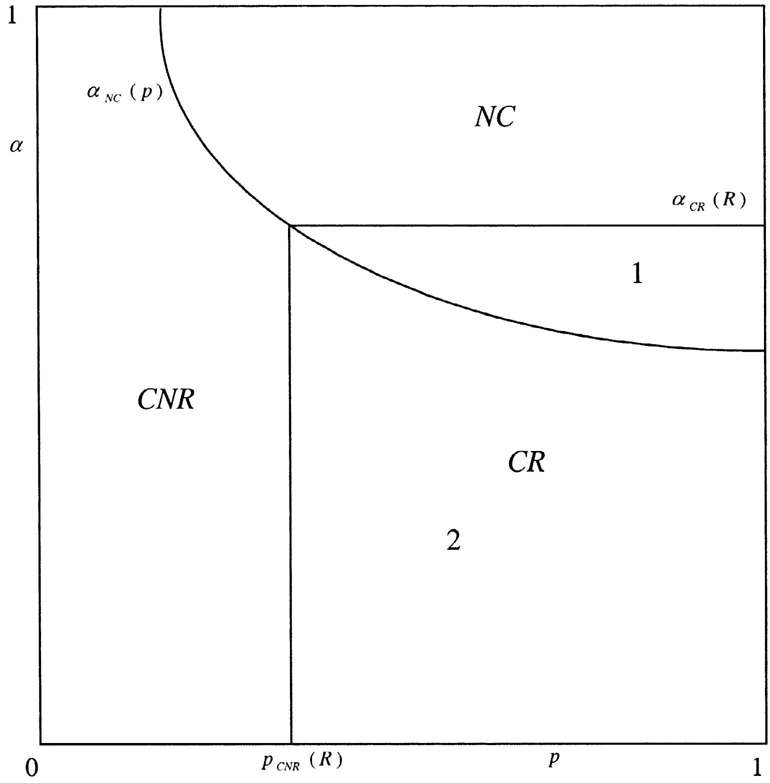

Fig. 2. Subgame perfect equilibria.

(1994). They show that to induce violators to self-report it is enough to set the reduced fine such that $pF \geqslant R$. In our setting, a firm would reveal if $p > p_{\mathrm{CNR}}(R)$, which can be rewritten as:

$$\frac{\delta}{1+\delta}pF - \frac{(1+\delta(1-p))}{1+\delta}(\Pi_M - \Pi_N) \geqslant R. \tag{14}$$

There are two reasons why this condition differs from the findings of Kaplow and Shavell (1994). The first just lies in the formal specification of the model, and the fact that when firms do not reveal they can be fined only with a one period lag and every two periods (hence the term $\delta/(1+\delta)$ which multiplies pF). The second one is more important, and is due to the fact that by revealing the firm also foregoes with certainty some (appropriately weighted and discounted) future profits that, by not revealing, it would get with probability $(1-p)$ if not found guilty (hence the second term in the inequality above). Therefore we should expect that whenever self-reporting diminishes the future profits of violators (which might occur not only in the case of interruption of a collusive practice, but also in cases where self-reporting increases the compliance of polluters, or reveals black assets to the tax authorities), fine discounts have to be more generous.

4. Optimal enforcement

In the previous section we have studied the firms' decisions given policy parameters. We now move to the endogenous choice of such parameters. The analysis of the optimal enforcement choices of the AA is built in two steps. We first consider the policy combinations (α, p) that the AA can implement given its enforcement technology, i.e. its budget constraint; then we derive the iso-welfare gains curves in the (α, p) space where the equilibrium outcomes of the game among firms were identified; finally we discuss under which conditions leniency programs should be used.

4.1. Budget constraint

In this section we specify the enforcement technology and derive the locus of implementable policies, that we define as the AA budget constraint. The AA is (exogenously) endowed with a sunk per-period budget; we assume that setting the fines at any level is not costly, while increasing the probability of enforcement requires resources.

In general we expect a trade-off between the monitoring rate α and the successful prosecution rate p implementable, in the sense that with given resources increasing the former requires a contraction in the latter. This might occur, for instance, because the officers must be assigned either to monitoring or to prosecution, or for any other technological constraint that makes the outcomes of these two tasks negatively related. The general conclusions we shall draw regarding the optimal policies hold for any downward sloping budget constraint. Here we propose a specific modelling of the enforcement technology that quite naturally conveys a negative relationship.

The total budget allows to hire L officers that are organised in teams of l units.[20] A team performs both the monitoring and (if firms do not reveal) the prosecution tasks. It opens an investigation by selecting randomly an industry within a pool of

[20] l can be interpreted as the overall time spent by the members of a team on a single case during the period.

N symmetric industries[21] which are potentially collusive. Given the total labour force L, choosing the dimension of the teams determines the number[22] of cases n that can be treated in a period, $n = L/l$. In the case of symmetric cartels that we are considering, all the industries choose the same equilibrium behaviour NC, CR or CNR: in this latter case the investigations last two periods. Hence, the AA opens n new cases each period (NC and CR) or every two periods (CNR).The probability that an industry is reviewed is therefore

$$\alpha = \frac{n}{N} = \frac{l_N}{l} \tag{15}$$

where $l_N = L/N$ is the dimension of (time spent by) a team if all the N industries were reviewed at the same time: since the interesting case is when the AA has scarce resources, in our discussion l_N will be very small.

The dimension of the team influences also the probability p of proving firms guilty when they do not cooperate. We assume that a minimum scale of the team (time spent) l_0 is needed in order to obtain some evidence, and that prosecution has decreasing returns, according to the function:

$$p = g(l - l_0) \tag{16}$$

with $g(0) = 0$, $\lim_{l \to \infty} g(\cdot) = 1$, $g'(\cdot) > 0$, and $g''(\cdot) < 0$ for $l \geqslant l_0$, and $\lim_{l \to l_0} g'(\cdot) = \infty$.

There is a trade-off in the enforcement policy between opening more reviews, which requires smaller teams, and being able to successfully conclude them, which is more likely if the teams are larger. To obtain the budget constraint, we can notice that $g(\cdot)$ is increasing and we can invert it, defining $f(p) = g^{-1}(p)$. Then $l = l_0 + f(p)$ is the team size needed to obtain a successful prosecution probability p: it is increasing and convex in p, with a vertical intercept at l_0.

