Rev Ind Organ (2018) 52:403–427
https://doi.org/10.1007/s11151-017-9586-8

Optimal Leniency Programs When Firms Have Cumulative and Asymmetric Evidence

Marc Blatter[1] · Winand Emons[2] · Silvio Sticher[3]

Published online: 29 July 2017

Abstract An antitrust authority deters collusion with the use of fines and a leniency program. Firms have imperfect cumulative evidence of the collusion. That is, cartel conviction is not automatic if one firm reports. Reporting makes conviction only more likely: the more that firms report, the more likely is conviction. Furthermore, the evidence is distributed asymmetrically among firms. This set-up allows us meaningfully to analyze three typical features of leniency programs: minimum-evidence standards; ringleader discrimination; and marker systems. Minimum-evidence standards provide high-evidence firms with proper incentives to report. They are better at deterring than is ringleader discrimination. Under a marker system only one firm reports so that the antitrust authority never gets the entire available evidence. Appropriate minimum-evidence standards make a marker system redundant.

Keywords Antitrust · Cartels · Deterrence · Leniency · Evidence

We thank Andreas Bachmann, Juan Beccuti, Caroline Buts, Giacomo Calzolari, Alain Egli, Rodrigo Londoño, Evgenia Motchenkova, Tom Ross, Giancarlo Spagnolo, Tobias Wenzel, Larry White, and two referees for helpful comments. The usual disclaimer applies.

✉ Winand Emons
winand.emons@vwi.unibe.ch

Marc Blatter
marc.blatter@vwi.unibe.ch

Silvio Sticher
silvio.sticher@sbb.ch

[1] Universität Bern, Swiss National Bank, Bern, Switzerland

[2] Departement Volkswirtschaftslehre, Universität Bern, CEPR, Swiss Competition Commission, Schanzeneckstrasse 1, Postfach 8573, 3001 Bern, Switzerland

[3] Swiss Federal Railways, Bern, Switzerland

JEL Classification D43 · K21 · K42 · L40

1 Introduction

A corporate leniency program reduces the sanctions for self-reporting cartel members. In 1993 the US Department of Justice significantly clarified its first-informant rule that guarantees amnesty for the first reporting firm even when an investigation has already started. This program has been so effective that many other countries followed suit.

The theoretical literature on leniency typically makes the following two assumptions: first, if one firm reports the illegal behavior, the cartel is convicted for sure; each firm thus possesses perfect evidence. Inducing one firm to report, no matter which one, is sufficient for the antitrust authority (AA) to convict the cartel. Second, if each firm has perfect evidence, the distribution of the evidence is automatically symmetric.

We relax both assumptions: our firms have imperfect and cumulative evidence. The evidence of one firm increases the probability of conviction, but not necessarily to one; the evidence of two firms leads to a higher probability of conviction than the evidence of one firm alone. With imperfect evidence we can allow for asymmetric evidence: one firm may have more evidence than does another firm. This framework allows us meaningfully to analyze minimum-evidence standards, a marker system, and ringleader discrimination—three typical features of leniency programs.

We first show that without these features the asymmetry of the evidence may indeed make it more difficult to deter collusion as compared to symmetric evidence. The high-evidence firm may prefer to remain silent while the low-evidence firm reports.

We then look at a minimum standard of evidence which a firm has to meet to get leniency. A standard such that the high-evidence firm qualifies while the low-evidence firm does not lowers the AA's enforcement cost. It provides the high-evidence firm with strong incentives to report.

Next we analyze a marker system. After a firm applies for the marker, the AA informs the firm about its position in line. If the firm is first in line, it gets the marker and thus receives leniency if it chooses to report. The marker system allows firms to report conditionally: they reveal their information if the marker is available and they do not reveal if the marker is no longer available. When both firms conditionally report, the AA gets only one report: it gets less evidence than in the no-marker case where both firms report. If firms are sufficiently asymmetric, the marker can, however, be useful. Yet, with appropriate minimum-evidence standards the marker system is ineffective. If only the high-evidence firm qualifies for leniency, it can be sure to avoid the fine, and a marker is not necessary for that privilege.

Finally, we assume that the high-evidence firm is the ringleader of the cartel and analyze the effects of denying leniency to the ringleader. If the ringleader's evidence is similar to the evidence of the other cartel members, ringleader discrimination lowers deterrence costs: without leniency it is less attractive for the ringleader to collude in the first place. Yet, if the ringleader has sufficiently more evidence than the other firms, ringleader discrimination is ineffective: the ringleader plans not to report anyway. Ringleader discrimination is less effective than minimum-evidence standards: shutting out the high evidence firm from leniency generates less deterrence than barring the low evidence firms.

1.1 Related Literature

Our paper builds on the analysis of leniency programs by Motta and Polo (2003), Spagnolo (2003), Aubert et al. (2006), and Harrington (2008).[1] This literature analyzes the effects of leniency on the frequency of collusion and derives optimal fine structures.

Our basic set-up is closest to Motta and Polo (2003). Besides in the nature of evidence, our framework differs from theirs in two other respects: first, we focus mainly on cartel deterrence. By contrast, they also consider the possibility that a leniency program may temporarily interrupt collusion which we deal with briefly in the Appendix. Second, in our setting the AA fully deters at minimum cost whereas in Motta and Polo (2003) the AA faces a budget constraint so that it may not be able to achieve complete deterrence.

The novel feature of our set-up is the cumulative and asymmetric nature of the evidence. Imperfect evidence has already been addressed, yet from a different angle. In Aubert et al. (2006) firms can destroy their evidence, thereby reducing the cartel's risk of being convicted. In our setting, the firms' evidence is exogenously given and if a firm blows the whistle it has to reveal everything. Yet, our firms may not possess all the evidence necessary that is to convict the cartel.

In Harrington (2008) the AA's stand-alone evidence varies over time, which affects the value of additional evidence that is provided by the firms. However, as in all of the above-mentioned papers, a cartel is convicted for sure as soon as one member turns in. By contrast, in our setting a firm that submits evidence increases the probability of conviction, but not necessarily to one.

Feess and Walzl (2010), Silbye (2010), Herre et al. (2012), and Charistos and Constantatos (2016) address asymmetric evidence. As in our framework, Feess and Walzl (2010) and Silbye (2010) consider two firms that possess different amounts of evidence. The AA chooses a fine structure that decreases with the amount of provided evidence. We touch on the issue of the optimal fine structure by considering minimum standards of evidence and ringleader discrimination. Nevertheless, following the winner-takes-all characteristics of the US leniency program, in our setting firms either get full or no leniency at all. We do not consider leniency as depending in a more subtle way on the amount of evidence to be of further interest due to the difficulties to implement such rules in practice.

Charistos and Constantatos (2016) study the effects of a marker system. In their set-up the AA does worse with a marker system than without because with a marker it does not get the entire available evidence.

As to ringleader discrimination, in Herre et al. (2012) the ringleader possesses perfect information, while the other cartel members possess incomplete (and

[1] Further theoretical research includes Cyrenne (1999), Harrington and Chang (2009), Harrington (2013), Sauvagnat (2015), and Harrington and Chang (2015); empirical and experimental research includes Bigoni et al. (2012, 2015), Brenner (2009), and Miller (2009). For a survey, see Spagnolo and Marvão (2016).

identical across firms) evidence. They analyze whether or not the ringleader should be granted leniency. Similar to our results they find that excluding the ringleader from leniency has a deterring effect when firms are sufficiently symmetric. A further comparison of their results with ours is difficult because they allow for side payments between members to stabilize the cartel, which we do not.

In Chen et al. (2015) the ringleader and his follower both posses perfect evidence. Ringleader discrimination can lead to increased or decreased levels of cartel conduct. On the one hand, ringleader discrimination undoes some of the destabilizing benefit of the leniency program. On the other hand, under discrimination the ringleader faces a more severe punishment which can reduce the incentive to instigate in the first place.[2]

To the best of our knowledge, we are the first to analyze minimum standards of evidence formally.

1.2 Institutional Background

In both the US and the EU, the institutional details of a leniency program are not determined by competition law, but instead by the internal policies of the AAs.[3] This implies that the AA can revise its leniency program as to what it sees fit.

Typically, leniency programs include the following features: (1) the assignment of leniency to the first applicant who reports participation in a cartel;[4] (2) a marker system that allows an applicant to secure its position in line; (3) the requirement of full disclosure of evidence; (4) an ongoing requirement to cooperate fully with the AA; (5) a requirement to cease collusive behavior; and (6) ringleader discrimination.

We take all of these features into account. In line with the US leniency program, in our model only the first evidence-providing firm receives leniency. We require the firms to reveal all evidence in their possession. Furthermore, a reporting firm has to cease its collusive conduct.

Both the US and the EU leniency programs specify requirements with respect to the evidence that an applicant has to provide. The Antitrust Division of the US Department of Justice requires the applicant to report “the wrongdoing with candor and completeness” and to provide “cooperation that advances the Division in its investigation.”[5] In Sect. 4, we deal with these requirements by introducing minimum standards of evidence.

[2] Perfect and symmetric evidence is a special case of our set-up. Yet, we do not find a negative effect of ringleader discrimination. This is due to the different sequencing. In Chen et al. (2015) firms set quantities and later decide whether to reveal or not. In our framework firms decide simultaneously about quantities and their reporting strategy; these decisions are thus perfectly coordinated. Therefore, we do not encounter many of the subgames that Chen et al. have to deal with.

