

On the Design of Leniency Programs

Zhijun Chen* Patrick Rey†

2012-01-04

Abstract

Leniency programs contribute to destabilizing collusion, however offering generous leniency might risk being exploited to perverse effects. This paper develops a simple model to resolve this trade-off in the design of leniency programs, which allows us to relate the optimal leniency policies to the frequency and effectiveness of investigations. We show that it is always desirable to offer some leniency before an investigation is launched, it is moreover optimal to keep offering leniency once an investigation is underway if investigations are less likely to succeed in uncovering cartels absent self-reporting. Our analysis also confirms the usefulness of restricting leniency to the first informant only; in contrast, it does not appear to support prohibiting leniency for repeated offenders.

JEL Classification: L41, K21

Keywords: Cartel, Leniency Program, Antitrust

*University of Auckland; e-mail: chenzj1219@gmail.com.

†Toulouse School of Economics (IDEI, GREMAQ and IUF); E-mail: prey@cict.fr

1 Introduction

Cartel detection and deterrence are among antitrust authorities' highest priorities. One of the most important developments in this area of antitrust policy is the introduction of leniency programs. First adopted in 1978 in the U.S., these programs allow corporations or individuals involved in illegal cartel activity to receive amnesty if they come forward and denounce the cartel. In 1993, the US amnesty program was revised to give firms more opportunities and higher incentives to cooperate with the Antitrust Division: the "first informant" rule now guarantees amnesty to the first reporting firm (and only to the first one), while the "post investigation amnesty" rule allows the first informant to remain eligible even once an investigation is already underway. This revised leniency program has been the most effective antitrust enforcement tool and it has helped the Antitrust Division to crack dozens of international cartels, convict U.S. and foreign executives, and enforce record-breaking corporate fines. This success has encouraged many other countries or jurisdictions to adopt their own leniency programs.¹

In spite of a great success in practice,² many open questions remain and, while the positive analysis has already made some progress, much remains to be done to study the optimal design of leniency programs. While leniency programs bring stronger incentives for cartel participants to break the collusive agreements and moreover denounce their cartel, it also broadens the range of collusive strategies; in particular cartels may abuse these programs if they are too generous. Such strategic exploitation has been discussed by economists such as Massimo Motta and Michele Polo, who argue for example that "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",³ as well as lawyers and practitioners, such as Wouter Wils, who notes: "Successful cartels tend to be sophisticated organizations, capable of learning. It is thus safe to assume that cartel participants will try to adapt their organization to leniency policies, not only so as to minimize the destabilizing effect, but also, where possible, to exploit leniency policies to facilitate the creation and maintenance of cartels. This raises the question whether there could be features of leniency programs that risk being exploited to perverse effect."⁴ For example, "[i]n situations

¹A leniency program has for example been adopted by the EU Commission in 1996, and revised in 2002; many countries have also adopted leniency programs. South Korea recently adopted a leniency program that can furthermore grant monetary rewards to individual informants.

²See Hammond (2005).

³See Motta and Polo (2003), pp. 349.

⁴See Wils (2008), pp.137.

where the same companies participate in a number of cartels in different markets, or repeatedly form cartels over time, one could imagine a system in which cartel participants take turns to apply for leniency, every time one of the cartels is (about to be) detected by the competition authority. In jurisdictions where cartels are not only punished by fines on companies, but also by imprisonment of individuals, such a system is most unlikely to be attractive, but in jurisdictions without such individual penalties it may work."⁵ This concern is further validated by recent evidence from laboratory experiments; for instance, Hinloopen and Soetevent (2008) show that collusive strategies such as "collude and report systematically" can be adopted when the leniency programs are too generous; reporting the cartel then becomes part of the collusive agreement: participants collude and apply for leniency in every period to benefit from the reduction of fines.

This risk of exploitation and perverse effects in turn puts into question whether antitrust authorities should grant amnesty or even rewards to eligible cartel members, and raises an issue about how much leniency should be offered so as to balance optimally its opposite effects on collusion. This paper establishes a simple model to analyze this trade-off for the optimal design of leniency program; it departs from the existing literature in several aspects. First of all, while most of the existing literature (with few exceptions, such as Motta and Polo (2003)) adopts an optimistic view which excludes the possibility that generous leniency might be exploited by cartels, we adopt a more cautious attitude that takes into account this possibility. Second, while most of the existing literature takes the leniency rate as given and studies the impact of leniency programs on cartel formation and sustainability, our paper looks for the optimal amnesty rates, both before and after an investigation is launched, taking into consideration several features of antitrust enforcement such as the probability that a cartel would be investigated and successfully prosecuted in the absence of reporting.

To study the effectiveness of leniency programs, we consider an environment where industries differ in their benefits from collusion. Firms will form a cartel if the benefits from collusion exceed the expected penalty under antitrust enforcement; deterring collusion "as much as possible" then amounts to maximizing the threshold on collusive benefits below which collusion is deterred. The optimal leniency programs balances two effects: (i) destabilizing usual collusion (of the form "collude and never report") by encouraging firms to deviate and denounce the cartel; and (ii) discouraging firms from exploiting the leniency program through "collude and report" strategies. Our simple framework allows us to relate the optimal leniency rates (the "carrot"), which is the solution of the trade-off just mentioned, to the effectiveness of random investigations

⁵See Wils (2008), pp. 137.

(the "stick"). We show that it is always desirable to offer some leniency, at least in the absence of any ongoing investigation; whether amnesty remains desirable once an investigation is underway depends however on the frequency of investigations as well as on the likelihood of success for these investigations: Whenever cartels are more patient and thus collusion is more robust, it is optimal to grant leniency post-investigation when investigations are unlikely to succeed; whenever cartels are less patient, however, granting amnesty post-investigation is desirable only if investigations are sufficiently frequent and unlikely to be successful. The analysis also shows that it is optimal to offer less leniency once an investigation is already underway, as it is the case with most leniency programs,⁶ when investigations are more likely to succeed once they are launched; when instead investigations are less likely to succeed, it can however be desirable to offer more amnesty once an investigation is underway, in order to make these investigations more effective.

The revision of the US amnesty program in 1993 introduced two main innovations. While the effectiveness of restricting leniency to the first informant is widely recognized in the existing literature, fewer analyses study the rule that allows the first informant to remain eligible even after an investigation is underway. This paper provides a formal analysis on post-investigation leniency which validates the desirability of offering post-investigation leniency when investigations are otherwise quite unlikely to succeed. Our simple model also allows us to compare the different variants of policies, and yields several policy implications. The most surprising is that withdrawing leniency for repeated offenders generates a countervailing effect which can make the leniency policy completely ineffective; this calls for a cautious use of heavy sticks.

This paper builds on the growing literature on leniency programs. In particular it is closely related to Motta and Polo (2003), who analyze the impact of leniency on collusion in a framework where the antitrust agency can launch random investigations that sometimes lead to successful prosecution. They take the leniency rates as given and study the deterrence as well as desistance effects of the amnesty program. Taking into account the possibility of that cartels may strategically exploit leniency policies, they find that the introduction of a leniency program generates two conflicting effects, as mentioned above, making it difficult to conclude that a leniency program unambiguously increase welfare. In contrast, our simple model allows us to characterize the optimal degree of leniency and show that both pre-and post-investigation leniency can help prevent the formation of some cartels and unambiguously increase welfare. They also investi-

⁶For example, the EU program grants a 75%–100% reduction of fines before investigation, but only a 50%–75% reduction once an investigation is already underway.

gate the most effective way to allocate antitrust resources between preliminary investigation and prosecution; in contrast, we take here the likelihood of investigations and successful prosecution as given, and study the optimal design of leniency policy; our framework can be used to investigate the optimal allocation of enforcement resources under a budget constraint, and thus to determine the optimal investigation rate and the relevant effectiveness of investigation, but we leave this work in the future research.