Since $\alpha = l_N/l$, the budget constraint of the AA, is

$$\alpha_{BC}(p) = \frac{l_N}{l_0 + f(p)} \tag{17}$$

for $0 \leqslant p \leqslant g(L - l_0)$, which is the highest feasible probability, obtained if all the

[21] As specified above, we assume that an industry can be investigated repeatedly, no matter whether it was previously found guilty or not; moreover, if collusion restarts after a sentence, a new investigation might be opened immediately finding evidence up to the current period regarding the new attempt to collude; finally, if an industry or a firm not colluding (i.e. behaving non-cooperatively or deviating) is reviewed, the investigation is closed in the same period. Moreover, we assume symmetric industries, which implies that a CR, a CNR or a NC equilibrium is implemented in each industry. All these assumptions imply that the number N of industries subject to monitoring when new reviews are opened remains constant over time. Hence, the game remains stationary even when we endogenize the choice of the policy parameters.

[22] To ease the exposition and analysis we treat n as defined over the real line.

officers L are allocated to a single case. The $\alpha_{BC}(p)$ is downward sloping, initially concave and then convex. In Appendix A these properties are formally established. Moreover, since $l_N = L/N$, an increase in the budget L shifts up parallelly the vertical intercept of $\alpha_{BC}(p)$.

Along the budget constraint we have different levels of the ex-ante probability of being fined, $\alpha_{BC}(p) \cdot p$. Notice that

$$\frac{\mathrm{d}\alpha_{BC}(p) \cdot p}{\mathrm{d}p} = \frac{l_N[l - pf'(p)]}{l^2} = \frac{l_N}{l}(1 - \epsilon_p)$$

where $\epsilon_p \equiv \mathrm{d}f/\mathrm{d}p \cdot p/l$ is the elasticity of the team size l with respect to the target probability p. Hence, when $\epsilon_p = 1$ the ex-ante probability of being fined is constant. Let us define p_1 the probability that makes $\epsilon_p = 1$. At $p = p_1$ a 1% increase in the probability of successful prosecution requires a team 1% larger, which in turn reduces by 1% the probability of being reviewed, leaving the ex-ante probability of being fined $\alpha_{BC}(p) \cdot p$ constant. The probability p_1 will play a crucial role in the analysis of the optimal enforcement.

4.2. Welfare gains

We can now derive a welfare measure of the antitrust activity. When cartels can arise in the market, antitrust intervention improves social welfare by preventing (deterrence) or temporarily interrupting (desistence) collusion: this latter case, as we argued above, corresponds to the restrictions and remedies that the AA can impose on the behaviour of guilty firms for a certain period through a sentence. We evaluate these effects through a utilitarian welfare function with equal weights on consumer and producer surplus. Moreover, the fines are assumed to be pure transfers that do not affect the aggregate welfare. Hence, in the evaluation of the welfare effects we focus on the deterrence and desistence properties of LP.

The traditional deadweight loss (DWL) measures the welfare gains associated with a successful intervention that induces a more competitive market equilibrium. We evaluate the welfare gains of antitrust enforcement by comparing the equilibrium outcomes NC, CR and CNR with the situation where collusion arises because no policy intervention is promoted.

As mentioned when describing the enforcement technology, the AA faces N industries which can potentially promote collusion. In a NC equilibrium, no cartel arises and therefore from $t = 1$ the AA realizes the following welfare gains:

$$W_{\mathrm{NC}} = N\frac{DWL}{1 - \delta} = NK \tag{18}$$

where K is the present value of avoiding the cartelization of an industry. When the firms coordinate on a Collude and Reveal equilibrium, the AA opens in each period n reviews which induce revelation and end up with the n industries behaving non cooperatively for the current period (with DWL gains), until the AA moves its attention to another industry. Hence, starting from $t = 1$, the AA obtains $nDWL$ gains per period, with a total welfare gain

$$W_{\mathrm{CR}} = \frac{nDWL}{1-\delta} = nK = \alpha W_{NC} \tag{19}$$

Comparing the welfare gains under NC and CR, the latter is lower because it interrupts cartels only with probability α in each period (and $\alpha \leqslant \alpha_{\mathrm{CR}} < 1$ in the CR region). Finally, in a Collude and Not Reveal equilibrium the AA interrupts n cartels for one period ($nDWL$) only with probability p, taking two periods to conclude the procedure. The welfare gains are therefore

$$W_{\mathrm{CNR}} = \frac{np\delta DWL}{1-\delta^2} = \frac{np\delta K}{1+\delta} = p\frac{\delta}{1+\delta}W_{\mathrm{CR}} = \alpha p\frac{\delta}{1+\delta}W_{\mathrm{NC}} \tag{20}$$

Compare the welfare gains of antitrust intervention with and without LP for given α in a right (CR) and left (CNR) neighbourhood of p_{CNR}: CNR gives a lower welfare gain because ex-post desistence occurs with a lower probability (p) once a case is opened, and because it takes two periods and a discount factor $\delta/(1+\delta)$ to reach a decision.

Consider now the iso-welfare gains curves in the (α, p) space associated with the three outcomes. We initially focus on the case in which no LP are used, i.e. $R = F$: no collusion (NC) and collude and not reveal (CNR) are the two possible outcomes. In all the region NC the welfare gains are the same, as they do not depend on α and p. In the CNR region, the welfare gains depend on both α and p and the indifference curves are equilateral hyperboles. If the welfare gain is W the iso-welfare gains curve in the CNR region is

$$\overline{\alpha}_{\mathrm{CNR}}(p) = \frac{(1+\delta)W}{\delta NKp} \tag{21}$$

Hence, $\overline{\alpha}_{\mathrm{CNR}}(p)$ is an equilateral hyperbole as the upper boundary of the CNR region $\alpha_{\mathrm{NC}}(p)$ is.[23]

Consider now the iso-welfare gains curves when LP are introduced with $R = 0$.

[23] In fact, it is easy to verify that the two curves overlap for a welfare gain in the CNR region equal to

$$W = NK\frac{\Pi_M - (1-\delta)\Pi_D - \delta\Pi_N}{\Pi_M - \Pi_N + F} \tag{22}$$

This welfare gain is the highest that can be realized in a CNR equilibrium, and corresponds to the highest iso-welfare curve in the CNR region.