[3] See US Department of Justice (2008) and European Commission (2006).

[4] The US system does not allow for leniency for a second-reporting firm. By contrast, the EU program offers reduced leniency also to all other firms that are not first to come forward, provided that the additional information is sufficiently valuable.

[5] See US Department of Justice (2008). Likewise, the European Commission requires that the evidence has to enable the Commission to find an infringement of Article 101 TFEU (European Commission 2006).

We also analyze the effects of a marker system: a standard element of most leniency programs, under which the AA informs the firm whether it is the first to seek leniency.[6] Common arguments for its use are legal certainty and transparency, and that it encourages a "race to the courthouse." In possession of the marker, a firm has, e.g., 30 days to collect the evidence necessary to "perfect the marker." For reasons of tractability, we ignore the time dimension of the marker system. In our model, the AA immediately informs a firm whether leniency is available.

In the US it is not possible for ringleaders to obtain a fine reduction. By contrast, in the EU ringleaders benefit from leniency as long as the undertaking did not take "steps to coerce other undertakings to join the cartel or to remain in it" (European Commission 2006, p. 13). We deal with the effects of ringleader discrimination in Sect. 6.

The rest of this paper is organized as follows: the next section describes the model. In Sect. 3 we derive the equilibria for the reference scenario. In Sect. 4 we extend the model to minimum standards of evidence, in Sect. 5 to the marker system, and in Sect. 6 to ringleader discrimination. Section 7 concludes.

2 Model

Each industry consists of two potentially colluding firms. The legislator specifies the antitrust framework that we take as exogenously given. Within this framework the AA chooses its policy so as to deter any collusion at minimal cost.

At the outset the legislator announces the fine $F > 0$ that a convicted firm pays whenever it communicated with the other firm in the period under consideration. The legislator grants leniency to the first reporting firm. To receive leniency, the reporting firm has to provide evidence of the conspiracy and it has to stop the collusive conduct immediately. If both firms choose to report, nature determines with equal probability who is first. Accordingly, in expectation each firm obtains half the leniency. We look at the case of full leniency so that the reporting firm ends up with no fine while the non-reporting firm pays F; if both firms report, each of them pays in expectation F / 2.

The AA's choice consists of two elements: first, the AA starts an investigation with probability $\alpha \in [0, 1]$; second, the AA decides how much effort $p \in [0, 1]$ it puts into the investigation.[7] This effort gives rise to the probability P of detecting and convicting a cartel; we will specify P as we move along. Moreover, in Sect. 4 the AA communicates the minimum standard of evidence that guarantees leniency. In Sect. 5 it announces the use of a marker system, and in Sect. 6 it establishes ringleader discrimination.

[6] "The Division frequently gives a leniency applicant a "marker" for a finite period of time to hold its place at the front of the line for leniency while counsel gathers additional information through an internal investigation to perfect the client's leniency application. While the marker is in effect, no other company can 'leapfrog' over the applicant that has the marker" (Department of Justice 2008).

[7] Actually, the AA determines α and p by the choice of the size and the allocation of its personnel; see Sect. 2.1.

Then an infinitely repeated game starts. The stage game in each period $t = 0, \ldots$ has the following structure: knowing α and p, first firm i, $i = a, b$, decides whether it wants to communicate with the other firm or not. If both firms choose to communicate, they create evidence that —if detected— leads to a conviction by the AA; unless both firms communicate, they do not engage in illegal behavior, and there is thus no evidence thereof.[8] The evidence dissolves at the end of the period.[9]

Then the AA starts an investigation with probability α, which leads to the investigation subgame; with probability $(1 - \alpha)$ the game continues with the no-investigation subgame. In both subgames the firms choose "classical" economic conduct variables such as prices under Bertrand competition, quantities under Cournot competition, or where to sell under exclusive territories. We will use prices under Bertrand competition to describe our framework. Simultaneously, the firms decide whether they report any communication (R) or not (N); firms make this decision knowing whether an investigation is ongoing or not.

If there is an investigation, the firms' reporting behavior together with the AA's investigation effort determines the probability of conviction,

$$P = \begin{cases} p, & \text{if no firm reports;} \\ p_i, & \text{if firm } i \text{ reports, } i = a, b; \\ p_2, & \text{if both firms report.} \end{cases}$$

To derive the specific form of P, suppose that the firms' communication has created sufficient evidence that, if totally uncovered, results in a certain conviction. Through its choice of effort the AA uncovers evidence; we describe the details thereof in the next section. Let p be the probability of conviction based on the AA's stand-alone evidence.

Besides the AA, each firm also possesses evidence. Let ρ_i denote the probability of conviction generated by firm i's stand-alone evidence, $i = a, b$. Let the pieces of evidence be independently distributed.[10] Then we have for the joint evidence (AA and one firm, or AA and both firms)

$$\begin{aligned} p_i &:= 1 - (1 - p)(1 - \rho_i), \quad i = a, b; \\ p_2 &:= 1 - (1 - p)(1 - \rho_a)(1 - \rho_b). \end{aligned}$$

If, say, only firm b reports, $p_b = p + \rho_b - p\rho_b$; in terms of evidence, expected total evidence is given by the evidence that is uncovered by the AA plus the evidence that is provided by firm b, minus the expected value of the evidence that is in joint possession.

[8] The AA thus does not make type I errors, i.e., punish non-colluding firms. See, e.g., Block and Sidak (1980) for a discussion of antitrust enforcement when courts make errors.

[9] Communication, even when it is not followed by anti-competitive behavior, is considered illegal.

[10] The assumption of independence is less restrictive than it perhaps seems. Suppose all of the evidence is in the firms' possession, scattered on hard drives. Through a dawn raid the AA seizes the hard drives. Yet, the relevant pieces of information are hidden in tons of terabytes that the AA has to scan tediously using its personnel. For example, in the recent Libor/Euribor/Tibor cases, the Swiss Competition Commission (COMCO) scanned more than 9 million pages of communication (www.newsd.admin.ch/newsd/message/attachments/46714.pdf). We rule out strategic destruction of evidence; see also Footnote 25. Accordingly, a firm submits either all of its available evidence, or it does not report at all.

Note that we have $p \le p_i \le p_2,\ i = a, b$, and subadditivity $p_2 \le p_a + p_b$. Furthermore, the probability of conviction is equal to one if and only if either the AA has perfect evidence $(p = 1)$ or at least one of the firms submits perfect evidence $(\rho_i = 1,\ i = a, b)$. Put differently, if $p < 1$ and both firms submit imperfect evidence, the probability of conviction is less than one.

We think $p_2 < 1$ rather than $p_2 = 1$ is the interesting case to consider. Leniency is an instrument for gathering evidence to prove an infringement.[11] The AA requires that the reporting firm submits all of the evidence in its possession. Yet, it does not require that the reporting firm pleads guilty. The AA still has to prove an infringement, the success of which is typically not guaranteed.[12] Moreover, the reporting firms do not have to waive their right to appeal the AA's decision; an appellate court may overturn its ruling.[13]

Without a loss of generality we assume $\rho_a \le \rho_b$, which implies $p_a \le p_b$. There are various reasons why firms may possess evidence of different quality. The employees of one firm may be particularly diligent in documenting their communication related to the cartel.[14] Furthermore, a cartel member might have little evidence because an employee involved in the cartel has left the firm in the meantime and all his documents containing information about the cartel have disappeared.[15] Finally, if we view our two-member model as a shortcut for a larger cartel, the more informed firm could be the ringleader that coordinates the other participants.[16] Thus, in practice, the overlap of the evidence in possession of the firms is typically not perfect and the AA has to rely upon the evidence of different firms in order to prove an infringement.[17]

[11] By contrast, settlement is a tool to speed up the procedure to reach a decision. Both the US Department of Justice and the European Commission require that the parties plead guilty to settle a cartel case.

[12] The US Department of Justice requires that leniency applicants "confess participation in a criminal antitrust violation" (Department of Justice 2008). However, the Department of Justice has to carry out the investigation and prove an infringement.

[13] In the air cargo cartel Lufthansa received full immunity from fines under the European Commission's leniency program because it was the first to provide information about the cartel (www.europa.eu/rapid/press-release_IP-10-1487_en.htm?locale=en). Nevertheless, Lufthansa filed an appeal "based on legal considerations" (www.bloomberg.com/news/2011-01-27/japan-airlines-appeals-48-8-million-antitrust-fine-at-eu-court.html).

[14] In the road construction bid-rigging cartel of the Swiss Canton of Argau one cartel member had painstakingly kept records of the ex ante agreed upon bids; see www.weko.admin.ch/aktuell/00162/index.html?lang=fr....

[15] In the gas insulated switchgear case the leniency applicant ABB informed the European Commission about a market sharing agreement between European and Japanese producers. In its decision the Commission relies not only on the evidence submitted by ABB, but additionally on the statements of Mr M., a former employee who had represented ABB at the operational level during the cartel period. Accordingly, ABB was not in possession of all the evidence due to staff turnover; see curia.europa.eu/juris/document/document.jsf?text=&docid=80216&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1.

[16] For more on this argument see Kobayashi (1992) and Herre et al. (2012).