Spagnolo (2004) also examines the effect of leniency program on cartels and shows that the antitrust authority should not impose a fine on firms that deviate from a cartel agreement; it should instead reward only the first informant; Spagnolo also notes that, while leniency can contribute to destabilize cartels, it can also have a negative impact, by reducing the expectation fines; Spagnolo concludes that it is therefore optimal to grant a reward (to the first informant) equal to the sum of fines levied from other cartel firms. Our analysis builds on his insights and takes into account similar adverse effects when studying the optimal leniency rates. Aubert, Rey and Kovacic (2006) compare the impact of reduced fines and positive rewards and argue that rewarding individuals can deter collusion in a more effective way. They also discuss possible adverse effects of whistleblowing programs on firms' behavior and incentives to innovate and cooperate. Harrington (2008) characterizes the leniency program in a framework that allows the probability of discovery and successful prosecution to change over time. He points out that offering leniency can trigger a "Race-to-the-courthouse" when detection becomes likely, which in turn increases the expected penalties from engaging in cartel activity; he also shows that it is optimal to restrict eligibility to the first informant and also often optimal (assuming away positive rewards) to grant full leniency to that first informant. Harrington and Chang (2009) studies the impact of leniency programs on cartel desistance as well as cartel deterrence. He develops a framework where industries differ in the benefits from deviation (for simplicity, we suppose instead that firms differ in their benefits from collusion as well as from deviation) and in which exposed cartels disappear until they have a new opportunity to form (a random event). This allows for a nice characterization of not only the equilibrium number of cartels, but also the distribution of cartel duration.

The rest of the paper is organized as follows. Section 2 sets up the model. Section 3 studies the basic trade-off between the two above-mentioned forces in a simple framework and discusses some policy implications. Section 4 extends the analysis to allow for both pre- and post-investigation leniency. We finally conclude in section 5.

2 The model

2.1 The collusion game

In each industry, two identical firms play an infinitely repeated "competition vs collusion" game. More precisely, the two firms use the same discount rate $\delta \in (0, 1)$, maximize the expected discounted sum of their profits and, in each period, choose whether to "collude" or "compete à la Bertrand"; the gross profit of a firm is:

- 0 if both firms compete,
- B if both firms collude,
- $2B$ for a firm that deviates from the collusive market scheme while the other colludes, in which case the other firm gets 0.

This stage game can for example be interpreted as representing a standard Bertrand duopoly, in which two firms produce the same good with the same constant unit cost c , and compete for a demand $D(p)$: static price competition then indeed yields zero profit, whereas the maximal benefit from collusion corresponds to half of the monopoly profits ($B = \pi^M/2 = \max_p (p - c) D(p)/2$); and deviating from such collusion yields a short-term gain that can be as large as the entire monopoly profit, i.e., twice as large as the benefit from collusion.⁷

Firms can try to sustain repeated collusion by returning to competition (which is both the static Nash equilibrium and the minimax) in case a firm deviates from the collusive outcome. In the absence of any antitrust policy, collusion is therefore sustainable if:

$$B(1 + \delta + \delta^2 + \dots) = \frac{B}{1 - \delta} \geq 2B + \delta \times 0(1 + \delta + \dots) = 2B,$$

that is, if

$$\delta > \frac{1}{2}.$$

We will assume throughout the paper that this condition holds, so that collusion is indeed a concern.

To study the effectiveness of the antitrust policy in deterring collusion "as much as possible", it is useful to introduce some heterogeneity among industries. For the sake of presentation we

⁷For this Bertrand duopoly, perfect collusion on the monopoly price is sustainable whenever *some* collusion is sustainable (i.e., whenever $\delta \geq 1/2$). In more general settings, some collusion might be sustainable even when perfect collusion is not. Our focus on binary decisions (compete or collude) admittedly overlooks this possibility, but allows us to keep the analysis tractable when introducing antitrust and leniency policies.

will assume that δ remains constant across industries, which instead vary in their stakes of collusion, B : the bigger B is, the more profitable is collusion, as well as the short-term gains from a deviation.

2.2 Antitrust enforcement

We assume that collusion leaves some evidence, which the antitrust authority can then find if it investigates the industry; however, due to budget and resource limitations, this happens only with probability ρ , where $0 < \rho < 1$; in addition, each firm can also bring this evidence to the antitrust authority. When a cartel is detected, either through an investigation or because a cartel member provided the incriminating evidence, each firm must pay a fine F . The antitrust policy parameters ρ and F are exogenously fixed. Without loss of generality, we can interpret F as the maximal fine that can be imposed on cartels, since optimal deterrence is usually achieved by setting indeed the fine to its maximal level.⁸ To keep the analysis simple, we assume that the evidence of collusion is generated only if both firms agree on collusion and it lasts only for one period, which implies that the cartel cannot be prosecuted for its past activity.

In each period, the timing of the game is thus as follows:

- Stage 0. Each firm chooses whether to enter into a collusive agreement. If at least one firm chooses not to collude, then competition takes place and the game ends for that period; otherwise:
- Stage 1. Each firm chooses whether to respect the agreement and "collude", or deviate and "compete" on the market. These decisions are not observed by rivals until the end of

⁸This is a standard result in the literature of optimal law enforcement; see e.g. Polinsky and Shavell (2000) for an excellent survey.

We moreover assume here that the maximal fine is independent from the stakes of collusion. In practice, maximal fines are set according to judicial principles, which vary across countries and may be related, directly or indirectly, to the nature and importance of the anticompetitive behavior. The link between fines and the collusion benefits is however often imperfect, since fines are subject to exogenous caps (e.g., 10% of the turnover in EU and \$100 million in the US) and driven by multiple considerations. For instance, the Commission (see European Commission (2006)) determines a first amount based on the value of sales affected by the collusion and on the number of years of infringement. It may then adjust that amount "on the basis of an overall assessment which takes account of all the relevant circumstances." To ensure that fines have a sufficiently deterrent effect, the Commission may moreover "increase the fine to be imposed on undertakings which have a particularly large turnover beyond the sales of goods or services to which the infringement relates."

the period;⁹ then:

- Stage 2. Each firm decides whether to report the evidence to the antitrust agency. The cartel is detected with probability 1 if at least one firm reports, in which case the first informant gets a reduced fine $(1 - q)F$, while the other pays F ; otherwise, the cartel is detected with probability ρ , in which case all firms pay the full fine F .

In the absence of any leniency program, firms never benefit from denouncing a cartel. Thus, in each period collusion brings a net profit of B , minus the expected fine ρF ; the expected discounted value of collusion is therefore equal to

$$V_N \equiv \frac{B - \rho F}{1 - \delta},$$

where the subscript N stands for "Normal collusion". This collusion is sustainable only if ¹⁰

$$V_N \geq 2B - \rho F,$$

or equivalently

$$B \geq \underline{B} \equiv \frac{\delta \rho F}{2\delta - 1}. \quad (1)$$

Collusion is therefore sustainable when its stake is sufficiently large; otherwise, each firm would find it profitable to deviate since the short-term gain from doing so, B , then exceeds the cost of foregone collusion benefits in the future, δV_N . The threshold \underline{B} thus characterizes the effectiveness of antitrust enforcement: antitrust enforcement becomes more successful when \underline{B} increases, as is for example the case when the probability of detection ρ and/or the fine F increase.

⁹If firms can detect deviations before the evidence of collusion becomes obsolete, they could "punish" deviations by denouncing the cartel (which is self-sustainable, since each firm is willing to expose the cartel when it expects the rival to do it anyway). However, when leniency is available, a deviator is likely to "run to the courthouse" and, by choosing the timing of the deviation, beat its rival in this race. As long as this is the case, whether deviations are detected before the end of the period does not affect the analysis.

¹⁰For the sake of exposition we focus on perfect collusion, where firms collude in every period. It can be checked that, as in standard pure Bertrand settings, perfect collusion is here sustainable as soon as firms can collude with positive probability in at least some periods (this is because deviating from collusion always generate the same short-gains, while the value of future collusion increases when it systematically occurs in all periods).

3 Optimal leniency under secret investigations

3.1 Collusive strategies

We now introduce a leniency program, which allows the *first informant* (and only the first one) to benefit from a reduced fine $(1 - q)F$ (or even from a positive reward, if $q > 1$). As we will see, leniency makes "normal" collusion more difficult, but also enlarges the set of collusive strategies. We first consider these two issues, and then characterize the optimal degree of leniency.