Now three outcomes can occur: NC, CNR and CR. In the CR region, W_{CR} depends only on α, and therefore the iso-welfare curves are horizontal. For a certain welfare gain W, the iso-welfare gains curve in the CR region is flat at

$$\overline{\alpha}_{\mathrm{CR}} = \frac{W}{NK} \tag{23}$$

To find the (same) iso-welfare gains curve that passes through the CNR and CR regions when LP are used, the following argument applies (refer to Fig. 3): let us fix a welfare gain $\tilde{W}$; in the CNR region the iso-welfare gains curve corresponds to the $\overline{\alpha}_{\mathrm{CNR}}(p)$ curve already discussed. Once entering the CR region, the iso-welfare

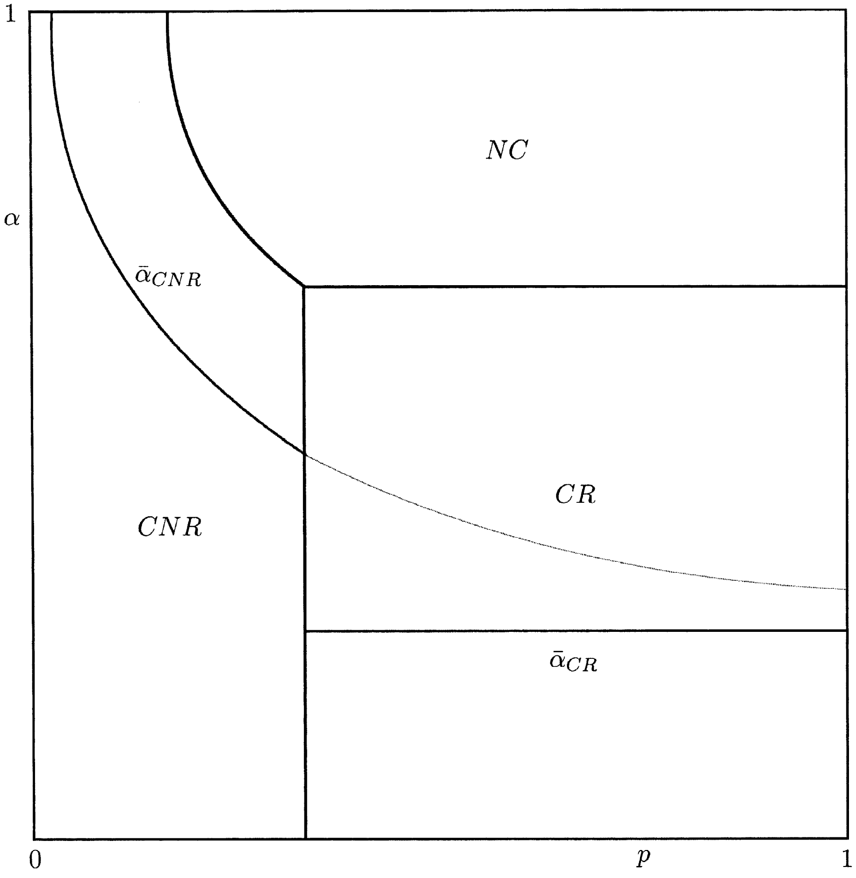

Fig. 3. Iso-welfare gains curves giving the same welfare level in the CNR and CR regions.

gains curve jumps down at $\overline{\alpha}_{\mathrm{CR}} = \tilde{W}/NK$ and becomes horizontal.[24] In Fig. 3 the thick line represents the iso-welfare gains curve passing through the CNR and CR regions.

4.3. Optimal policies

We can now identify the optimal policies given the iso-welfare gains curves, the budget constraint and the incentive compatibility constraints that identify the equilibrium outcomes in the game played by the firms for given policy parameters. We first characterize the optimal policy when the AA wants to implement one of the three outcomes NC, CNR and CR. Then we compare the implementable outcomes and select the best one.

As a general point in all the equilibrium outcomes, it is always optimal to set $F = \overline{F}$ since increasing the fines is not costly and allows to obtain more favourable (lower) boundaries $\alpha_{\mathrm{NC}}(p)$ and $p_{\mathrm{CNR}}(R)$.

Proposition 2. *The optimal policies that implement the NC, CNR and CR outcomes are:*

– *If $\alpha_{BC}(p) \geqslant \alpha_{\mathrm{NC}}(p)$ for some $p \in [0, 1]$ the NC outcome can be implemented. The optimal policy picks up the tangency point of the $\alpha_{\mathrm{NC}}(p)$ and $\alpha_{BC}(p)$ curves at $p = p_1$, where the ex-ante probability of being fined, $\alpha_{BC}(p_1) \cdot p_1$, remains constant. If no tangency point exists, the optimal policy entails the corner solution at $\alpha_{\mathrm{NC}}(p) = 1$.*

– *The optimal policy to implement CNR sets $R = F$ and picks up the tangency point between the iso-welfare gains curve $\overline{\alpha}_{\mathrm{CNR}}(p)$ and the budget constraint $\alpha_{BC}(p)$ at the point $p = p_1$, where the ex-ante probability of being fined, $\alpha_{BC}(p_1) \cdot p_1$, remains constant. If no tangency point exists, the optimal policy entails the corner solution $p = g(l_N - l_0)$ and $\alpha = 1$ or $p = g(L - l_0)$ and $\alpha = l_N/(l_0 + f(p))$.*

– *If $g(L - l_0) \geqslant p_{\mathrm{CNR}}(0)$, a CR outcome can be implemented. The optimal policy sets $R = 0$, $p = p_{\mathrm{CNR}}(0)$ and $\alpha = l_N/(l_0 + f(p))$.*

Proof. See Appendix B. □

Proposition 2 describes the optimal combination of policy parameters (α, p) which implements each of the three possible sub-game perfect equilibrium outcomes. This amounts to finding, within each region, the highest iso-welfare gains curve subject to a given budget constraint. For regions NC and CNR the optimal point will be given by the tangency point between the budget curve and the iso-welfare curves (or by a corner solution, as stated in Proposition 2). Since

[24] Notice that $\overline{\alpha}_{\mathrm{CR}} = \tilde{W}/NK < \overline{\alpha}_{\mathrm{CNR}}(1) = (1+\delta)\tilde{W}/\delta NK$, i.e. the flat portion $\overline{\alpha}_{\mathrm{CR}}$ of the iso-welfare gains curve is below the value of the downward sloping portion when prolonged to $p = 1$, $\overline{\alpha}_{\mathrm{CNR}}(1)$.

the iso-welfare gains curves in these two regions are equilateral hyperboles with slope α/p while the budget constraint's slope is $\alpha\varepsilon_p/p$, the tangency point is at $p = p_1$. Intuitively, if at the optimal policy that implements CNR or NC the ex-ante probability $\alpha \cdot p$ were not constant along the budget constraint it would be welfare improving to modify the policy mix obtaining a more effective enforcement by reducing the expected profits from collusion.