[17] In the airfreight cartel decision the European Commission writes: "Accordingly, many contacts which do not amount to decisive evidence of an infringement in themselves are nevertheless relevant, when assessed with other contacts, to establishing the single and continuous infringement"; see www.ec.europa.eu/competition/antitrust/cases/dec_docs/39258/39258_7008_7.pdf.

Next let us describe how firms collude: by communicating, firms fix prices: they set the monopoly price q_M. This leads to a stage profit of π_M for each firm, which is half the monopoly profit of the industry. Firms support the collusive behavior with grim-trigger strategies. If a firm deviates from q_M, Nash punishment with price q_C equal to marginal cost starts and continues forever; each firm makes the static Nash profit $\pi_C = 0$. If a firm deviates while the other firm colludes, the deviating firm slightly undercuts q_M with q_D and its profit is $\pi_D = 2\pi_m$; the non-deviating firm's profit is $\underline{\pi} = 0$.[18]

The cartel is stable in the absence of the AA. If δ denotes the firms' common discount rate, this means that $\pi_M/(1-\delta) > 2\pi_M$, or $\delta > 1/2$: getting π_M forever is better than getting $2\pi_M$ in the first round and from then on nothing. Recall that to receive leniency, a reporting firm has to cease its collusive conduct. Therefore, reporting automatically triggers punishment because the reporting firm has to cease setting q_M.[19]

2.1 Enforcement Technology

Legislation stipulates the fine F as well as full leniency. The AA strives to deter all cartels at minimum cost. It chooses the size of its staff L; normalizing the wage to 1, L also measures the AA's cost. Given N industries, the AA determines the $n \leq N$ ones to be investigated: $\alpha = n/N$. The AA then allocates L / n staff to each industry.[20] The stand-alone probability of conviction p increases linearly with the personnel that is allocated to that industry:

$$p(L,n) = \begin{cases} \gamma L/n, & \text{if L/n} \leq 1/\gamma; \\ 1, & \text{otherwise,} \end{cases}$$

with $\gamma > 0$. If the AA uses manpower efficiently — $L/n \leq 1/\gamma$ — then $\alpha p = \gamma L/N$. Therefore, the AA's cost is $L = \alpha p N/\gamma =: C(\alpha p)$. With this enforcement technology it is, for example, possible to monitor all industries ($\alpha = 1$) with very low p. This policy implies that the no-investigation subgame is never reached.[21]

To summarize the model:

[18] Using the Bertrand example greatly simplifies the notation because $\underline{\pi} = \pi_C = 0$ and $\pi_D = 2\pi_M$. All of our results hold for the general case with $\underline{\pi} \leq \pi_C < \pi_M < \pi_D$; see the earlier version of the paper: papers.ssrn.com/sol3/papers.cfm?abstract_id=2474998.

[19] In Motta and Polo (2003) firms can also collude on the reporting strategy. They agree ex ante to collude by setting q_M if there is no investigation and by reporting and setting q_C if there is an investigation. After an investigation with agreed upon reporting, they continue to play their collusive strategy. Through an investigation the AA temporarily "interrupts" collusion: what is called "cartel desistence" in the literature. We have chosen the harshest punishment, mainly for reasons of tractability; for a similar approach see, e.g., Harrington (2008). Allowing firms to collude on their reporting strategy makes it more difficult to deter collusion. Our qualitative results do not, however, change. See the Appendix for an analysis of this collude and reveal strategy.

[20] See Motta and Polo (2003) for a similar approach; in their set-up the staff size is exogenously given, so that the AA may not be able completely to deter collusion.

[21] As we will show later, this policy is, however, not optimal.

- The legislator determines the fine F and grants full leniency for the first reporting firm; this antitrust framework is exogenously given.
- The AA announces α and p. (In Sect. 4 it additionally announces the minimum standard of evidence. In Sect. 5 it also announces the use of a marker system, and in Sect. 6 it establishes ringleader discrimination.)
- Then the stage game begins:
  - Firms decide whether they communicate or not.
  - The AA starts an investigation with probability α, which results in an investigation or a no-investigation subgame.
  - In both subgames firms choose prices and whether they report or not.
  - If firms communicated, they are convicted with probability P.

Firms maximize profits with respect to their communication, price, and reporting decision. The AA chooses α and p so as to achieve full deterrence at minimal cost.[22]

3 Equilibrium Analysis

In this section we look at the scenario without a minimum standard of evidence, without a marker system, and without ringleader discrimination: any amount of evidence is sufficient to be granted leniency, and at the time of reporting a firm does not know whether it will get leniency or not. We will first derive the firms' behavior in the investigation and the no-investigation subgames. Then we analyze under which conditions firms collude and how the AA can deter communication. Finally, we determine the cost-minimizing deterrence policy as a function of the fine F.

3.1 The Firms' Behavior

3.1.1 Investigation Subgame

Let us now analyze the subgame that begins after the firms have communicated and the AA has started its investigation. Firms choose their prices. Simultaneously, they decide whether to report or not. As we will see below, their reporting strategies determine their prices. Therefore, we identify the firms' strategies only by their reporting behavior and examine the following equilibrium candidates: both firms report (R, R); both firms do not report (N, N); and one firm reports while the other does not (R, N) and (N, R). We will next determine under which conditions each candidate can indeed be an equilibrium; if several possible equilibria exist, we pick the Pareto-superior one.

If both firms report (R, R), the probability of conviction is p_2. The fact that both firms report triggers competition in all future periods. Therefore, their prices are

[22] Following the literature we assume that the deadweight loss from the cartel exceeds enforcement costs so that full deterrence maximizes total surplus.

determined solely by the stage game, the only equilibrium of which is the competitive one with both firms setting q_C. Accordingly, (R, R) yields $-p_2F/2$ for each firm.

For (R, R) to be an equilibrium, firm b (say) must have no incentive to deviate. If b deviates while a reports, the firms still compete from the next period onward; thus, b sets q_C in the current period. Therefore, if firm b deviates to N, it chooses q_C, which yields 0. Its payoff from deviating is thus $-p_aF$. Firm b pays the fine for sure rather than with probability 1 / 2; in return it reduces the probability of conviction from p_2 to p_a. Accordingly, a necessary condition for (R, R) to be a Nash equilibrium in the subgame is $p_2 \leq 2p_i,\ i = a, b$. Since $p_a \leq p_b$, (R, R) is an equilibrium if

$$p_2 \leq 2p_a \quad \text{or} \quad p \geq 1 - 1/((1-\rho_a)(1+\rho_b)) =: \hat{p}. \tag{1}$$

Note that $\hat{p} \leq 1/2$.

If no firm reports (N, N), the probability of conviction is p. Both firms collude with respect to q_M because this Pareto dominates (N, N) together with q_C. Moreover, if it is optimal to collude in the current period, it is also optimal to not report and collude in future periods. Thus, a firm's payoff is $\pi_M - pF + \sum_{t=1}^{\infty} \delta^t(\pi_M - \alpha pF)$.[23] If a firm deviates to R and q_D, it earns $2\pi_M$ and pays no penalty in the current period. From the next period onward the firms compete so that there is no risk to pay the fine. The firm's deviation profit is thus $2\pi_M$. A necessary condition for (N, N) to be a Nash equilibrium in the subgame is

$$p \leq \pi_M(2\delta - 1)/(1 - \delta + \delta\alpha)F =: p_{N,N}(\alpha). \tag{2}$$

If a reports while b does not (R, N), the probability of conviction is p_a. Firm a's reporting triggers competition in all future periods. Thus, the stage game determines the equilibrium and both firms choose q_C. Since a reports while b does not, a is granted full leniency, which yields a payoff 0. Suppose firm a unilaterally deviates to N. Then competition is still triggered from the next period onwards because firm b plays q_C in the current period. Firm a would thus play N together with q_C. However, this cannot be optimal because a loses the leniency. Firm b obtains $-p_aF$ in (R, N). If b also reports, its payoff is $-p_2F/2$. Accordingly, a necessary and sufficient condition for (R, N) to be played in the subgame is that the expected fine for b when reporting exceeds the expected fine when not reporting, or

$$p_2 > 2p_a. \tag{3}$$

Analogously, as a condition for (N, R) to be played in the subgame we obtain $p_2 > 2p_b$. The assumption $\rho_a \leq \rho_b$ implies that this condition is never satisfied, so that (N, R) is never played. If (1) is satisfied, (3) does not hold. The two equilibrium candidates (R, R) and (R, N) thus exclude each other. Moreover, one of the candidates always exists.

[23] This assumes that firms also play (N, N) in the no-investigation subgame. In the next section we show that this is indeed the case.

Table 1 Payoff matrix of the investigation subgame

		firm b	
		R	N
firm a	R	$-p_2F/2$ $-p_2F/2$	$-p_aF$ 0
	N	0 $-p_bF$	$\pi_M - pF + \sum_{t=1}^{\infty}\delta^t(\pi_M - \alpha pF)$ $\pi_M - pF + \sum_{t=1}^{\infty}\delta^t(\pi_M - \alpha pF)$

If (2) is satisfied, the equilibrium (N, N) also exists, so that the issue of equilibrium selection arises. Recall that for (N, N) to be an equilibrium, the equilibrium payoff must be greater or equal than the deviation payoff $2\pi_M$ which in turn is greater than 0. In (R, R) both firms get $-p_2F/2$. In (R, N) firm a gets 0 and firm b gets $-p_aF$. Both payoffs are less than the payoffs in (N, N) which, therefore, Pareto-dominates all other possible equilibria. Thus, if (2) is satisfied with strict inequality, the firms indeed play (N, N).