Normal collusion.

Firms can still try to collude in every period and never report any evidence to the antitrust agency. Firms then get as before V_N if they stick to such collusion and $2B - \rho F$ if they cheat and compete on the product market; normal collusion can thus again be sustained only when $B \geq \underline{B}$. But when deviating a firm can now denounce the cartel to benefit from leniency; it will indeed have an incentive to do so if the amnesty rate reduces the expected fine that it faces, i.e., if:

$$q > \underline{q} \equiv 1 - \rho > 0.$$

When this condition holds, normal collusion is sustainable only when:

$$V_N = \frac{B - \rho F}{1 - \delta} \geq 2B - (1 - q)F,$$

that is:

$$B \geq B_N^r(q) \equiv \frac{\rho - (1 - \delta)(1 - q)}{2\delta - 1}F, \quad (2)$$

where the superscript r stands for "report collusion". The threshold $B_N^r(q)$ increases with the amnesty rate and is indeed higher than \underline{B} when $q > \underline{q}$.

Alternative collusive strategies.

Firms may however attempt to take advantage of the leniency program and use it to reduce the expected fines they face. They could for example take turns for denouncing the cartel; this supposes that cartels can readily start afresh once they have been exposed. In reality, one would expect the antitrust agency to keep such an industry under close scrutiny, making it difficult to collude for at least some time. Yet firms could start colluding later on and apply again for leniency at some point; more realistically, they may apply for amnesty when they feel that an investigation becomes likely or that the cartel will collapse. For the sake of exposition, we will stick here to the assumption that the antitrust policy is stationary and treats all industries alike; we consider later on the possibility that exposed cartels are subject to a closer scrutiny,

as well as the possibility for the firms to denounce a cartel only when an investigation is already underway.

Given our stationarity assumptions, a relevant alternative strategy is to collude and report systematically the cartel. Assuming that both firms are equally likely to be the first informant, the value of such collusion is given by

$$V_R(q) \equiv \frac{B - \left(1 - \frac{q}{2}\right) F}{1 - \delta},$$

where the subscript R stands for "collude and Report". It is clear that reporting is self-sustainable: if a firm anticipates that the other will report the cartel, it is better to report and apply for leniency as well.

This alternative form of collusion is therefore sustainable as long as firms have no incentive to deviate and compete in the product market:¹¹

$$V_R(q) \geq 2B - \left(1 - \frac{q}{2}\right) F,$$

that is, whenever

$$B \geq B_R(q) \equiv \frac{\delta \left(1 - \frac{q}{2}\right) F}{2\delta - 1}. \quad (3)$$

The threshold $B_R(q)$ *decreases* as the amnesty rate increases: offering additional leniency makes this form of collusion more attractive (V_R increases) and, by the same token, more robust to deviation. In particular, excessive leniency would allow the firms to reduce the expected fine they face and would then foster collusion; this occurs when

$$1 - \frac{q}{2} < \rho,$$

or

$$q > \bar{q} \equiv 2(1 - \rho),$$

in which case this alternative form of collusion is more robust than normal collusion absent leniency: $B_R(q) < \underline{B}$ for any $q > \bar{q}$.

¹¹Conversely, we assume away any coordination problem in the race to reporting, so that no firm can try to beat the other when they both plan to apply for leniency. This excludes the possibility that a firm can deviate in the product market and moreover benefit from denouncing the cartel unilaterally. This assumption seems reasonable here given the fact that denouncing a cartel is a scheduled action of the collusive strategy and each firm will report at the first time.

3.2 Optimal amnesty rate

To sum-up, "normal collusion" is sustainable when

$$B \geq B_N(q) \equiv \max \{\underline{B}, B_N^r(q)\},$$

while "collude and report" is sustainable when $B \geq B_R(q)$. Conversely, it can be checked that no other form of collusion is sustainable if these are not.¹² We now seek to characterize the optimal degree of leniency. The antitrust authority aims to deter cartels in as many industries as possible; the amnesty rate q should therefore maximize the deterrence threshold

$$B(q) \equiv \min \{B_N(q), B_R(q)\},$$

which appears in bold in Figure 1.

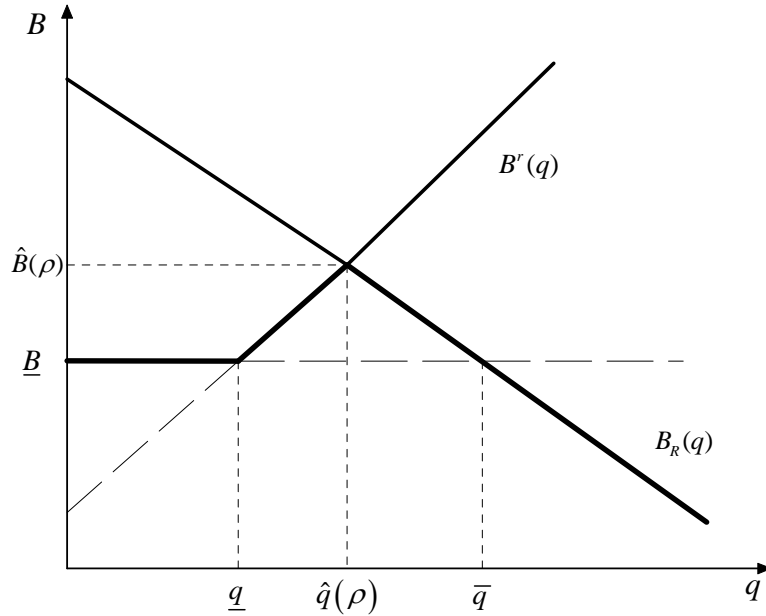


Figure 1: Optimal Amnesty Rate

As noted above, introducing leniency makes normal collusion more fragile as soon as $q > \underline{q}$, and does not excessively foster alternative forms of collusion as long as $q < \bar{q}$; since $\bar{q} = 2\underline{q} > \underline{q}$, it is optimal to offer an amnesty rate $q \in (\underline{q}, \bar{q})$, so as to deter any collusion in industries where, absent leniency, normal collusion could prevail. And since increasing q increases B_N^r (i.e.,

¹²As usual, the two firms should behave symmetrically in order to maximize the scope for collusion, and colluding in every period maximizes the value of future collusion, which contributes to make it more robust to deviations. In addition, randomizing between reporting or not (even using a public lottery to preserve symmetry) is not sustainable when neither "not reporting" nor "always reporting" can be sustained.

destabilizes normal collusion) but decreases B_R (i.e., facilitate "collude and report" strategies), the optimal amnesty rate is such that the two thresholds coincide:

$$B_N^r(q) = \frac{\rho - (1 - \delta)(1 - q)}{2\delta - 1}F = B_R(q) = \frac{\delta \left(1 - \frac{q}{2}\right)}{2\delta - 1}F,$$

which is achieved for

$$q = \hat{q}(\rho) \equiv \frac{1 - \rho}{1 - \frac{\delta}{2}}. \quad (4)$$

From the above analysis, the rate \hat{q} is strictly between $\underline{q} > 0$ and \bar{q} ; it increases as ρ decreases, and it may be desirable to reward informants ($\hat{q} > 1$) when random investigations are not very effective ($\rho < \delta/2$).

The threshold $\hat{B} = B_N^r(\hat{q}) = B_R(\hat{q})$, which characterizes the effectiveness of the leniency program, is equal to

$$\hat{B}(\rho) \equiv \frac{\delta(1 - \delta + \rho)}{(2\delta - 1)(2 - \delta)}F, \quad (5)$$

and is indeed higher than \underline{B} .