Fig. 4 shows how the equilibria NC, CR and CNR can be optimally implemented for different budget constraints. Although a budget constraint may pass through more than one region NC, CR and CNR, we have drawn three different budget constraints, one for each implemented policy, to make the picture clearer. Given a budget α^1_{BC}, the best policy combination that implements NC is described by point E^1, i.e. at the tangency point of the budget constraint and the $\alpha_{\mathrm{NC}}(p)$ curve corresponding to $p = p_1$. Similarly, the CNR outcome is optimally im-

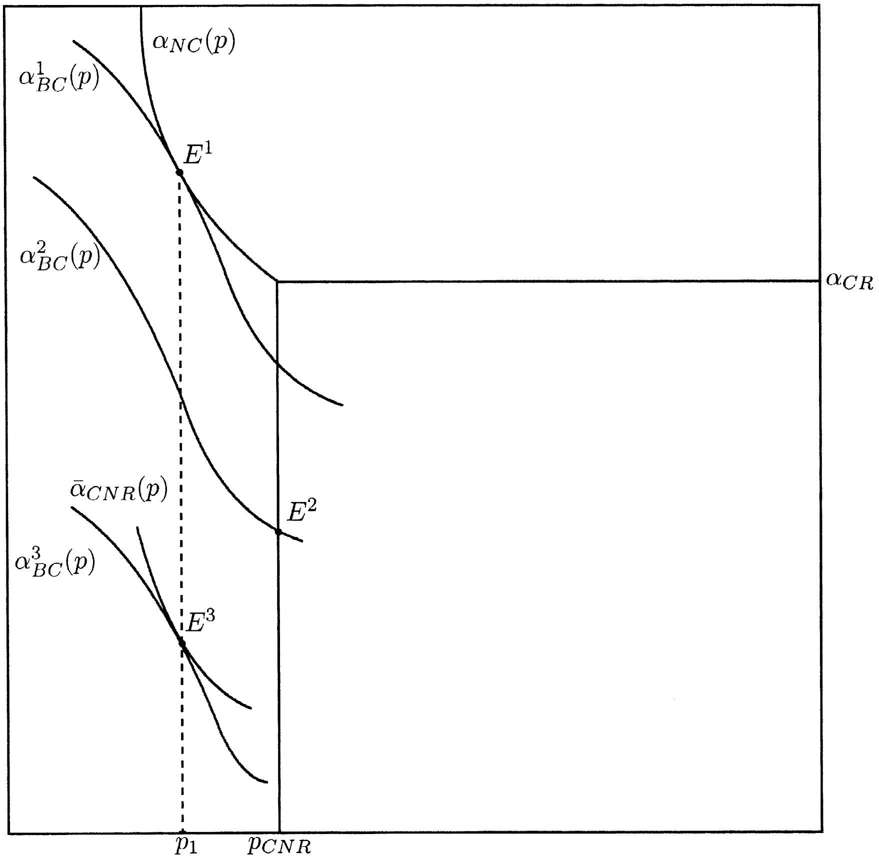

Fig. 4. The optimal policies to implement NC (point E^1), CR (point E^2), and CNR (point E^3).

plemented for budget α^3_{BC} at the tangency point with the iso-welfare curve $\overline{\alpha}_{\text{CNR}}(p)$, again for $p = p_1$.

Given a (downward sloping) budget α^2_{BC}, the highest (horizontal) iso-welfare gains curve attainable in the region CR is reached at the corner solution E^2 with $p = p_{\text{CNR}}(0)$. Note that the optimal leniency scheme calls for $R = 0$. To understand why this is the case, suppose that $R > 0$. By setting $p = p_{\text{CNR}}(R)$ the AA makes firms indifferent between revealing or not. By giving a more generous discount $R = 0$ firms strictly prefer to reveal, and the AA can still induce revelation reducing p. In turn, this frees up some resources of the AA, which can organise smaller teams and open more reviews, still obtaining firms' revelation. Note also that—although the welfare gains in a CR equilibrium depend only on α, the optimal policy has to satisfy the constraint $p = p_{\text{CNR}}$ to be credible. If the AA, relying on firms' cooperation, chose to open a very large number of reviews (to reach a very high α) by organising very small teams, then it would be unable to successfully conclude any of the reviews with a probability p sufficiently high to induce revelation.

We have found the optimal policies needed to implement the three different outcomes NC, CNR and CR. It is intuitive that, for any downward sloping budget constraint, the same qualitative features hold: a tangency point for the NC or CNR outcome and a corner solution with $R = 0$ when CR is implemented. Our last step is to identify the conditions for selecting the outcome associated to the highest welfare gain. The following Proposition states the result.

Proposition 3. *If the NC outcome can be implemented, i.e. $\alpha_{BC}(p) \geqslant \alpha_{\text{NC}}(p)$ for some $p \in [0, 1]$, the optimal policy entails implementing the NC outcome. If only CR and CNR can be implemented, i.e. $\alpha_{BC}(p) < \alpha_{\text{NC}}(p)$ for any $p \in [0, 1]$, the optimal policy selects, among the two, the outcome that gives the higher welfare gains: if $\alpha_{BC}(p_{\text{CNR}}(0)) \geqslant \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$, the optimal policy implements a CR outcome. If $\alpha_{BC}(p_{\text{CNR}}(0)) < \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$, or if $g(L - l_0) < p_{\text{CNR}}(0)$, the optimal policy implements a CNR outcome.*

Proof. See Appendix B. □

Proposition 3 identifies the conditions under which the Antitrust Authority selects the outcome NC, CNR or CR using the optimal policies analyzed in Proposition 2. If the budget constraint is sufficiently high to implement the NC outcome, this is the optimal policy. For intermediate budgets, only the CNR and CR outcomes are implementable and we have to compare the welfare gains obtained in the two cases: the condition $\alpha_{BC}(p_{\text{CNR}}(0)) \geqslant \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$ ensures that the CR outcome can be implemented with higher welfare gains than the CNR outcome, i.e. that the budget constraint reaches a higher iso-welfare gains curve in the CR than in the CNR region. Finally, when this condition fails to hold, or for very low budgets, the CNR outcome is the best policy.

Let us discuss briefly the case of intermediate budget levels. In Fig. 5 we have an example in which $\alpha_{BC}(p_{\mathrm{CNR}}(0)) \geqslant \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$, which makes CR the preferred outcome (recall that we have already proved that the best leniency policy calls for $R = 0$). Point A, where the budget constraint and the iso-welfare gains curve in the CNR region are tangent at $p = p_1$, is the best CNR outcome implementable. The associated welfare gain is $W^m = NK\delta p_1 \alpha_{BC}(p_1)/(1+\delta)$. The same welfare level is obtained, in the CR region, along the flat iso-welfare gains curve, setting $\alpha^m = W^m/NK = \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$. But we see that the budget constraint enters in the CR region at $\alpha^M = \alpha_{BC}(p_{\mathrm{CNR}}(0))$ (point B) which is higher than α^m: welfare gains W^Mat point B (in region CR) are therefore higher than W^m at point A (in region CNR).