Table 1 summarizes the investigation subgame via the reduced normal form:

3.1.2 No-Investigation Subgame

As in the investigation subgame, it suffices to identify the firms' strategies by their reporting behavior. Moreover, not all possible strategies are of interest: if firms choose to report in the no-investigation subgame, they will certainly not communicate in the first place so that there is no need to deter cartel formation. Therefore, we will only derive the necessary conditions for firms not to report in the no-investigation case.

We follow the different cases of the investigation subgame, i.e., we analyze ((N, N), (R, R)); ((N, N), (R, N)); and ((N, N), (N, N)); in each of the three strategy combinations the first pair of strategies is played in the no-investigation subgame and the second pair in the investigation subgame.

Let us start with the case where (R, R) is the equilibrium in the investigation subgame: (1) holds and (2) is violated. Suppose firms play (N, N) together with q_M when there is no investigation. Then the ex-ante expected profit from communicating is

$$\pi((N,N),(R,R)) = \sum_{t=0}^{\infty} \delta^t (1-\alpha)^t (-\alpha p_2 F/2 + (1-\alpha)\pi_M).$$

Firms start with communication and continue to do so if there was no investigation in the preceding period; if there is an investigation, they stop to communicate and play q_C. Now consider the no-investigation subgame. If a firm does not report, it makes profit $\pi_M + \delta\pi((N,N);(R,R))$. If it reports and chooses q_D, its profit is $2\pi_M$. Accordingly, the firms play (N, N) in the no-investigation subgame if

$$\begin{aligned} &2\pi_M < \pi_M + \delta\pi((N,N),(R,R)) \quad \text{or} \\ &\pi((N,N),(R,R)) > \pi_M/\delta. \end{aligned} \tag{4}$$

If (4) is not satisfied, at least one firm reports and picks q_C.

Next consider the case where (*R*, *N*) is the equilibrium in the investigation case: (1) and (2) do not hold. Again, we want to determine under which condition both firms do not report in the no-investigation case. Recall that firm *b* which does not report in the investigation case does worse than the reporting firm *a* which gets leniency. Therefore, if the high-evidence firm *b* plays *N* in the no-investigation subgame, the low-evidence firm *a* will certainly do so too. Suppose that firms play (*N*, *N*) together with q_M when there is no investigation. Then the ex-ante expected profit from communicating for firm *b* is

$$\pi_b((N,N),(R,N)) = \sum_{t=0}^{\infty} \delta^t (1-\alpha)^t (-\alpha p_a F + (1-\alpha)\pi_M).$$

By the same reasoning as above, the firms play (*N*, *N*) in the no-investigation subgame if

$$\pi((N,N),(R,N)) > \pi_M/\delta. \tag{5}$$

Finally, consider the case in which (*N*, *N*) is the equilibrium of the investigation subgame: (2) is satisfied. Suppose that firms play (*N*, *N*) together with q_M when there is no investigation. Then the ex-ante expected profit from communicating is $\pi((N,N),(N,N)) = \sum_{t=0}^{\infty} \delta^t (\alpha[\pi_M - pF] + (1-\alpha)\pi_M)$. The preceding argument yields that the firms do not report in the no-investigation subgame if

$$\pi((N,N),(N,N)) > \pi_M/\delta.$$

This condition is satisfied if (2) holds.

3.1.3 Communication Stage and Deterrence

If a firm does not communicate, it obtains 0 in the current period. Stationarity implies that the firm does not communicate in all future periods as well. The present value from not communicating is 0.

Firms communicate and then play ((*N*, *N*), (*R*, *R*)) rather than not communicate if $\pi((N,N),(R,R)) > 0$. The AA rules out this communication profile if it makes sure that the inequality does not hold. Straightforward computations show that this is possible for sufficiently high values of α and p. In our set-up, however, α and p are costly. It is clearly cheaper to make sure that (4) is not satisfied: then firms report in the no-investigation subgame and this communication equilibrium does not exist. In the appendix we derive the function $p_{R,R}(\alpha)$ such that if $p \geq p_{R,R}(\alpha)$, (4) is not satisfied and this communication equilibrium does not exist.[24]

[24] A proper notation would be $p_{(N,N),(R,R)}(\alpha)$ where (*N*, *N*) denotes the firms' strategies in the no-investigation and (*R*, *R*) in the investigation subgame. Since for all relevant deterrence constraints firms play (*N*, *N*) in the no-investigation subgame, we suppress (*N*, *N*) and write as a shortcut $p_{R,R}(\alpha)$.

Firms communicate and play ((*N*, *N*), (*R*, *N*)) if $\pi((N,N),(R,N)) > 0$. A similar argument as in the previous paragraph shows that the AA deters efficiently by making sure that (5) does not hold. It does so by setting $p \geq p_{R,N}(\alpha)$, which we also derive in the Appendix.

Firms communicate and play ((*N*, *N*), (*N*, *N*)) if $\pi((N,N),(N,N)) > 0$, which holds if $p < \pi_M/\alpha F$. This condition is satisfied if (2) holds strictly. Thus, if $p \geq p_{N,N}(\alpha)$, the firms report in the investigation subgame, and this communication equilibrium does not exist.

To sum up: the AA achieves complete deterrence if

$$p \geq \max\{p_{N,N}(\alpha), p_{R,R}(\alpha), p_{R,N}(\alpha)\}, \tag{6}$$

where $p_{R,R}(\alpha)$ and $p_{R,N}(\alpha)$ are given by (A1) and (A3).

3.2 Optimal Deterrence

The AA's optimal policy minimizes $C(\alpha p)$ subject to (6). Proposition 1, which we prove in the appendix, characterizes the solution (α^*, p^*):

Proposition 1 *There exist* $0 < \underline{F} < \overline{F}$ *such that:*

(a) if $F < \underline{F}$, $\alpha^* = (\pi_M(2\delta - 1) - (1-\delta)F)/\delta F$ *and* $p^* = 1$;
(b) if $F \geq \underline{F}$, $\alpha^* \in (0,1)$ *and* $p^* \in (0,1)$:
   (i) *for* $\hat{p} \leq 0$, α^* *is defined by* $p_{N,N}(\alpha^*) = p_{R,R}(\alpha^*)$ *with* $\hat{p}$ *given by* (1);
   (ii) for $\hat{p} > 0$ *and* $F \in (\underline{F}, \overline{F}]$, α^* *is defined by* $p_{N,N}(\alpha^*) = p_{R,R}(\alpha^*)$;
   (iii) for $\hat{p} > 0$ *and* $F > \overline{F}$, $p_{N,N}(\alpha^*) = p_{R,N}(\alpha^*)$ *defines* α^*.

To explain this result, first note that the AA must always ensure that the communication equilibrium ((*N*, *N*), (*N*, *N*)) is not played by firms. To see this, suppose on the contrary that $p \geq p_{N,N}(\alpha)$ is not binding. If the AA only has to deter the ((*N*, *N*), (*R*, *R*)) and ((*N*, *N*), (*R*, *N*)) communication equilibria, it can do so by setting, e.g., $\alpha = 1$ and p arbitrarily small, which results in arbitrarily small enforcement cost. With this policy the no-investigation subgame is never reached. In the investigation subgame firms make the competitive profit 0 minus the expected fine. Hence, they do better by not communicating in the first place. Yet, with p small, firms prefer not to report in the investigation subgame: they play ((*N*, *N*), (*N*, *N*)).

Part (a) of Proposition 1 is illustrated in Fig. 1. With small F the legislator endows the AA with limited punishment possibilities. Since fines are low, not reporting in the investigation subgame dominates reporting. Therefore, the AA only has to deter the ((*N*, *N*), (*N*, *N*)) communication equilibrium. To do so, high values of α and p are necessary: $p_{N,N}(\alpha)$ is large. It is cheaper to deter (*N*, *N*) in the investigation subgame than in the no-investigation subgame. In the investigation

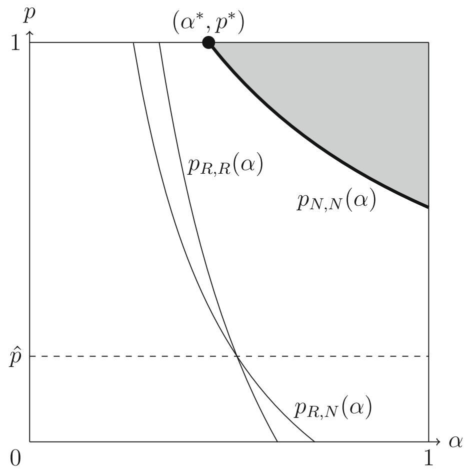


Fig. 1 The deterrence region in case (a)

subgame p has a stronger deterrence effect than α. Therefore, the AA optimally sets $p^* = 1$ and $\alpha^* < 1$.