The following proposition summarizes the analysis:

Proposition 1 *It is always desirable to offer some leniency; moreover the optimal amnesty rate is determined so as to deter normal collusion, without encouraging collusion with reporting: it is characterized by (4) and increases as the probability of prosecution, ρ , decreases.*

The above analysis shows that, in addition to the "stick" (the probability ρ of investigations), it is useful to offer "carrot" (the amnesty rate q), and all the more so as the stick becomes weaker (\hat{q} increases when ρ decreases).¹³ The best way to fight collusion is to induce firms to cheat and to report the cartel activity, which is why leniency is desirable: $\hat{q}(\rho) > 0$. However, offering leniency encourages firms to "collude and report"; the optimal leniency rate \hat{q} (which, by construction, lies between \underline{q} and \bar{q}) reflects precisely the trade-off between destabilizing normal collusion and not encouraging alternative strategies and is such that, for the "marginal industry" $B = \hat{B}(\rho)$, decreasing q would allow firms to collude in a standard fashion, without fearing a deviation and denunciation, whereas increasing q would allow the firms to "collude and report", without fearing a deviation: $B_N^r(\hat{q}) = B_R(\hat{q}) = \hat{B}(\rho)$.

The same trade-off drives the impact of random audits on the optimal amnesty rate: increasing the frequency of investigations or their performance destabilizes normal collusion and thus

¹³In particular, *rewards* are required in the absence of a stick: $\underline{q} = 1$ for $\rho = 0$, implying that pure leniency programs ($q \leq 1$) are ineffective in the absence of a stick.

tilts the balance in favor of lower amnesty rates. As illustrated in Figure 2, increasing the probability of successful audits from ρ to ρ' has no impact on "collude and report" strategies, and thus does not affect $B_R(q)$, but destabilizes normal collusion ($B_N^r(q; \rho)$ moves up) in the marginal industry and neighboring ones (that is, for B slightly larger than $\hat{B}(\rho)$). A small reduction in the leniency rate q then deters also "collude and report" strategies, while still deterring normal collusion, in these additional industries.

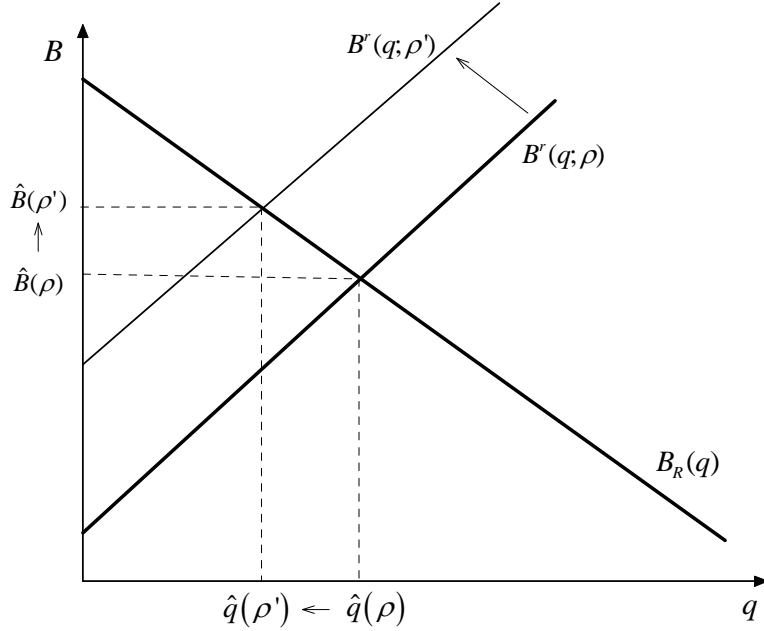


Figure 2: Comparative Statics

The above analysis reflects the trade-off between the main benefit of leniency programs, which is to obtain evidence and contribute in this way to discourage cartel activities, and the often voiced concern that, by reducing the expected fines, leniency may instead foster these cartel activities. To be sure, this concern is taken into consideration in a rather crude way here, since a binding constraint is to prevent cartel members from repeatedly applying for leniency. It could be argued that such strategy is rarely seen in practice. We will discuss some variants in the next sections but note here that, in our model, we would not expect firms to adopt this strategy when the leniency rate is set at its optimal level: industries with collusive stakes lower than \hat{B} cannot collude and, while both forms of collusion are sustainable in industries with larger stakes, "normal collusion" is more valuable:

$$V_N = \frac{B - \rho F}{1 - \delta} > V_R(\hat{q}) \equiv \frac{B - \left(1 - \frac{\hat{q}}{2}\right) F}{1 - \delta} = \frac{B - \left(\frac{1 - \delta + \rho}{2 - \delta}\right) F}{1 - \delta}.$$

If the antitrust authority was instead offering excessive leniency ($q > \hat{q}$), then firms may prefer

to "collude and report" and, in those industries with collusion stakes satisfying $B_R(q) < B < B'_N(q)$, they would indeed rather do so since normal collusion would not be sustainable.

Interestingly, this analysis is validated by laboratory experiments conducted by Hinloopen and Soetevent (2008). They focus on the case $\rho = 0.4$ and $\delta = 0.8$, for which our model predicts that the competition agency should offer full leniency ($\hat{q} = 1$), and compare two treatments. In the first treatment, the amnesty rate is set to the optimal level ($q = 1$) whilst in the second treatment the leniency program is excessively generous, with a rate amounting to $q = 1.8$.¹⁴ They find that players never "collude and report" in the first treatment, while more than 70% of subjects do so in every period in the second treatment, in line with our prediction.

In practice, leniency programs usually grant a reduction of the fine (up to 100%) but rarely offer rewards. According to the above analysis, these leniency rates might well lie below the optimal level, and thus a fortiori below the level which could be exploited by cartels. There is however a growing literature calling for more generous programs, in order to induce more cartels to self-report. For instance, Spagnolo (2004) argues that first-best deterrence could be achieved with high enough fines, by promising first informants a reward equal to the sum of fines levied from other cartel firms (which corresponds to $q = 2$ in our framework). In contrast, this paper calls for a cautious use of high rewards, since these risk being exploited and generate perverse effect (indeed, we find that the optimal leniency rate satisfies $\hat{q} < \bar{q} < 2$).

Remark: First informant rule. We have assumed so far that only the first informant can benefit from leniency. Allowing more than one firm to benefit from amnesty does not affect normal collusion but makes "collude and report" more attractive and therefore more robust, which reduces the effectiveness of the leniency programme. The threshold for "collude and report" strategies decreases when both informants can benefit from leniency:

$$B'_R(q) \equiv \frac{\delta(1-q)F}{2\delta-1} < B_R(q),$$

which leads to a lower leniency rate $q = \underline{q}$ in equilibrium. As a result the equilibrium threshold decreases to \underline{B} , so it is actually optimal to offer no leniency in this case. The leniency program thus performs less well in the absence of "first-informant-wins" rule; this result may explain why the original version of the US leniency program did not contribute much to defeat cartels before

¹⁴More precisely, in this second treatment, the leniency program grants full amnesty to the informant when only one player reports (as in the case of a deviation), but still grants each player a 90% reduction of the fine when both players report. This corresponds indeed to $q = 1.8$ in our framework and thus encourages the players to "collude and report" rather than "collude and never report".

the 1993 revision.¹⁵

3.3 Repeated offenders

The above analysis supposes that firms can repeatedly collude and benefit from leniency. However, in some jurisdictions (such as Greece, for example), amnesty is never offered to a repeated offender.¹⁶ This policy can prevent cartels from adopting repeated "collude and report" strategies, but may also trigger other forms of collusion, such as reporting once and never after. The following analysis shows that this form of collusion may actually be more robust than "collude and report" in the absence of any specific rule for repeated offenders; therefore, ruling out leniency for repeated offenders may actually weaken antitrust enforcement.