It is now easier to see how the budget of the AA determines the optimal policy. Note first that, as proved in Appendix A, different levels of the budget L shift up or down the $\alpha_{BC}(p)$ curve but leave unchanged the probability p_1 which is chosen

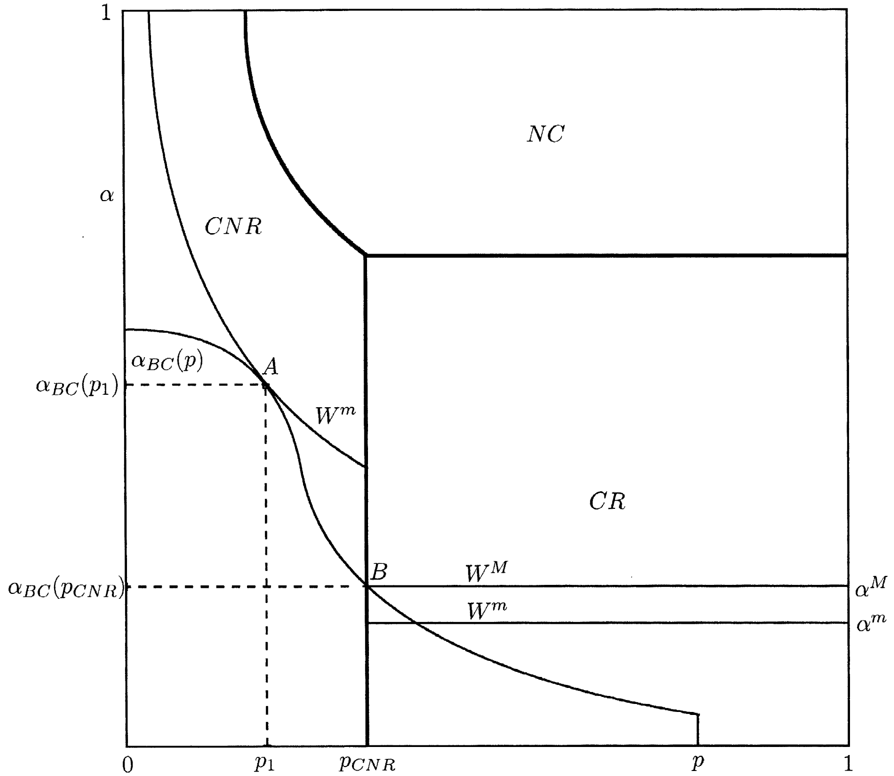

Fig. 5. Welfare comparison of CNR (point A) and CR (point B) optimal policies when CR is preferred.

in both the NC and CNR outcomes. Let us start from a rich budget L such that the NC outcome can be implemented. In this case the optimal policy selects the NC outcome and sets $p = p_1$ and $\alpha = \alpha_{\mathrm{NC}}(p_1)$ as shown in Fig. 4. Now let us decrease the budget L so that only CR and CNR are implementable, and let the condition $\alpha_{BC}(p_{\mathrm{CNR}}(0)) \geqslant \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$ be satisfied: CR is now preferred to CNR (refer again to Fig. 4). Since p_1 does not change as the budget L continues to shrink, CR continues to be the optimal outcome of the policy until the budget is so poor that $p_{\mathrm{CNR}}(0)$ cannot be obtained even working with a single team, i.e. until $g(L - l_0) < p_{\mathrm{CNR}}(0)$. At this point, since a CR outcome cannot be chosen, we revert to a CNR outcome, which is implemented at the tangency solution $p = p_1$. Hence, decreasing budgets L induce the sequence NC–CR–CNR of policy outcomes.[25]

4.3.1. Asymmetric cartels

We have assumed in this paper perfectly symmetric cartels, characterized in each industry by the same relative profits from collusion $(\Pi_M - \Pi_N)$ and from deviation $(\Pi_D - \Pi_N)$, to simplify the model. Considering asymmetric industries might break stationarity, and make the model much more complex.

In Motta and Polo (1999), where we use a slightly simpler setting, relaxing the assumption of symmetric cartels does not change the qualitative results of the analysis. Heterogeneous cartels are described in that paper by assuming a distribution of types with respect to the variables $(\Pi_M - \Pi_N)$ and $(\Pi_D - \Pi_N)$. We obtain there that not all the cartels choose the same equilibrium strategy for given policy parameters. The (α, p) space can be divided in five regions: three of them correspond to all the cartels choosing the same NC, CR or CNR strategy, as in the symmetric case. However, there are also policy parameters that induce a CNR/CR or a CNR/NC equilibrium, with the more profitable cartels choosing in both cases to collude and not reveal. In these mixed regions, the proportion of cartel types that choose the CNR strategy depends on the policy parameters. The design of the optimal policies takes into account how the marginal type is influenced: we obtain therefore, at the margin, the same welfare improving (firms move from CNR to CR) and welfare decreasing (cartels moving from NC to CR) effects of leniency programs that in a symmetric setting apply to the choices of all the cartels in the economy.

[25] This sequence continues to hold as long as the minimum team size l_0 is high enough, i.e. independent prosecution is not a trivial task. When l_0 is very low, p_1 is low as well and the tangency point occurs at the upper boundary of the $\alpha_{\mathrm{NC}}(p)$ curve: we cannot exclude that NC is implemented close to (or at) the corner solution and that, once the budget line goes below the $\alpha_{\mathrm{NC}}(p)$ curve, the condition $\alpha_{BC}(p_{\mathrm{CNR}}(0)) \geqslant \delta p_1 \alpha_{BC}(p_1)/(1+\delta)$ fails to hold, inducing a sequence NC–CNR.

5. Fine reductions only before the inquiry is opened

As mentioned in the introduction, the initial leniency program introduced in the US in 1978 entitled firms to a reduction in fines only if the cooperation started *before* an inquiry was opened. Similarly, the regime chosen in the EU with the July 1996 Notice applies mainly to firms which reveal information before the AA has opened an official investigation. It is therefore interesting to analyze whether granting leniency treatment only to firms which report before an investigation starts can be justified in terms of enforcement effectiveness. We shall show in this section that this is not the case.

In a 'fine reduction only before the inquiry is opened' regime, the AA initially sets the policy parameters and the LP eligibility rules. Then firms decide to collude or to deviate; in the following period, before an inquiry has been opened, the firms choose whether to reveal the existence of the cartel to the AA; if no firm reveals the AA reviews the cartel with probability α and concludes the prosecution stage in the next period condemning the firms with probability p. Once a case is concluded the game restarts.