Increasing F allows the AA to reduce α^*, which leads us to part (b) of Proposition 1. With α small, the no-investigation subgame becomes sufficiently likely, which, in turn, induces firms to collude even though they know that collusion breaks down when there is an investigation. The communication equilibria ((*N*, *N*), (*R*, *R*)) and ((*N*, *N*), (*R*, *N*)) are therefore attractive for firms: they do not report and make profits π_M in the no-investigation subgame. Firm b determines which of the two equilibria is played in the investigation subgame. If firm b reports, it increases the probability of conviction by $(p_2 - p_a)$, while it reduces its expected fine by half. Thus, if $p_2 \leq 2p_a$, firm b reports. Recall that $p_2 \leq 2p_a$ is equivalent to $p \geq \hat{p}$ with $\hat{p}$ given by (1); accordingly, if $p < \hat{p}$, firm b prefers not to report and vice versa for $p \geq \hat{p}$. Since $\hat{p} \leq 1/2$, for large values of p, firm b reports and $p \geq p_{R,R}(\alpha)$ defines the second binding constraint; see Fig. 2. To deter ((*N*, *N*), (*R*, *R*)), the AA has to increase α; the higher fine F enables the AA to decrease p.

Further increasing F allows the AA to lower p until eventually $p < \hat{p}$ (provided $\hat{p} > 0$); see Fig. 3. The lower is p, the higher is the increase in the probability of conviction if b reports. Consequently, for low p firm b will not report in the investigation subgame, and $p \geq p_{R,N}(\alpha)$ is the second binding constraint. If, however, $\hat{p} \leq 0$, $p \geq p_{R,R}(\alpha)$ continues to define the second binding constraint.

The value of $\hat{p}$ thus plays a crucial role as to which constraint is binding. $\hat{p}$ is positive if $\rho_b/(1+\rho_b) > \rho_a$, which holds for ρ_b sufficiently larger than ρ_a. If the two stand-alone probabilities of conviction are of similar size, firm b will report in the investigation subgame like firm a. Call this the case of symmetric firms. Let us now make the firms asymmetric by increasing ρ_b while holding ρ_a constant. This exercise increases b's expected fine when reporting while b's fine when not reporting is unaffected. Thus, for ρ_b large enough firm b prefers not to report, and $p \geq p_{R,N}(\alpha)$ defines the second binding constraint for the AA. $p_{R,N}(\alpha) > p_{R,R}(\alpha)$ means that it is more difficult for the AA to deter cartels with asymmetric firms than

Fig. 2 The deterrence region in case (b) (ii)

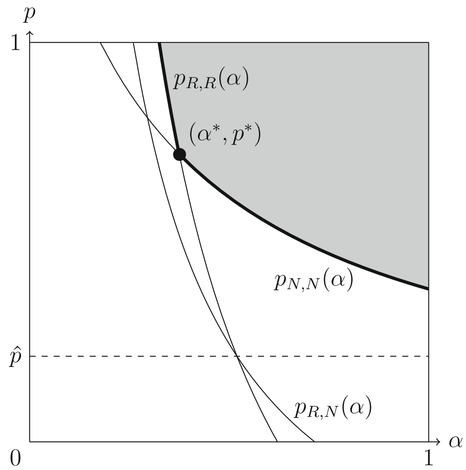
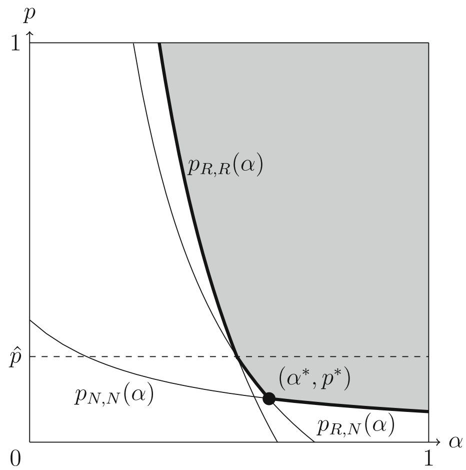


Fig. 3 The deterrence region in case (b) (iii)



cartels with symmetric firms. This case distinguishes our analysis from previous work.

To push this point somewhat further: the enforcement cost $C(\alpha^* p^*)$ decreases monotonically in F. If the legislator wants to minimize enforcement costs in the spirit of Becker (1968), it chooses a high F. If firms are asymmetric, we are in the case of Fig. 3. The AA has to deter the firm with the greater evidence from not reporting in the investigation subgame.

4 Minimum Standards of Evidence

Let us now analyze a minimum standard of evidence $\hat{\rho}$. A firm is granted leniency if and only if the evidence provided is at least $\hat{\rho}$. If $\hat{\rho} < \rho_a$, the minimum standard has no bite, and our preceding analysis applies; $\hat{\rho} > \rho_b$ corresponds to having no leniency at all, which makes deterrence more expensive than with leniency.

This leaves us with the case $\rho_a < \hat{\rho} \leq \rho_b$. Firm a can never avoid the fine so that firm b is the only one to obtain leniency. Accordingly, the communication equilibria ((N, N), (R, R)) and ((N, N), (R, N)) no longer exist. To deter the ((N, N), (N, N)) communication equilibrium, the AA efficiently induces firm b to report in the investigation subgame because it qualifies for leniency. Firm b's payoffs from not reporting and deviating are as described above. Therefore, the AA deters this communication equilibrium by setting $p \geq p_{N,N}(\alpha)$, with $p_{N,N}(\alpha)$ defined by (2).

Next consider ((N, N), (N, R)). In the no-investigation subgame, both firms have the same deviation profit: $2\pi_M$. Firm a, however, has the smaller continuation profit since it does not qualify for leniency. Therefore, the AA optimally induces firm a to deviate. In the no-investigation subgame, firm a's payoff from colluding is $\pi_M + \sum_{t=1}^{\infty} \delta^t (1-\alpha)^{t-1}[-\alpha p_b F + (1-\alpha)\pi_M]$. Firm a deviates if

$$p \geq \frac{\pi_M(2\delta(1-\alpha)-1) - \rho_b \delta \alpha F}{(1-\rho_b)\delta\alpha F} =: p_{N,R}(\alpha).$$

Note that $p_{N,R}(\alpha)$ equals $p_{R,N}(\alpha)$, with ρ_b substituted by ρ_a. Accordingly, $p_{N,R}(\alpha)$ qualitatively has the same properties as $p_{R,N}(\alpha)$. The AA achieves complete deterrence if

$$p \geq \max\{p_{N,N}(\alpha), p_{N,R}(\alpha)\}. \tag{7}$$

As in Proposition 1, for $F \geq \tilde{\underline{F}} > \underline{F}$ the AA's optimal choice (α^*, p^*) satisfies $p_{N,N}(\alpha^*) = p_{N,R}(\alpha^*)$: both restrictions hold with equality. Straightforward computations show that $p \geq p_{N,R}(\alpha)$ is less stringent than both $p \geq p_{R,R}(\alpha)$ and $p \geq p_{R,N}(\alpha)$. Consequently, $C(\alpha^* p^*)$ is lower with the minimum standard than without.

Given that firm a cannot get leniency, firm b enjoys immunity whenever it reports. If firm b reports, the AA gets ample evidence. This, in turn, implies that firm a ex ante faces a high expected fine. By contrast, without minimum standards either both firms pay F / 2 when convicted along ((N, N), (R, R)); or the conviction rate is $p_a < p_b$ because only the low-evidence firm a reports along ((N, N), (R, N)).

To sum up: a minimum standard reduces the AA's enforcement cost if firms are asymmetric and the standard is such that firm b qualifies for leniency while firm a does not. Such a minimum standard provides the high-evidence firm with strong incentives to report: it can be sure to avoid the fine if it chooses to report; the low-evidence firm cannot interfere.

Nevertheless, a minimum standard requires that the evidence is observable and verifiable. Moreover, the AA needs to know how much evidence the firms have, in order to get the standard right. If the AA faces, e.g., uncertainty as to the evidence that is possessed by the firms, it runs the risk of setting the standard too high so that it actually does worse than with no minimum standard at all.[25]

[25] Note that with minimum standards firms have an incentive to keep evidence in order to qualify for leniency (see also Aubert et al. 2006; Agisilaou 2012). Therefore, our assumption that evidence cannot be destroyed is less restrictive with minimum standards than without. This may be seen as a further argument in favor of minimum standards of evidence.

5 Marker System

5.1 Conditional Reporting

Under a marker system the AA informs a firm upon request about its position in line with respect to leniency. If the firm is the first one in line, it gets the marker and leniency if it chooses to report. By contrast, if the firm is second in line and the first firm reports, the AA informs the second firm that the marker and thus leniency are no longer available. If both firms ask for the marker simultaneously, each of them gets it with equal probability.

The marker system creates additional strategies for firms as compared to the no-marker case. They can request the information and then condition their reporting behavior on whether they get the marker (m) or not (n).

For example, a firm can play the following strategy: report if the marker is available, and do not report if the marker is not available: ($R|m$, $N|n$). This strategy weakly dominates unconditional reporting R. If, say, firm a is the first one to report, both strategies yield the same payoff. If, however, firm a is second in line, R yields a payoff that is strictly lower than ($R|m$, $N|n$). Since the firm does not get the marker, it has to pay the fine F if convicted under both strategies. If it does not report, the probability of conviction is lower than if it reports. Likewise, the strategy ($N|m$, $R|n$) is weakly dominated by unconditional not reporting N.