Suppose for example that the leniency program is eligible only once in any given industry. This "once only" policy has no direct impact on normal collusion, and prevents firms from colluding and reporting repeatedly. But the cartel can then turn to alternative strategies, such as "report Once and never again" (O); the value of this collusion is given by

$$V_O = B - (1 - \frac{q}{2})F + \delta V_N = B - (1 - \frac{q}{2})F + \delta \frac{B - \rho F}{1 - \delta}$$

After the first report, firms can no longer benefit from leniency and thus have no incentive to report again the cartel; collusion is then sustainable as long as it resists deviations in the product market, i.e. whenever:

$$B \geq \underline{B}.$$

In the first period of this collusion path, firms report the cartel anyway; collusion is thus sustainable as long as it resists again deviations in the product market, i.e., as long as:

$$V_O \geq 2B - (1 - \frac{q}{2})F,$$

which boils down again to

$$B \geq \frac{\delta \rho F}{2\delta - 1} = \underline{B}.$$

Prohibiting leniency for repeated offenders thus creates quite robust alternative collusion strategies: by reporting once, cartel members can make sure that no one has an incentive to report afterwards, which thus stabilizes normal collusion in the future; and since normal collusion is more profitable than alternative collusion strategies, this also contributes to stabilize collusion

¹⁵This is in line with previous insights; see for example Spagnolo (2004) and Harrington (2008).

¹⁶See Wils (2008), footnote 139 at p. 138.

in the first period. As a result, "collude and report once" is sustainable whenever normal collusion is sustainable absent leniency, which renders the leniency program completely ineffective:

Proposition 2 *Restricting leniency to first-time offenders makes it ineffective in deterring collusion.*

Wouter Wils, too, has argued that "[e]xcluding repeat offenders from leniency would mean that leniency would no longer work at all if the same group of companies that was found to have formed a cartel originally subsequently enters a new cartel. If all these companies cannot apply for leniency because they are repeat offenders, this second cartel will thus be more stable than the first. Paradoxically, excluding recidivists from leniency may thus encourage recidivism".¹⁷ Our analysis validates this concern and suggests that the antitrust authority should be cautious before refusing to grant leniency to repeated offenders, unless it can deter exposed cartels from returning to collusion.

3.4 Tighter supervision for cartelized industries

We also assumed so far that an industry is able to collude again immediately once the cartel has been exposed. In practice, however, we would expect the antitrust authorities to keep the industry under closer scrutiny, at least for some time, in order to prevent the industry from colluding again.¹⁸ In practice, this can take the form of more intense monitoring (higher ρ), greater fines (higher F) for repeated offenders, and so forth. To fix ideas, suppose for example that once a cartel is revealed, either by self-reporting or by random investigations, then during T periods the industry is subject to a tight supervision that prevents it from colluding.¹⁹ This reduces the value of "collude and report" strategies to

$$V_R(q; T) \equiv \frac{B - \left(1 - \frac{q}{2}\right) F}{1 - \beta(T)},$$

¹⁷See Wils (2008) at footnote 139, p. 138.

¹⁸More generally, while we have restricted attention to "stationary" antitrust policies, Frezal (2006) points out that non-stationary policies may be more effective even in the absence of leniency programs: targeting specific industries in sequence may prevent firms from colluding for some time, which in turn reduces the attractiveness of collusion and contributes to make it more fragile. A complete analysis of non-stationary investigation and leniency policies remains however beyond the scope of the present paper.

¹⁹Systematic monitoring ($\rho' = 1$) and fines absorbing the observed benefits from collusion ($F' = B$) would clearly achieve this. More generally, the expected fine should be large enough to violate (1) (that is, $\rho' F' > (2\delta - 1) B/\delta$).

where $\beta(T) \equiv \delta^{T+1} < \delta$ represents the relevant discount factor for future collusion in that case, and the value of normal collusion to

$$V_N(T) \equiv \frac{B - \rho F}{1 - \gamma(T)},$$

where $\gamma(T) \equiv (1 - \rho)\delta + \rho\beta(T) > \beta(T)$.

Following the same logic as before, normal collusion is now defeated if:

$$B \geq B_N(q; T) \equiv \max\{B_N^r(q; T), \underline{B}(T)\},$$

where (for readability purposes, we will often omit the argument T in $\beta(T)$ and $\gamma(T)$):

$$B_N^r(q; T) \equiv \frac{\rho - (1 - \gamma)(1 - q)}{2\gamma - 1} F,$$

and

$$\underline{B}(T) \equiv \frac{\gamma\rho F}{2\gamma - 1}.$$

The threshold B_N^r still increases with q and, as before, leniency contributes to further destabilize collusion (i.e., $B_N^r(q; T) > \underline{B}(T)$) as soon as $q > \underline{q} = 1 - \rho$.

Similarly, "collude and report" strategies are deterred if:²⁰

$$B \geq B_R(q; T) \equiv \frac{\beta\left(1 - \frac{q}{2}\right)}{2\beta - 1} F.$$

The optimal amnesty rate q is determined by the same logic as before:

$$B_N^r(q; T) = B_R(q; T),$$

which yields

$$q = \hat{q}(T) \equiv \frac{2(1 - \rho)(\beta + \delta - 1)}{(2\beta - 1) + (2\gamma - 1)(1 - \beta)},$$

and

$$\hat{B}(T) = B_N^r(\hat{q}(T); T) = B_R(\hat{q}(T); T).$$

Since increasing the duration T of scrutiny reduces β and the value of collusion, both $B_N^r(q; T)$ and $B_R(q; T)$, and thus the optimal deterrence threshold $\hat{B}(T)$, increase. But since longer periods of scrutiny hurt "reporting" strategies even more than normal collusion, they allow to destabilize collusion further by offering higher amnesty rates: this deters normal collusion in additional industries, without triggering anymore alternative forms of collusion in these industries:

²⁰We assume that the scrutiny duration T satisfy $\beta(T) < 1/2$ (if $\gamma(T) > 1/2 > \beta(T)$, then the "collude and report" strategy cannot be sustained and it is thus optimal to offer the maximal leniency or reward).

Proposition 3 *Increasing the duration of close scrutiny for exposed cartels not only further destabilizes collusion ($\hat{B}(T)$ increases with T), but also calls for higher amnesty rates ($\hat{q}(T)$ also increases with T).*

Proof. See Appendix A. ■

This proposition shows that our analysis still applies when antitrust authorities can place exposed cartels under closer supervision: offering some leniency (for industries not currently under close supervision) still contributes to deter collusion, and the optimal leniency is actually more generous than in the absence of such closer scrutiny.

We conclude this discussion with two brief remarks.²¹ First, when exposed cartels are subject to closer scrutiny, ruling out leniency for repeated offenders still makes the leniency program less effective, although it does not render it completely ineffective. Second, while the emphasis has been placed so far on preventing the formation of cartels (the deterrence effect of antitrust policy), placing uncovered cartels under tighter supervision also allows antitrust agencies to stop the activity of cartels once they are formed (desistance effect); leniency programs can then contribute to foster desistance as well as deterrence.

4 Amnesty before and after investigations

We now study whether it can be optimal to keep offering leniency once an investigation is already underway. For that purpose, we will refine the modelling of random audits and distinguish between the launch of an investigation and its outcome. More precisely, we will suppose that, in each period:

- in the absence of any report, the antitrust authority launches an investigation with probability α , where $0 < \alpha < 1$;
- when an investigation is launched, in the absence of any report it succeeds in uncovering cartels with probability p , where $0 < p < 1$.

In practice, one would expect α and p to be quite small, due to resource constraints and the inherent difficulties in uncovering hidden evidence.

²¹See Chen and Rey (2007) for further analysis of these two aspects.

4.1 Open or secret investigations?

When the antitrust authority launches an investigation, it can do so openly or try to keep it secret. When investigations are launched secretly, the situation is essentially the same as in the previous section: firms anticipate that a cartel will be caught with probability $\rho = \alpha p$, and the optimal antitrust policy consists in offering the amnesty rate $\hat{q}(\alpha p)$ characterized by Proposition 1;²² it is thus optimal to introduce a leniency program when the overall probability of conviction is small, and the optimal amnesty rate then deters cartels such that

$$B < \hat{B}(\alpha p).$$

When investigations are instead launched publicly, cartel members may choose to report the cartel either before or after an investigation is launched; conversely, the antitrust authority can also adopt different amnesty rates for these two situations. Let q_b and q_a denote respectively the amnesty rates offered to a first informant that would report the cartel *before* and *after*, respectively, an investigation is launched; in each period, the timing of the game becomes:

- Stage 0. Each firm chooses whether to enter into a collusive agreement. If at least one firm chooses not to collude, then competition takes place and the game ends for that period; otherwise:
- Stage 1. Each firm chooses whether to respect the agreement and "collude", or deviate and "compete" on the market. These decisions are again not observed by rivals until the end of the period; then:
- Stage 2. Each firm decides whether to report the evidence to the antitrust agency. The cartel is detected with probability 1 if at least one firm reports, in which case the first informant gets a reduced fine $(1 - q_b)F$, while the others pay F ; otherwise:
- Stage 3. With probability $1 - \alpha$, the antitrust agency launches no investigation and the game ends for that period; with probability α , the antitrust agency launches an investigation and:
- Stage 4. Each firm decides whether to report the evidence to the antitrust agency. The cartel is detected with probability 1 if at least one firm reports, in which case the first

²²If the amnesty rate differs once an investigation is already underway, the relevant amnesty rate is the expected one, $q = \alpha q_a + (1 - \alpha) q_b$.

informant gets a reduced fine $(1 - q_a)F$, while the others pay F ; otherwise, the cartel is detected with probability p , in which case all firms pay the full fine F .