In this setting we are interested to check whether both a collude and not reveal (CNR) and a collude and reveal (CR) subgame perfect equilibrium exist. The following proposition shows that only the former one exists.

Proposition 4. *Consider the repeated game played by the firms from $t = 1$ on, for given policy parameters (F, R, α, p) and under the 'reduced fines only before an inquiry is opened' rule. A CR subgame perfect equilibrium never exists. For any fine reduction R, if $\alpha < \alpha_{NC}(p)$ a CNR equilibrium exists. Otherwise, a NC equilibrium exists.*

Proof. See Appendix B. □

This proposition says that under this regime LPs are not effective, since firms will never reveal after having colluded, and the condition which determines whether firms do not collude or collude (and do not reveal) is the same as when LPs do not exist. Everything is as if a fine reduction was not in place. It is not difficult to understand why. Consider the benchmark case where firms were entitled to fine discounts *after* the opening of an investigation. There the expected profit from collusion decreases when the event 'opening of an investigation' realizes, since the probability of being condemned to the full fines jumps up from αp to p. Firms may react, if the rules of the LP allow, revealing information in exchange of reduced fines. Hence, it is the increase in the probability of being fined that triggers, in the benchmark case, firms to reveal after they have started to collude.

If instead collaboration with the AA is rewarded only for reporting before an investigation starts, firms have no incentive to reveal information to the AA after the investigation starts. But they have no incentive to reveal before the investigation starts either. Before observing if a review is opened, the probability of being condemned is still αp, the same as when deciding on colluding or not. If αp is sufficiently high, firms will abstain from collusion; but if it is low enough, they will collude and not reveal. If nothing new happens between the moment firms decide on collusion and the moment they are asked to report to the AA, they will have no incentive to defect.[26]

This result depends in part on the stationary structure of the game we are using, as the discussion suggests. We can find, out of the setting of the model, reasons why even fine reductions that are granted only before an inquiry is opened might be effective.

For instance, if the full fine F depends on the duration of the cartel agreement, firms might find it convenient to start colluding and reveal after some periods even if not monitored (a particular type of CR strategy), in order to avoid the huge cumulated fine they would pay in case of an adverse sentence. However, even in this case extending fine reductions to late revelations (i.e., revelations after an investigation is started) should make LP more effective: when the full fines increase over time, if the firms can pay R after a late revelation, they might reveal if monitored even if the cumulated full fines F are not yet so high to induce early revelation.[27]

The effectiveness of reduced fines granted also to late revelations is consistent with the US experience, where initially the LP was used only for firms which spontaneously offered evidence before the inquiry was opened. In this initial regime the program was quite ineffective while, once allowed in 1993 for reduced fines even after the inquiry was opened, the number of cases in which the firms cooperated with the judges increased significantly.[28]

[26] Malik and Schwab (1991, pp. 30–31) also noticed that the introduction of a tax amnesty alone (without an increase in compliance effort, for instance) will not lead a taxpayer to report additional income: ‘He has already chosen the optimal level of evasion, and if he has not received any additional information, then this optimal level of evasion remains unchanged’.

[27] A further example of non-stationarity: if the perception of the risk of being caught colluding changes over time (or a more risk-adverse management takes over), then firms might report to the AA even if the LPs apply only to revelations which occur before the investigation starts.

[28] In the 1994 Annual Report of the Antitrust Division it is stressed that in the first year of the new regime ‘an average of one corporation per month came forward with information on unilateral conspiracies, compared to an average of one per year under the previous policy. The policy thus allowed the Division to extend the reach of its criminal enforcement activities with relatively little expenditure of resources’ (Antitrust Division, 1994, pp. 6–7). More recently, ‘[a]mnesty applications over the past year have been coming in at the rate of approximately two per month’ (Spratling, 1999, p. 2).

6. Conclusions

In this paper we have analyzed the effects of leniency programs on the incentives of firms to collude and to reveal information that helps the Antitrust Authority to prove illegal behaviour. We have showed that, by reducing the expected fines, leniency programs may induce a pro-collusive reaction: combinations of policy parameters which, without leniency programs, would prevent collusion, may induce firms to collude (and reveal if monitored) when fine reductions are given. Hence, if the resources available to the AA are sufficient to prevent collusion using full fines, leniency programs should not be used.

However, when the AA has limited resources, leniency programs may be optimal in a second best perspective. Fine reductions, inducing firms to reveal information once an investigation is opened, increase the probability of ex-post desistence and save resources of the AA, thereby raising welfare. The optimal scheme requires maximum fine reduction, that is, the firms that collaborate with the Authority should not pay any fine.

We have then showed that allowing fine reductions only to firms which report to the Antitrust Authority *before* an inquiry is opened (as initially established in the US policy in 1978, and in the spirit of the EU Notice on the non-imposition of fines), is inferior to a regime where firms are entitled to fine discounts even if they reveal *after* an inquiry is opened.

We believe that, despite the simple setting, our paper sheds some light on the desirable features of actual leniency programs. In particular, our analysis indicates that a leniency program should be *equally* applicable (and generous) to information disclosed before and after an investigation has started. The US experience, where after the 1993 policy revision a corporation is granted leniency after an investigation has begun, shows that the extension of the leniency program to post-investigation amnesty is a crucial ingredient for success.

Acknowledgements

We benefited from discussions with and comments from Luis Cabral, Chiara Fumagalli, Federico Ghezzi, Jennifer Reinganum, Patrick Rey, Giancarlo Spagnolo, Thomas Von Ungern-Sternberg, two anonymous referees, and seminar participants at the University of Salerno, Bocconi University, the European University Institute and the 1999 EARIE Meeting.

Appendix A. The budget constraint

In Appendix A we formally prove the main properties of the budget constraint discussed in Section 4.1. Since $\alpha_{BC}(p) = l_N/(l_0 + f(p))$, in the set $(\alpha, p) \in [0, 1]^2$, depending on the value of $l_N = L/N$, i.e. of the AA total budget L, the budget constraints intersects the boundaries at the following points: when $l_N < l_0$ we have a vertical intercept for $p = 0$ at $\alpha_{BC}(0) = l_N/l_0$; when $l_N \geqslant l_0$ the budget constraint starts at $\alpha = 1$ and $p = g(l_N - l_0)$. The budget constraint is downward sloping, with slope

$$\frac{\mathrm{d}\alpha_{BC}}{\mathrm{d}p} = -\frac{l_N f'(p)}{(l_0 + f(p))^2} = -\frac{\alpha\epsilon_p}{p} < 0$$

and the second derivative of $\alpha_{BC}(p)$ is:

$$\frac{\mathrm{d}^2\alpha_{BC}}{\mathrm{d}p^2} = l_N \frac{2[f'(p)]^2 - f''(p)[l_0 + f(p)]}{[l_0 + f(p)]^3} \tag{A.1}$$

Since $\lim_{p\to 0} f'(p) = 0$ and $f'' > 0$, for low values of p the curve is concave, while it becomes convex when p becomes large.