Assuming that firms do not play weakly dominated strategies, we are left with the two strategies ($R|m$, $N|n$) and N for both the investigation and the no-investigation subgame. As a shortcut we will denote ($R|m$, $N|n$) by RN. The two communication equilibria ((N, N), (N, N)) and ((N, N), (RN, RN)) are thus of interest. The analysis of the ((N, N), (N, N)) equilibrium is along the same lines as in the no-marker case. The deviation to consider is RN rather than R. When first in line, the deviation to RN generates the same payoffs as the deviation to R. Thus, the AA deters this equilibrium if $p \geq p_{N,N}(\alpha)$, where $p_{N,N}(\alpha)$ is defined in (2). Let us now turn to the ((N, N), (RN, RN)) communication equilibrium.

5.2 The Firms' Behavior

5.2.1 Investigation Subgame

If both firms play RN, the probability of conviction is with equal probability p_a or p_b, depending on who is first in line. Firm b, for example, gets a payoff $-p_aF/2$; if b deviates to N, its payoff amounts to $-p_aF$. Consequently, (RN, RN) is always an equilibrium in the investigation subgame. If (N, N) is an equilibrium in the investigation subgame, it Pareto-dominates (RN, RN), and the firms play (N, N). This is the case if $p < p_{N,N}(\alpha)$.

5.2.2 No-Investigation Subgame

Suppose firms play (*N*, *N*) together with q_M when there is no investigation. Then firm *a*'s ex-ante expected profit from communicating is

$$\pi_a((N,N),(RN,RN)) = \sum_{t=0}^{\infty} \delta^t (1-\alpha)^t(-\alpha p_b F/2 + (1-\alpha)\pi_M).$$

Note that *b*'s profit is higher than *a*'s because $p_a < p_b$. Now consider the no-investigation subgame. If firm *a* does not report, it makes profit $\pi_M + \delta\pi_a((N,N),(RN,RN))$. If it reports and chooses q_D, its profit is $2\pi_M$. Accordingly, (*N*, *N*) is an equilibrium of the no-investigation subgame if

$$\pi_a((N,N),(RN,RN)) \geq \pi_M/\delta. \tag{8}$$

If (8) is not satisfied, at least one firm reports and picks q_C.

5.2.3 Communication Stage and Deterrence

To deter ((*N*, *N*), (*N*, *N*)) the AA sets $p \geq p_{N,N}(\alpha)$. To deter ((*N*, *N*), (*RN*, *RN*)) the AA has to make sure that (8) is not satisfied: then firms report in the no-investigation subgame and this communication equilibrium does not exist. This is the case if

$$p \geq (2\pi_M(2\delta(1-\alpha)-1) - \rho_b\delta F)/(1-\rho_b)\delta\alpha F =: p_{RN,RN}(\alpha).$$

To sum up the marker case: the AA achieves complete deterrence if

$$p \geq \max\{p_{N,N}(\alpha), p_{RN,RN}(\alpha)\}. \tag{9}$$

5.3 Optimal Deterrence

The optimal policy of the AA minimizes $C(\alpha p)$ subject to (9). Proposition 2 characterizes the solution (α^*, p^*):[26]

Proposition 2 *There exists* $\underline{F} > 0$ *such that:*

(a) *if* $F < \underline{F}$, $\alpha^* = (\pi_M(2\delta - 1) - (1-\delta)F)/\delta F$ *and* $p^* = 1$;
(b) *if* $F \geq \underline{F}$, $\alpha^* \in (0,1)$ *and* $p^* \in (0,1)$ *with* α^* *defined by* $p_{N,N}(\alpha^*) = p_{RN,RN}(\alpha^*)$.

In Fig. 4, the bold dashed line depicts the set of the optimal deterrence combinations (α^*, p^*); for the sake of clarity we have not included the $p_{N,N}$-curve. Let us now compare the enforcement cost under the no-marker and the marker system. If $F < \underline{F}$ so that $p^* = 1$, the AA deters with the same (α^*, p^*) combinations under both systems. Accordingly, the AA's deterrence costs are the same.

[26] Since the proof of Proposition 2 is similar to the proof of Proposition 1, we skip it.

If $F > \underline{F}$, $p^* = p_{RN,RN}(\alpha^*) < 1$. Furthermore, $p_{R,R}(\alpha) < p_{RN,RN}(\alpha)$ whenever the latter is less than 1. This means that for the cases (b) (i) and (ii) of Proposition 1 deterrence is more expensive in the marker scenario than in the no-marker scenario.

In the no-marker scenario the AA deters the ((*N*, *N*), (*R*, *R*)) equilibrium where in the investigation subgame the probability of conviction is p_2. In the marker scenario the AA deters the ((*N*, *N*), (*RN*, *RN*)) equilibrium where in the investigation subgame the expected probability of conviction is $(p_a + p_b)/2 < p_2$. The expected payoff from colluding is thus higher in the marker than in the no-marker case and a higher p is necessary to deter firms from communicating.

This leaves us with the case (b) (iii) of Proposition 1. Here the outcome depends on whether $p_{RN,RN}(\alpha)$ and $p_{R,N}(\alpha)$ intersect at some $\bar{\alpha} > 0$. If they do not intersect, a marker is detrimental. If they intersect, then for $\alpha^* > \bar{\alpha}$, $p_{RN,RN}(\alpha^*) < p_{R,N}(\alpha^*)$, and a marker is beneficial; see Fig. 4.

In the ((*N*, *N*), (*R*, *N*)) equilibrium, the probability of conviction is p_a in the investigation subgame. In the ((*N*, *N*), (*RN*, *RN*)) equilibrium, the probability of conviction is $(p_a + p_b)/2 > p_a$ in the investigation subgame. The expected payoff is thus higher in the no-marker case than in the marker case and a higher p is necessary to deter firms from colluding.

To sum up: deterrence is cheaper under the marker than the no-marker system if and only if $\hat{p}$ and F are sufficiently large. $\hat{p}$ is large when firms are sufficiently asymmetric: $\rho_b \gg \rho_a$. In the no-marker case, only firm *a* reports whereas in the marker case both firms report with equal probability. Colluding is thus less attractive in the marker case, and the AA has an easier time to deter firms from communicating. In all other cases the marker system does worse because the second-in-line firm withholds its evidence.

Finally, let us look at a minimum standard of evidence $\hat{\rho}$ under the marker system. If $\hat{\rho} < \rho_a$, the minimum standard has no bite, and the preceding analysis applies; $\hat{\rho} > \rho_b$ corresponds to no leniency. If $\hat{\rho} \in (\rho_a, \rho_b]$, firm *b* is the only that can obtain leniency; firm *a* cannot interfere. Firm *b* does not need a marker to forestall being queue-jumped by firm *a*; therefore, our analysis from Sect. 4 applies.

6 Ringleader Discrimination

Finally, let us look at the case where the high-evidence firm *b* is the cartel's ringleader. Moreover, suppose that, as in the US, the ringleader does not qualify for leniency. Accordingly, the communication strategies ((*N*, *N*), (*R*, *R*)) and ((*N*, *N*), (*N*, *R*)) no longer exist. Ringleader discrimination is orthogonal to minimum-evidence standards: here the high-evidence firm is denied leniency; there the low-evidence firm is denied leniency.[27]

[27] One of the reasons that the Department of Justice denies leniency to ringleaders is its concern about recidivism. There is evidence that recidivism is a problem in the EU; see Zhou (2015) and Marvão (2016). A proper analysis of recidivism requires a dynamic model that includes the cartel formation process. We follow here the approach taken in the literature and employ a static framework.

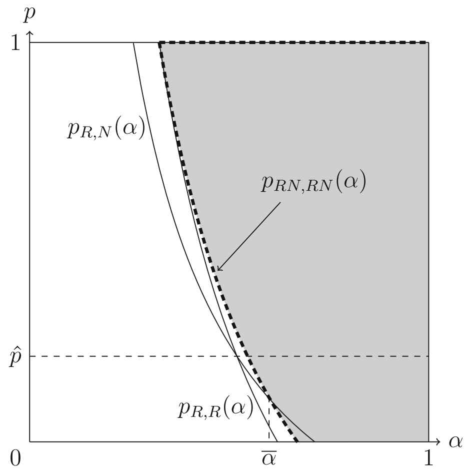


Fig. 4 The deterrence region in the marker scenario

To deter ((*N*, *N*), (*N*, *N*)), the AA efficiently induces firm *a* to report in the investigation subgame because it qualifies for leniency. Firm *a*'s payoffs from not reporting and deviating are as described in Sect. 3. Therefore, the AA deters this communication equilibrium by setting $p \geq p_{N,N}(\alpha)$, with $p_{N,N}(\alpha)$ defined by (2).

Next consider ((*N*, *N*), (*R*, *N*)). In the no-investigation subgame, both firms have the same deviation profit $2\pi_M$. Firm *b*, however, has the smaller continuation profit since it does not qualify for leniency. Therefore, the AA optimally induces firm *b* to deviate. In the no-investigation subgame, firm *b*'s payoff from colluding is $\pi_M + \sum_{t=1}^{\infty} \delta^t (1-\alpha)^{t-1}[-\alpha p_a F + (1-\alpha)\pi_M]$. The AA induces firm *b* to deviate by setting $p \geq p_{R,N}(\alpha)$ with $p_{R,N}(\alpha)$ given by (A3). Thus, the AA achieves complete deterrence if

$$p \geq \max\{p_{N,N}(\alpha), p_{R,N}(\alpha)\}. \tag{10}$$

(10) differs from (6) in that the constraint $p \geq p_{R,R}(\alpha)$ is missing. Since $p_{R,N}(\alpha) < p_{R,R}(\alpha)$, the AA deters at lower cost with ringleader discrimination rather than without whenever $p_{R,R}(\alpha)$ is binding in Proposition 1; see Figs. 1, 2 and 3. In all of the other cases of Proposition 1, the deterrence cost with and without ringleader discrimination is the same.