Making investigations public creates additional forms of collusion, since firms can try to abuse the program by reporting for example only when an investigation is launched. However, the antitrust agency can also adjust the amnesty rate once it launches an investigation, and this actually allows antitrust enforcement to remain as effective as with secret investigations. To see this, suppose that the agency grants no leniency once an investigation is started (i.e., $q_a = 0$). Then, firms cannot benefit from reporting the cartel once an investigation is underway, since doing so would increase the probability of prosecution (from p to 1), without any reduction in the fine. Thus, cartel members' only relevant choice is between "never reporting" and "reporting before an investigation is launched". But this choice is essentially the same as the one they face (between "normal collusion" and "collude and report") when investigations are launched secretly, and thus the antitrust agency can still perform as well as with open investigations as it can with secret ones.

We now study whether the antitrust agency can perform strictly better with open investigations than with secret ones. In the light of the above discussion, this will be the case whenever it is optimal to offer some leniency even once an investigation is already underway.

4.2 Collusive strategies

Three types of collusive strategies become relevant in the case of open investigations: besides the previous ones, i.e., normal collusion, where firms never report the cartel, and "collude and report", where firms systematically report the cartel to benefit from reduced fines, a new form of collusion consists in reporting only in case of investigation (I). We now characterize conditions under which firms can sustain these forms of collusion.

Normal collusion (N)

The value of normal collusion is now equal to

$$V_N = \frac{B - \alpha p F}{1 - \delta}.$$

As before, deviating in the product market (and not reporting) is not profitable if:²³

$$B \geq \underline{B} = \frac{\delta \alpha p F}{2\delta - 1},$$

²³For the sake of exposition, we keep the same notation for the thresholds already analyzed in the previous section, although ρ is now replaced with αp .

whereas deviating and reporting at stage 2 is not attractive if:

$$B \geq B_N^r(q_b) = \frac{\alpha p - (1 - \delta)(1 - q_b)}{2\delta - 1} F.$$

In addition, cartel members should not benefit from defecting and reporting in case of investigation, which requires:

$$V_N \geq 2B - \alpha(1 - q_a)F,$$

or equivalently:

$$B \geq B_N^i(q_a) \equiv \frac{\alpha p - \alpha(1 - \delta)(1 - q_a)}{2\delta - 1} F.$$

Finally,²⁴ a cartel member who did not cheat in the product market could choose to report in case of an investigation; this is not profitable if:

$$-pF + \delta V_N \geq -(1 - q_a)F$$

or equivalently

$$B \geq B_N^c(q_a) \equiv \frac{p(1 - \delta + \alpha\delta) - (1 - \delta)(1 - q_a)}{\delta} F.$$

Hence, normal collusion is sustainable if and only if:

$$B \geq B_N(q_b, q_a) \equiv \max \{ B_N^i(q_a), B_N^r(q_b), B_N^c(q_a), \underline{B} \}. \quad (6)$$

Collude and Report systematically (R)

This strategy is again self-sustainable at stage 2 (even in case of a deviation in the product market), since it is again a best response to report when others will report anyway. This strategy is therefore sustainable when it resists deviations in the product market, which is the case when:

$$B \geq B_R(q_b) = \frac{\delta \left(1 - \frac{q_b}{2}\right) F}{2\delta - 1}.$$

Collude and report in case of Investigation is launched (I)

Reporting in case of investigation is self-sustainable, irrespective of whether a firm deviates in the product market or not: given that the other will report, reporting is profitable since it reduces the expected fine by $q_a/2$. Therefore, to be sustainable this new form of collusion only need to resist two types of deviations, which consist in deviating and reporting at stage 2 or only in case of investigation.²⁵

²⁴Another possible deviation, "collude but report before investigation", is dominated by "deviate and report before investigation".

²⁵The other possible deviations are "collude but report at stage 2", which is dominated by "deviate and report at stage 2", and "deviate but never report", which is dominated by "deviate and report at stage 2".

Firms have no incentives to deviate and report at stage 2 if

$$V_I = \frac{B - \alpha \left(1 - \frac{q_a}{2}\right) F}{1 - \delta} \geq 2B - (1 - q_b) F,$$

that is:

$$B \geq B_I^r(q_b, q_a) \equiv \frac{\alpha \left(1 - \frac{q_a}{2}\right) - (1 - \delta)(1 - q_b)}{2\delta - 1} F.$$

Similarly, deviating and reporting in case of investigation is not profitable if:

$$V_I \geq 2B - \alpha \left(1 - \frac{q_a}{2}\right) F,$$

or:

$$B \geq B_I^i(q_a) \equiv \frac{\delta \alpha \left(1 - \frac{q_a}{2}\right)}{2\delta - 1} F.$$

Hence, this collusion is sustainable if and only if:

$$B \geq B_I(q_b, q_a) \equiv \max \{B_I^i(q_a), B_I^r(q_b, q_a)\}. \quad (7)$$

4.3 Optimal leniency policy

To deter collusion in as many industries as possible, the amnesty rates q_b and q_a should maximize the deterrence threshold:

$$B(q_b, q_a) \equiv \min \{B_N(q_b, q_a), B_I(q_b, q_a), B_R(q_b)\}.$$

In the absence of post-investigation leniency, the optimal (pre-investigation) leniency rate is $\hat{q}_b = \hat{q}(\alpha p)$, characterized by $B_N^r(\hat{q}_b) = B_R(\hat{q}_b)$ and equal to

$$\hat{q}(\alpha p) = \frac{1 - \alpha p}{1 - \frac{\delta}{2}};$$

the deterrence threshold is then:

$$\hat{B}(\alpha p) = \frac{1 - \delta + \alpha p}{2 - \delta} \frac{\delta F}{2\delta - 1}. \quad (8)$$

Offering some leniency during investigations ($q_a > 0$) provides another way to destabilize collusion, by inducing cartel members to report in case of investigation, but may also encourage firms to "collude and report in case of investigation". As we will see, offering some leniency during investigations is indeed useful whenever these investigations are unlikely to succeed in the absence of cooperation.