Finally, since $l = l_0 + f(p)$ is increasing and convex with vertical intercept at $l = l_0$, $\epsilon_p \geqslant 1$, i.e. $f' > l/p$, for $p \geqslant p_1$and $\epsilon_p < 1$ for $p < p_1$. Hence, the budget constraint is flatter to the left, and steeper to the right of $p = p_1$. Finally, since p_1 is such that $f'(\ p_1) = (l_0 + f(p_1))/p_1$ it is easy to check that p_1 does not depend on l_N and the total budget L while it is increasing in l_0.

Appendix B. Proofs

Proof: Proposition 1

Let us consider the different curves $\alpha_{\mathrm{NC}}(p)$, $\alpha_R(p, R)$, $\alpha_{\mathrm{CR}}(R)$, $p_{\mathrm{CNR}}(R)$ and $p_R(R)$ for different values of the reduced fine, starting from $R = F$. When R is close to F, $\alpha_{\mathrm{NC}}(p) < \alpha_R(p, R)$ for $p \in [0, 1]$ and $p_{\mathrm{CNR}}(R) > p_R(R) > 1$ while $\alpha_{\mathrm{CR}}(R)$ is close to zero. Hence, CNR is preferred to CR—$p_{\mathrm{CNR}}(R) > p_R(R)$—and the NC region is bounded below by $\alpha_{\mathrm{NC}}(p)$. For lower values of R, $p_R(R) < 1 < p_{\mathrm{CNR}}(R)$, which means that a CR region exists above the curve $\alpha_R(p, R)$ (where NC is preferred to CNR) and below $\alpha_{\mathrm{CR}}(R)$ (where CR is preferred to NC). Decreasing further the reduced fine R we get $p_R(R) < p_{\mathrm{CNR}}(R) < 1$: for $p \in [p_R(R), p_{\mathrm{CNR}}(R)]$ the CR region is delimited by $\alpha_R(p, R)$ and $\alpha_{\mathrm{CR}}(R)$ as before, while for $p \in [p_{\mathrm{CNR}}(R), 1]$ CR is preferred to both CNR and NC when $\alpha < \alpha_{\mathrm{CR}}(R)$. Finally, when $R < \delta(\Pi_D - \Pi_N)$ we have $p_{CNR}(R) < p_R(R) < 1$ and CR is preferred to NC and CNR for $\alpha < \alpha_{\mathrm{CR}}(R)$ and $p \in [p_{\mathrm{CNR}}, 1]$, with a rectangular region CR. The conditions stated in the Proposition summarize these different cases. □

Proof: Proposition 2

We characterize the optimal policies to implement NC, CNR and CR. Any point above the $\alpha_{\mathrm{NC}}(p)$ curve is equivalent in terms of welfare gains, allowing to completely deter collusion. We select a point on the boundary of the NC region, i.e. along the curve $\alpha_{\mathrm{NC}}(p)$ which allows to save resources, i.e. to implement NC with the minimum budget. Suppose that a tangency point exists between $\alpha_{\mathrm{NC}}(p)$ and $\alpha_{BC}(p)$. The slope of the $\alpha_{\mathrm{NC}}(p)$ curve is

$$\frac{\partial \alpha_{\mathrm{NC}}}{\partial p} = -\frac{(1+\delta)(\Pi_M - (1-\delta)\Pi_D - \delta\Pi_N)}{\delta^2 p^2(\Pi_M - \Pi_N + F)} = -\frac{\alpha}{p} \tag{B.1}$$

Since the slope of $\alpha_{BC}(p)$ is $-\alpha\epsilon_p/p$, the tangency occurs when $\epsilon_p = 1$, i.e. at $p = p_1$. Since $\epsilon_p > 1$ for $p > p_1$ and $\epsilon_p < 1$ for $p < p_1$, the budget constraint curve is below the $\alpha_{NC}(p)$ curve for any other p. Hence, $p = p_1$ and $\alpha = l_N/(l_0 + f(p_1))$ are the optimal policy. If l_0 is very low it may be that $\alpha_{\mathrm{NC}}(p_1) > 1$. In this case[29] the optimal policy entails the corner solution $\alpha_{\mathrm{NC}}(p) = 1$.

The implementation of a CNR equilibrium is very similar: since the slope of the iso-welfare gains curves, all of them equilateral hyperboles, is $-\alpha/p$ while that of the budget constraint curve is $-\alpha\epsilon_p/p$, the optimal policy is at the tangency point $p = p_1$. For very low values of l_0 it may be that $\alpha_{BC}(p_1) > 1$, i.e. in the CNR region the budget constraint is always steeper than the iso-welfare curves. In this case we have a corner solution at $\alpha = 1$ and $p = g(l_N - l_0)$. Finally, for low values of the budget L it may be that $g(L - l_0) < p_1$, which implies that the budget constraint, for all $p \leqslant g(L - l_0)$ is always flatter than the iso-welfare gains curves. In this case the corner solution entails $p = g(L - l_0)$ and $\alpha = l_N/(l + f(p))$.

Finally, let us consider the optimal implementation of a CR outcome. The conditions stated in the Proposition imply that the budget constraint passes through the CR region. By choosing R we determine how wide is this area. Since the iso-welfare gains curves are horizontal and the budget constraint is downward sloping, we'll pick up a corner solution at the left boundary of the CR region. Setting $R = 0$ we get the widest region, i.e. the lowest $p_{\mathrm{CNR}}(R)$ and the highest α along the budget constraint in the CR region. The CR outcome can be implemented as long as $p_{\mathrm{CNR}}(0)$ is feasible along the budget constraint, i.e. if $g(L - l_0) \geqslant p_{\mathrm{CNR}}(0)$. □

Proof: Proposition 3

If the budget constraint is tangent or intersects the lower boundary of the NC region, the NC outcome can be implemented. Since it is associated with the highest welfare gains, this is the optimal policy. If the $\alpha_{NC}(p)$ curve is always above the budget constraint, we have to choose between two possible outcomes, CR and CNR. If a CNR outcome is implemented, we have to choose $p = p_1$ and $\alpha = \alpha_{BC}(p_1)$. The welfare gains are $W^m = NK\delta p_1 \alpha_{BC}(p_1)/(1+\delta)$. The same level