Denying leniency to the ringleader is thus an especially good idea if the two firms are rather symmetric in terms of evidence so that $\hat{p} < 0$. Without leniency it is less attractive for the ringleader to set up the cartel in the first place. If, by contrast, firms are sufficiently asymmetric, the ringleader intends not to report anyway, so denying leniency has no effect on deterrence.[28] Since $p_{R,N} > p_{N,R}$, the AA fares better with minimum-evidence standards than with ringleader discrimination: Obtaining plenty of evidence from the ringleader provides more deterrence than getting little evidence from the other cartel members.

[28] Herre et al. (2012) also derive the deterring effect of ringleader discrimination when firms are sufficiently symmetric.

7 Conclusions

Nearly the entire literature on leniency examines the case where all members of a cartel have perfect evidence of the illegal conduct. This assumption implies that if one member reports, no matter which one, the cartel is convicted for sure. If all firms have perfect evidence, the distribution of evidence is symmetric, and either all firms apply for leniency or no firm applies; this is a feature that is typically not observed in practice. Moreover, within such a framework one cannot meaningfully analyze minimum-evidence standards, marker systems, and ringleader discrimination, which are three typical features of actual leniency programs.

Our analysis highlights the importance of appropriate minimum standards of evidence. In a setting with high fines and asymmetric evidence, deterrence is costly since the high-evidence firm has no incentive to report. The AA can lower its enforcement cost by choosing a minimum standard of evidence such that the high-evidence firm qualifies for leniency while the low-evidence firm does not. Such a minimum standard restores the incentives to report for the high-evidence firm. It can be sure to avoid the fine, because the low-evidence firm is, in fact, denied leniency.

The challenge for the AA in practice is to get the minimum standard right. If the standard is so low that both firms fulfill the conditions for leniency, it remains ineffective. By contrast, a standard so high that no firm reaches the threshold renders the whole leniency program ineffective.

When minimum-evidence standards are feasible, there is no need for a marker system: the high-evidence firm is protected from being queue-jumped anyway. When they are not feasible, a marker system may make sense.[29]

With a marker, firms never reveal the entire evidence: the AA gets only one report: with equal probability the high- or the low-evidence report. This increases the deterrence cost compared to the no-marker set-up if both firms report without the marker; this is the case if the firms have similar evidence. If, however, without the marker only the low-evidence firm reports, the marker lowers the AA's deterrence cost; this happens if firms possess sufficiently different evidence. Our model, therefore, suggests the use of a marker system if and only if the AA expects asymmetric distributions of the evidence that is in the possession of the colluding firms.

We have provided the first theoretical analysis of marker systems. But we have not dealt with all the arguments made in favor of a marker system. For example, this instrument may create legal certainty and transparency, or the possibility to secure a marker with preliminary evidence may induce a race to the courthouse. The analysis of these arguments, perhaps with our framework as a cornerstone, is left for future research.

As to policy implications: minimum standards of evidence, if applicable, seem a good idea because they create proper incentives for the high-evidence firm to report. Ringleader discrimination may reduce deterrence costs. But minimum-evidence standards allow for better deterrence than does ringleader discrimination.

[29] If the AA is constrained to use a uniform standard across all industries, for certain industries the standard may be so low that all firms qualify for leniency. Then the use of a marker may also make sense.

Combining minimum-evidence standards with ringleader discrimination is not advisable because this may effectively result in no leniency at all. Under a marker system the AA never gets the entire evidence. This may increase or reduce deterrence costs depending on the distribution of the evidence. A marker system is redundant when appropriate minimum-evidence standards apply.

Appendix

Properties of $p_{R,R}$ *and* $p_{R,N}$

Using $p_2(p) = 1 - (1-p)(1-\rho_a)(1-\rho_b)$, write (4) as

$$p \geq \frac{2\pi_M(2\delta(1-\alpha)-1) - (\rho_a + \rho_b - \rho_a\rho_b)\delta\alpha F}{(1-\rho_a)(1-\rho_b)\delta\alpha F} =: p_{R,R}(\alpha). \tag{A1}$$

$p_{R,R}(\alpha)$ is continuous and decreasing in α. Rewriting (A1) yields

$$\delta\alpha F[p(1-\rho_a)(1-\rho_b) + (\rho_a + \rho_b - \rho_a\rho_b)] \geq 2\pi_M(2\delta(1-\alpha)-1). \tag{A2}$$

The LHS of (A2) equals 0 for $\alpha = 0$ and increases in α. The RHS is positive for $\alpha = 0$ and negative for $\alpha = 1$. Hence, (4) holds for values of α below some threshold and does not hold for values of α above this threshold.

Using $p_a(p) = 1 - (1-p)(1-\rho_a)$, write (5) as

$$p \geq \frac{\pi_M(2\delta(1-\alpha)-1) - \rho_a\delta\alpha F}{(1-\rho_a)\delta\alpha F} =: p_{R,N}(\alpha). \tag{A3}$$

$p_{R,N}(\alpha)$ is continuous and decreases in α for $p_{R,N}(\alpha) \geq 0$. Rewrite (A3) as

$$\delta\alpha F[p(1-\rho_a) + \rho_a] \geq \pi_M(2\delta(1-\alpha)-1). \tag{A4}$$

(A4) holds for values of α close to 1 and is violated for values of α small. Hence, (5) holds for values of α below some threshold and does not hold for values of α above this threshold.

Proof of Proposition 1 Rewrite the AA's minimization problem as $(\alpha^*, p^*) = \arg\min_{\alpha,p} C(\alpha p)$, s.t.

$$p \geq p_{N,N}(\alpha) \quad \text{or} \quad \alpha \geq \alpha_{\mathrm{N,N}}(\mathrm{p}), \tag{A5}$$

$$p \geq p_{R,R}(\alpha) \quad \text{or} \quad \alpha \geq \alpha_{\mathrm{R,R}}(\mathrm{p}), \tag{A6}$$

$$\begin{aligned} p \geq p_{R,N}(\alpha) \quad &\text{or} \quad \alpha \geq \alpha_{\mathrm{R,N}}(\mathrm{p}), \\ 0 \leq \alpha \leq 1, \quad &\text{and} \quad 0 \leq \mathrm{p} \leq 1, \end{aligned} \tag{A7}$$

where

$$\alpha_{N,N}(p) := \frac{\pi_M(2\delta - 1)}{\delta pF} - \frac{1-\delta}{\delta},$$
$$\alpha_{R,R}(p) := \frac{\pi_M(2\delta - 1)}{\delta([p(1-\rho_a)(1-\rho_b) + \rho_a + \rho_b - \rho_a\rho_b]F/2 + 2\pi_M)},$$
$$\alpha_{R,N}(p) := \frac{\pi_M(2\delta - 1)}{\delta([p(1-\rho_a) + \rho_a]F + 2\pi_M)}.$$

First, note that $p^* \geq p_{N,N}(1) > 0$ because $p_{N,N}$ is decreasing in α; $\alpha^* \geq \alpha_{R,R}(1) > 0$ because $\alpha_{R,R}$ is decreasing in p.

Second, note that (A5) is always binding. Consider the relaxed problem of minimizing the cost subject to (A6) and (A7). Then $\alpha = 1$ and $p = \epsilon$, with ϵ positive and small, satisfy (A6) and (A7). The no-investigation subgame is never reached. In the investigation subgame firms always report. They earn 0 and pay the fine with positive probability ϵ. They do better by not reporting in the first place. This policy gives rise to the cost $C(1\epsilon) = \epsilon L/\gamma$. Obviously, $C(\alpha^* p^*) \leq \epsilon L/\gamma$. Now suppose (A5) is slack: using $\alpha^* \geq \alpha_{R,R}(1) > 0$ we have

$$\alpha^* p^* \geq \frac{\pi_M(2\delta - 1)}{F((1-\delta)/\alpha^* + \delta)} > \frac{\pi_M(2\delta - 1)}{F\big((1-\delta)/\alpha_{R,R}(1) + \delta\big)} > 0,$$

which means $\alpha^* p^* > \epsilon L/\gamma$ for ϵ sufficiently small, which contradicts $C(\alpha^* p^*) \leq \epsilon L/\gamma$.

Third, note that $(\alpha^*, p^*) = (\hat{\alpha}, 1)$ with $\hat{\alpha} = (\pi_M(2\delta - 1) - (1-\delta)F))/\delta F$ if and only if (A6) and (A7) are not binding. Consider the Lagrangian of minimizing C subject to (A5). Solving the first-order conditions with respect to α and p yields $1 - \delta = 0$: a contradiction. Hence, either $\alpha^* \in \{0, 1\}$ or $p^* \in \{0, 1\}$. We know already that $\alpha^* > 0$ and $p^* > 0$. Consequently, either α^* or p^* equals 1. Suppose $\alpha^* = 1$ and $p^* = (\pi_M(2\delta - 1))/F$: straightforward computations show that decreasing α by $d\alpha$ and increasing p by $d\alpha(1-\epsilon)p/\alpha$ reduces the cost without violating (A5). Therefore, $p^* = 1$ and $\alpha^* = \hat{\alpha}$.