Adopting a post-investigation leniency $q_a > 0$ can destabilize normal collusion by inducing cartel members to deviate and report in case of investigation, or to collude but report in case of investigation; we consider these two possibilities in turn. Inducing additional cartel members to "deviate and report in case of investigation" requires $B_N^i(q_a) > B_N^r(\hat{q}_b)$, or:

$$\alpha(1 - q_a) < 1 - \hat{q}(\alpha p).$$

At the same time, the policy prevents firms from adopting the "collude and report in case of investigation" strategy whenever $B_I^r(\hat{q}_b, q_a) > B_N^r(\hat{q}_b)$, or:

$$p < 1 - \frac{q_a}{2}.$$

These two conditions can be met when

$$p < \tilde{p}_1(\alpha) \equiv \frac{\alpha(2 - \delta) - \delta}{2\alpha(1 - \delta)}. \quad (9)$$

Thus, when random investigations are unlikely to succeed, offering some leniency during investigations can further destabilize "normal collusion" and the "collude and report in case of investigation" strategy (that is, there exists \tilde{q}_a such that $B_N(\hat{q}_b, \tilde{q}_a), B_I(\hat{q}_b, \tilde{q}_a) > B(\hat{q}_b, 0) = \hat{B}(\alpha p)$). Slightly reducing the pre-investigation leniency rate, from \hat{q}_b to \tilde{q}_b , then maintains these thresholds above $\hat{B}(\alpha p)$ while discouraging as well additional cartels from adopting the "collude and report systematically" strategy ($B_R(\tilde{q}_b) > B_R(\hat{q}_b)$ since $B_R(q_b)$ increases as q_b decreases). The leniency policy $(\tilde{q}_b, \tilde{q}_a)$ is thus more effective:

$$B(\tilde{q}_b, \tilde{q}_a) = \min\{B_N^i(\tilde{q}_a), B_I^r(\tilde{q}_b, \tilde{q}_a), B_R(\tilde{q}_b)\} > \hat{B}(\alpha p).$$

Since $B_I^r(q_b, q_a)$ decreases while $B_N^i(q_a)$ increases with q_a , this discussion suggests that the best amnesty rate q_a is such that these two thresholds coincide: $B_I^r(q_b, q_a) = B_N^i(q_a)$. Similarly, since $B_R(q_b)$ decreases while $B_I^r(q_b, q_a)$ increases with q_b , the best amnesty rate q_b is such that

these two thresholds also coincide, as illustrated by Figure 3:

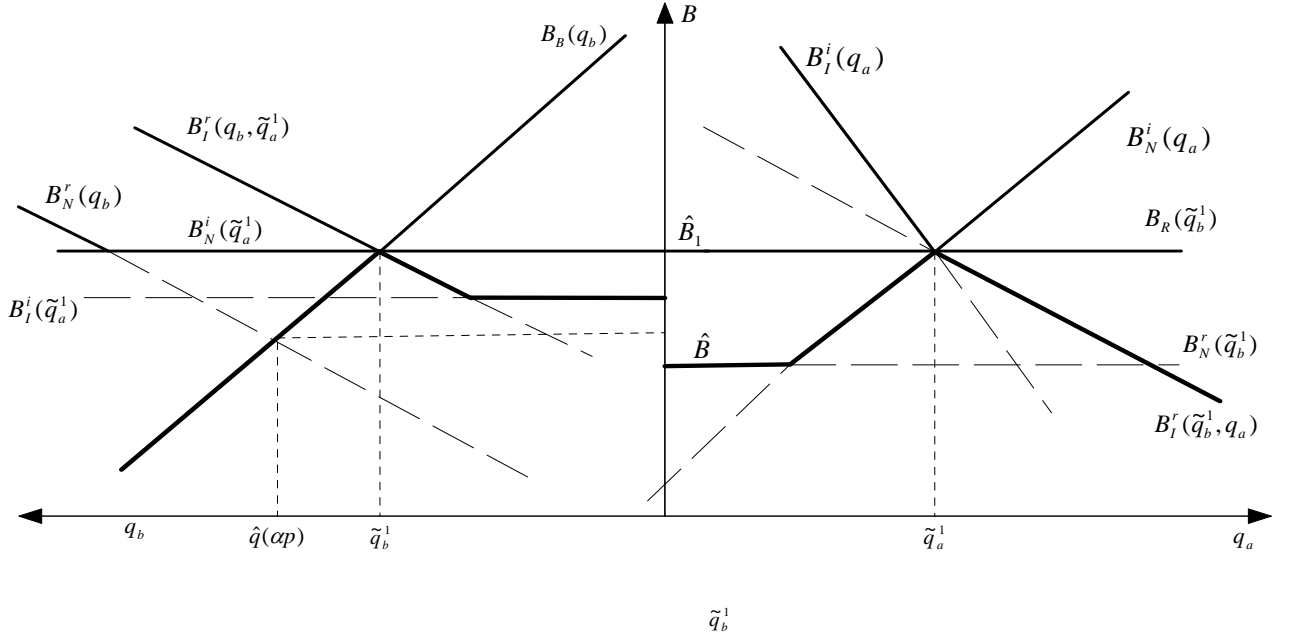


Figure 3: Optimal Amnesty Rates

The candidate optimal amnesty policy $(\tilde{q}_b^1, \tilde{q}_a^1)$ is thus characterized by:

$$\tilde{B}_1 = B_N^i(\tilde{q}_a^1) = B_I^r(\tilde{q}_b^1, \tilde{q}_a^1) = B_R^r(\tilde{q}_b^1),$$

which yields:

$$\begin{aligned} \tilde{q}_b^1(\alpha, p) &\equiv 2 \frac{(1-\alpha)(2-\delta) + \alpha(1-p)}{2(1-\delta)^2 + (2-\delta)}, \\ \tilde{q}_a^1(\alpha, p) &\equiv \frac{2\delta(1-\delta)(1-\alpha) + \alpha(1-p)(2-\delta)}{\alpha(2(1-\delta)^2 + 2-\delta)}. \end{aligned} \quad (10)$$

The resulting threshold of deterrence is

$$\tilde{B}_1(\alpha, p) \equiv \frac{2(1-\delta)^2 + \alpha(1-\delta+p)}{2(1-\delta)^2 + 2-\delta} \frac{\delta F}{2\delta-1}, \quad (11)$$

where $\tilde{B}_1(\alpha, p) \geq \hat{B}(\alpha p)$ if and only if $p \leq \tilde{p}_1(\alpha)$.

Alternatively, destabilizing normal collusion by inducing cartel members to "collude but report in case of investigation" requires $B_N^c(q_a) > B_N^r(\hat{q}_b)$, or

$$q_a > \frac{\delta \hat{q}_b}{2\delta-1} - \frac{1-\delta + (2\delta-1-2\alpha\delta)p}{2\delta-1}.$$

Avoiding the "collude and report in case of investigation" strategy still requires $p < (1 - \frac{q_a}{2})$, which is possible if:

$$p \leq \tilde{p}_2(\alpha) \equiv \frac{(1-\delta)(3\delta-2)}{(2\delta-1)(2-\delta) + 2\alpha\delta(1-\delta)}, \quad (12)$$

where $\tilde{p}_2(\alpha)$ is positive when $\delta > 2/3$. Conversely, under this condition there exists a leniency rate q_a satisfying $B_N^c(q_a) > B_N^r(\hat{q}_b)$ and $B_I^r(\hat{q}_b, q_a) > B_N^r(\hat{q}_b)$. A slight reduction in the pre-investigation leniency rate then maintains these inequalities and destabilizes as well the "collude and report systematically" strategy in additional industries, thereby increasing the deterrence threshold. Using the same logic as above, the candidate optimal leniency policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ is such that:

$$\tilde{B}_2 = B_N^c(\tilde{q}_a^2) = B_I^r(\tilde{q}_b^2, \tilde{q}_a^2) = B_R(\tilde{q}_b^2),$$

and is thus given by

$$\begin{aligned}\tilde{q}_b^2(\alpha, p) &\equiv \frac{2[2(1-\delta)(2\delta-1) + \alpha(3\delta^2 - 3\delta + 1) - (2\delta-1)(1-\delta + \alpha\delta)\alpha p]}{2(2\delta-1)(1-\delta)(2-\delta) + \alpha\delta^2}, \\ \tilde{q}_a^2(\alpha, p) &\equiv \frac{2[(1-\delta)(5\delta - \delta^2 - 2) + \alpha\delta^2 - (2\delta-1)(2-\delta)(1-\delta + \alpha\delta)p]}{2(2\delta-1)(1-\delta)(2-\delta) + \alpha\delta^2};\end{aligned}\quad (13)$$

while the resulting deterrence threshold is

$$\tilde{B}_2(\alpha, p) \equiv \frac{2(1-\delta)^2 + \alpha(1-\delta) + (1-\delta + \alpha\delta)\alpha p}{2(2\delta-1)(1-\delta)(2-\delta) + \alpha\delta^2} \delta F, \quad (14)$$

where $\tilde{B}_2(\alpha, p) \geq \hat{B}(\alpha p)$ if and only if $p \leq \tilde{p}_2(\alpha)$.