[29] The case $p_1 = 1$, which would induce a corner solution at $p = 1$, is not relevant as it would require a minimum team of infinite dimension.

of welfare W^m can be obtained in a CR equilibrium setting $\alpha = W^m/NK$. Solving for α we obtain

$$\alpha^m = \frac{\delta p_1 \alpha_{BC}(p_1)}{1+\delta} \tag{B.2}$$

Hence, the CR outcome induced by $(\alpha^m, p_{\mathrm{CNR}}(0))$ is welfare equivalent to the CNR outcome $(p_1, \alpha_{BC}(p_1))$. To select one of the two outcomes, we need to check whether the budget constraint passing through $(p_1, \alpha_{BC}(p_1))$ allows to implement a CR outcome preferable to $(\alpha^m, p_{\mathrm{CNR}}(0))$. If $\alpha_{BC}(p_{\mathrm{CNR}}(0)) \geqslant \alpha^m$, by moving into the CR region along the budget constraint we can implement (at least) an equivalent CR outcome. This case is shown in Fig. 5, where the welfare level $W^M > W^m$ can be attained in the CR region. If on the contrary the budget constraint passes below α^m at $p = p_{\mathrm{CNR}}(0)$, the CNR outcome is preferred. Finally, the CNR outcome is selected if CR cannot be obtained, i.e. if the highest probability $g(L - l_0)$ is lower than $p_{\mathrm{CNR}}(0)$. □

Proof: Proposition 4

To find the conditions under which not revealing is an equilibrium, we have to check two incentive constraints. The first requires that a firm prefers to collude and not reveal rather than deviate, and is the same as in the benchmark case analyzed in the first part of the paper. It amounts to the condition $\alpha \leqslant \alpha_{\mathrm{NC}}(p)$. The second incentive constraint imposes that the firm, after having joined the cartel, prefers not to reveal rather than reveal at the beginning of period 1, just before an inquiry may be opened. Notice that if firms do not reveal and an inquiry is opened (with probability α) during the same period, the prosecution stage will occur in the next period with firms proved guilty with probability p; after two periods the game restarts. Then, the value of the game if firms do not reveal is

$$\begin{aligned} V_{\mathrm{NR}}|\alpha &= \Pi_M + \delta\alpha[p(\Pi_N - F) + (1-p)\Pi_M] + \delta(1-\alpha)\Pi_M + \delta^2 V_{\mathrm{NR}}|\alpha \\ &= V_{\mathrm{CNR}} \end{aligned} \tag{B.3}$$

i.e. the value of the game is exactly the same that summarizes the expected value of a CNR equilibrium. The value of the game if the firm reveals, breaking the agreement, is

$$V_R|\alpha = \frac{\Pi_N}{1-\delta} - R < \Pi_D + \frac{\delta \Pi_N}{1-\delta} = V_D \tag{B.4}$$

Hence, if the first incentive holds, $V_{\mathrm{CNR}} \geqslant V_D$, the second constraint $V_{\mathrm{NR}}|\alpha \geqslant V_R|\alpha$ never binds. □

References

Andreoni, J., 1991. The desirability of a permanent tax amnesty. Journal of Public Economics 45, 143–159.

Antitrust Division, 1994. Annual Report for Fiscal Year 1994.

Arlen, J., Kraakman, R., 1997. Controlling corporate misconduct: an analysis of corporate liability regimes. New York University Law Review 72 (4), 687–779.

Das-Gupta, A., Mookherjee, D., 1996. Tax amnesties as asset-laundering devices. The Journal of Law, Economics & Organization 12 (2), 408–431.

Easterbrook, F.H., Landes, W.M., Posner, R.A., 1980. Contribution and claim reduction among antitrust defendants: a legal and economic analysis. The Journal of Law and Economics, 331–370.

European Union, 1996. Notice on the non-imposition or reduction of fines in cartel cases. Official Journal 207, 4.

Friedman, J., 1971. A non-cooperative equilibrium for supergames. Review of Economic Studies 38 (113), 1–12.

Grossman, G.M., Katz, M.L., 1983. Plea bargaining and social welfare. American Economic Review 73 (4), 749–757.

Innes, R., 1999. Remediation and self-reporting in optimal law enforcement. Journal of Public Economics 72 (3), 379–393.

Kaplow, L., Shavell, S., 1994. Optimal law enforcement with self-reporting of behavior. Journal of Political Economy 102 (3), 583–606.

Kobayashi, B.H., 1992. Deterrence with multiple defendants: an explanation to unfair plea bargains. RAND Journal of Economics XXIII (4), 507–517.

Kornhauser, L.A., Revesz, R.L., 1994. Multidefendant settlements under joint and several liability: the problem of insolvency. Journal of Legal Studies 23, 517–542.

Landes, W.M., 1971. An economic analysis of the courts. Journal of Law and Economics 14, 61–108.

Livernois, J., McKenna, C.J., 1999. Truth or consequences: enforcing pollution standards with self-reporting. Journal of Public Economics 71, 415–440.

Malik, A.S., 1993. Self-reporting and the design of policies for regulating stochastic pollution. Journal of Environmental Economics and Management 24, 241–257.

Malik, A.S., Schwab, R.M., 1991. The economics of tax amnesties. Journal of Public Economics 46, 29–49.

Motta, M., Polo, M., 1999. Leniency programs and cartel prosecution. Working Paper ECO No. 99/23, European University Institute.

Polinsky, M.A., Shavell, S., 1981. Contribution and claim reduction among antitrust defendants: an economic analysis. Stanford Law Review 33, 447–471.

Reinganum, J.F., 1998. Plea bargaining and prosecutorial discretion. American Economic Review 78 (4), 713–728.

Rey, P., 2000. Towards a theory of competition policy. Mimeo, available at http://www.univ-tlse1.fr/idei/Commun/Articles/Rey/seattle1026.pdf

Schrag, J., Scotchmer, S., 1997. The self-enforcing nature of crime. International Review of Law and Economics 17, 325–335.

Spagnolo, G., 2000. Optimal leniency programs. Mimeo, Stockholm School of Economics.

Spratling, G.R., 1999. The corporate leniency policy: answers to recurring questions. Speech of the Deputy Assistant Attorney General, presented at the Bar Association of the District of Columbia, February 16, 1999.

Tirole, J., 1988. The Theory of Industrial Organization. MIT Press, Cambridge, MA and London.

Tokar, S., 2000. Whistleblowing and corporate crime. Mimeo, European University Institute.