Fourth, note that $(\alpha^*, p^*) = (\hat{\alpha}, 1)$, or equivalently (A6) and (A7) do not hold, if and only if $F \leq \underline{F}$. Since $\alpha_{R,R}(1) \geq \alpha_{R,N}(1)$, (A6) and (A7) are not binding at $(1, \hat{\alpha})$ if

$$2(\pi_M(2\delta - 1))2\pi_M \geq (\pi_M(2\delta - 1))F + (1-\delta)F^2.$$

The LHS is positive. The RHS is 0 for $F = 0$ and strictly increasing in F. Thus, there exists a unique $\underline{F}$ with the desired properties.

Fifth, note that for $F > \underline{F}$, $\alpha^* \in (0, 1)$ and $p^* \in (0, 1)$. We know from the previous step that for $F > \underline{F}$ either (A6) or (A7) is binding. Since $\alpha_{R,N}(p) \in (0, 1)$ and $\alpha_{R,R}(p) \in (0, 1)$ for all $p \in [0, 1]$, $\alpha^* \in (0, 1)$. From our first step we know that $p^* > 0$. From our third step we know that in this case $\alpha^* = \hat{\alpha}$. In this case, however, if $p^* = 1$, step 4 implies $F < \underline{F}$: a contradiction.

Sixth, straightforward computations show that $\alpha_{R,R}(p) \geq \alpha_{R,N}(p)$ or equivalently $p_{R,R}(\alpha) \geq p_{R,N}(\alpha)$ if and only if $p \geq \hat{p}$ with $\hat{p}$ defined in (1).

Seventh, if $F > \underline{F}$ and $\hat{p} \leq 0$, $p^* = p_{N,N}(\alpha^*) = p_{R,R}(\alpha^*)$. From step 2 we know that $p^* = p_{N,N}(\alpha^*)$ and steps 3 and 4 imply $p^* = \max\{p_{R,R}(\alpha^*), p_{R,N}(\alpha^*)\}$. By step 6,

$p^* = p_{R,R}(\alpha^*)$ if and only if $p^* \geq \hat{p}$. Since $\hat{p} \leq 0$ and $p^* > 0$, this is always true.

Eighth, if $\hat{p} > 0$, $p^* = p_{N,N}(\alpha^*) = p_{R,R}(\alpha^*)$ for $F \in [\underline{F}, \overline{F}]$ and $p^* = p_{N,N}(\alpha^*) = p_{R,N}(\alpha^*)$ for $F \geq \overline{F}$. For $F > \underline{F}$ we have $p^* = p_{N,N}(\alpha^*) = \max\{p_{R,R}(\alpha^*), p_{R,N}(\alpha^*)\}$. As can be easily shown, the three functions have unique intersections. Therefore, p^* is unique and, moreover, continuous and decreasing in F. For $F = \underline{F}$, $p^* = 1$. If we increase F slightly, by continuity p^* falls slightly. Hence, $p^* > 1/2 \geq \hat{p}$ and by step 6 $p_{R,R} > p_{R,N}$. For F sufficiently large, p^* is arbitrarily small, meaning $p^* < \hat{p}$. Consequently, $p_{N,R} > p_{R,R}$. □

The Collude and Reveal Strategy:

Let us briefly discuss the collude and reveal strategy: Collude in each period; if there is no investigation, play q_M and do not reveal; if there is an investigation, play q_C and reveal; deviations are punished by grim-trigger strategies. Denote this collude and reveal strategy by $((N,N),(\bar{R},\bar{R}))$.

This strategy yields each period profit π_M with probability $(1-\alpha)$ and $-p_2F/2$ with probability α. Hence, $\pi((N,N),(\bar{R},\bar{R})) = [(1-\alpha)\pi_M - \alpha p_2 F/2]/(1-\delta)$. In the no-investigation subgame playing N and q_M yields profit $\pi_M + \delta\pi((N,N),(\bar{R},\bar{R}))$. Deviating to R and q_D yields profit $2\pi_M$. Firms play $((N,N),\ (\bar{R},\bar{R}))$ if $2\pi_M < \pi_M + \delta\pi((N,N),(\bar{R},\bar{R}))$.

The AA deters this collude and reveal strategy if

$$\alpha \geq \alpha_{\bar{R},\bar{R}}(p) := \frac{\pi_M(2\delta - 1)}{\delta([p(1-\rho_a)(1-\rho_b) + \rho_a + \rho_b - \rho_a\rho_b]F/2 + \pi_M)}.$$

Note that $\alpha_{\bar{R},\bar{R}} > \alpha_{R,R}$, which means that $\alpha_{R,R}$ is no longer binding: deterrence is more expensive if firms use this strategy than if they do not. From $\alpha_{\bar{R},\bar{R}} = \alpha_{R,N}$ we compute the intersection $\bar{p} = \hat{p} - 2\pi_M/(F(1-\rho_a)(1+\rho_b))$. $\bar{p}$ is positive for $\hat{p} > 0$ and F sufficiently large. If $\bar{p} > 0$, $\alpha_{R,N}$ binds for $p < \bar{p}$.

References

Agisilaou, P., (2012). Keep to sustain or keep to exploit? Why firms keep hard evidence, Munich CCP Working Paper 12-5, http://mpra.ub.uni-muenchen.de/39109/.

Aubert, C., Kovacic, W., & Rey, P. (2006). The impact of leniency programs on cartels. *International Journal of Industrial Organization*, *6*, 1241–1266.

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

Bigoni, M., Fridolfsson, S., Le Coq, C., & Spagnolo, G. (2012). Fines, leniency, and rewards in antitrust. *Rand Journal of Economics*, *43*, 368–390.

Bigoni, M., Fridolfsson, S., Le Coq, C., & Spagnolo, G. (2015). Trust, leniency, and deterrence. *Journal of Law, Economics, and Organization*, *31*, 663–689.

Block, M., & Sidak, J. (1980). The cost of antitrust deterrence: Why not Hang a price fixer now and then? *Georgetown Law Journal*, *68*, 1131–1139.

Brenner, S. (2009). An empirical study of the European corporate leniency program. *International Journal of Industrial Organization*, *27*, 639–645.

Charistos, K., & Constantatos, C. (2016). On leniency and markers in antitrust: How many informants are enough?, http://aphrodite.uom.gr/econwp/pdf/dp022016.pdf.

Chen, C., Ghosh, S., & Ross, T. (2015). Denying leniency to cartel instigators: Costs and benefits. *International Journal of Industrial Organization*, *41*, 19–29.
Cyrenne, P. (1999). On antitrust enforcement and the deterrence of collusive behaviour. *Review of Industrial Organization*, *14*, 257–272.
European Commission. (2006). Commission notice on immunity from fines and reduction of fines in cartel cases. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52002XC0219%2802%29.
Feess, E., & Walzl, M. (2010). Evidence dependence of fine reductions in corporate leniency programs. *Journal of Institutional and Theoretical Economics*, *166*, 573–590.
Harrington, J. (2008). Optimal corporate leniency programs. *Journal of Industrial Economics*, *56*, 215–246.
Harrington, J. (2013). Corporate leniency programs when firms have private information: The push of prosecution and the pull of pre-emption. *Journal of Industrial Economics*, *61*, 1–27.
Harrington, J., & Chang, M. (2009). Modeling the birth and death of cartels with an application to evaluating competition policy. *Journal of the European Economic Association*, *7*, 1400–1435.
Harrington, J., & Chang, M. (2015). When can we expect a corporate leniency program to result in fewer cartels? *Journal of Law and Economics*, *58*, 417–449.
Herre, J., Mimra, W., & Rasch, A. (2012). Excluding Ringleaders from Leniency Programs, http://ssrn.com/abstract=1342549.
Kobayashi, B. (1992). Deterrence with multiple defendants: An explanation for "Unfair" plea bargains. *Rand Journal of Economics*, *23*, 507–517.
Marvão, C. (2016). The EU leniency programme and recidivism. *Review of Industrial Organization*, *49*, 1–27.
Miller, N. (2009). Strategic leniency and cartel enforcement. *American Economic Review*, *99*, 750–768.
Motta, M., & Polo, M. (2003). Leniency programs and cartel prosecution. *International Journal of Industrial Organization*, *21*, 347–379.
Sauvagnat, J. (2015). Prosecution and leniency programs: The role of bluffing in opening investigations. *Journal of Industrial Economics*, *63*, 313–338.
Silbye, F., (2010). Asymmetric evidence and optimal leniency programs, Ph.D. Thesis, University of Copenhagen, Copenhagen (pp. 41–61).
Spagnolo, G., (2003). Divide et Impera: Optimal deterrence mechanisms against cartels and organized crime, http://sites.google.com/site/giancarlospagnoloshomepage/research-2/working-papers.
Spagnolo, G., & Marvão, C. (2016). Cartels and leniency: Taking stock of what we learnt, http://swopec.hhs.se/hasite/papers/hasite0039.pdf.
US Department of Justice. (2008). Frequently asked questions regarding the antitrust division's leniency program and model leniency letters, http://www.justice.gov/atr/frequently-asked-questions-regarding-antitrust-divisions-leniency-program.
Zhou, J. (2015). The rise and fall of cartels with multi-market colluders. *Review of Industrial Organization*, *48*, 381–403.