It remains to compare the two candidate policies. When investigations are frequent, cartel members would rather "deviate and report in case of investigation" than "collude but report in case of investigation" if they intend to defect from normal collusion, since in the latter case collusion is preserved with only a relatively small probability $(1-\alpha)$, making its expected value less attractive than the short-term gain from a deviation in the product market. Indeed, $\tilde{B}_1(\alpha, p) \geq \tilde{B}_2(\alpha, p)$ if and only if:

$$p \geq \tilde{p}(\alpha) \equiv \frac{2(1-\delta)(2\delta-1) + \alpha(2\delta^2-1) - \alpha^2\delta}{\alpha(2\delta-1)(3-2\delta) + 4\alpha^2(1-\delta)^2},$$

which always holds for any $\alpha \geq \alpha_1 (< 1)$, where α_1 solves $\tilde{p}(\alpha_1) = 0$ and is given by

$$\alpha_1 = \frac{2\delta-1}{\delta}.$$

We can therefore distinguish two cases:

Case (1): $\delta > 2/3$. When firms are sufficiently patient, either policy can be optimal, as illustrated in Figure 4. By construction, the loci of $\tilde{p}_1(\alpha)$, $\tilde{p}_2(\alpha)$, and $\tilde{p}(\alpha)$ coincide at a common point (α_0, p_0) , and the set of parameters (α, p) can be divided into three regions. When investigations are likely to be successful (Region A), namely, when $p \geq \tilde{p}_2(\alpha)$ for $\alpha \leq \alpha_0$ and $p \geq \tilde{p}_1(\alpha)$ for $\alpha \geq \alpha_0$, it is optimal to stop offering leniency once an investigation has been

launched; the optimal policy is then given by $(q_b^* = \hat{q}(\alpha p), q_a^* = 0)$. If instead investigations are less likely to be successful (in particular, whenever $p < p_0$), then it is desirable to keep offering some leniency even when an investigation is already underway. If investigations are sufficiently frequent (Region B), it is then optimal to offer $(q_b^* = \tilde{q}_b^1, q_a^* = \tilde{q}_a^1)$, so as to destabilize normal collusion by inducing cartel members to "deviate and report in case of investigation"; if investigations are infrequent (Region C), however, it is instead optimal to offer $(q_b^* = \tilde{q}_b^2, q_a^* = \tilde{q}_a^2)$, so as to destabilize normal collusion by inducing cartel members to "collude but report in case of investigation".

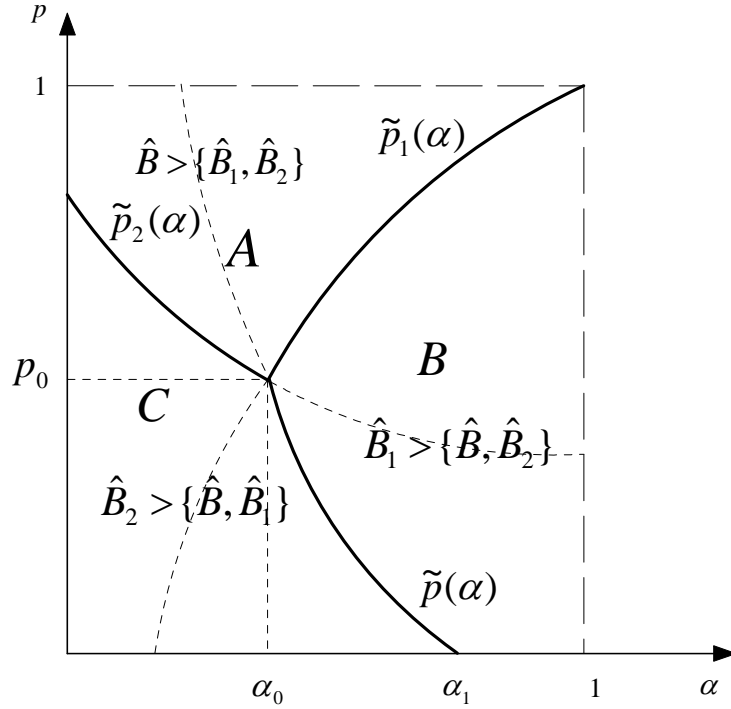


Figure 4: Optimal Leniency Policies: $\delta > 2/3$

Case (2): $1/2 < \delta \leq 2/3$. When firms are less patient, $\tilde{p}_2(\alpha) \leq 0$ and the policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ is thus always dominated by $(\hat{q}(\alpha p), 0)$ for any p and α . In this case, illustrated in Figure 5, when investigations are quite infrequent (namely, $\alpha \leq \underline{\alpha} \equiv \delta/(2 - \delta)$), it is optimal to deny leniency once an investigation is underway, whatever the effectiveness p of these investigations; when investigations are more frequent, however, it is optimal to keep offering some leniency during an

investigation if it is unlikely to succeed otherwise (namely, when $p \leq \tilde{p}_1(\alpha)$).

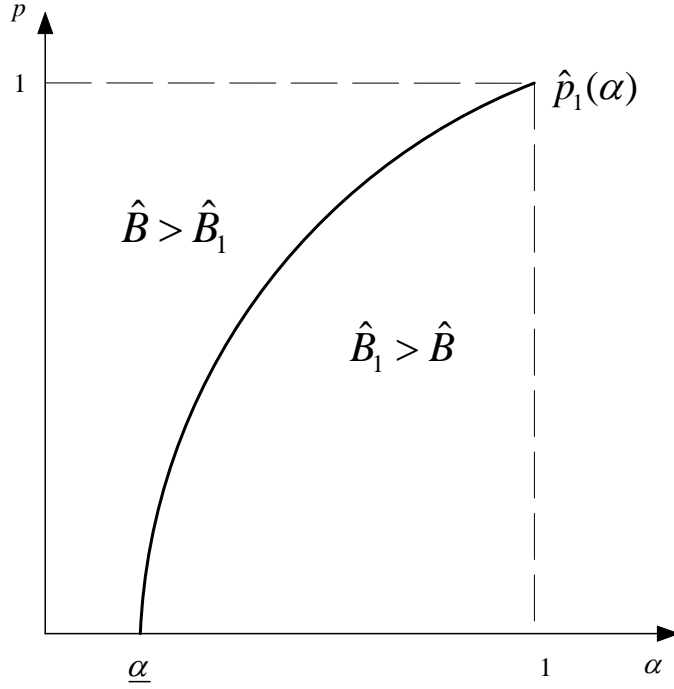


Figure 5: Optimal Leniency Policies: $\delta \leq 2/3$

These insights are formally confirmed by the following proposition:

Proposition 4 *It is always optimal to offer leniency before investigations; moreover:*

- *When firms are patient ($\delta > 2/3$), it is optimal to keep offering some leniency in case of investigation if and only if $p \leq \max\{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$; the optimal policy is then $(q_b^*, q_a^*) = (\tilde{q}_b^1(\alpha, p), \tilde{q}_a^1(\alpha, p))$ if $p \geq \tilde{p}(\alpha)$ and $(q_b^*, q_a^*) = (\tilde{q}_b^2(\alpha, p), \tilde{q}_a^2(\alpha, p))$ if $p \leq \tilde{p}(\alpha)$;*
- *When instead firms are impatient ($\delta \leq 2/3$), it is optimal to keep offering some leniency in case of investigation if only if $\alpha \geq \underline{\alpha}$ and $p \leq \tilde{p}_1(\alpha)$, in which case the optimal policy is $(q_b^*, q_a^*) = (\tilde{q}_b^1(\alpha, p), \tilde{q}_a^1(\alpha, p))$.*

Proof. See Appendix B. ■

This proposition characterizes the optimal leniency policy, as a function of the frequency of investigations (α) and the probability that an investigation is successful in the absence of informant (p). In particular, it provides conditions under which it is optimal to keep offering leniency even when an investigation is already underway. When firms are patient and collusion is thus robust, it is optimal to grant leniency post-investigation whenever these investigations are unlikely to succeed; when firms are less patient, however, granting amnesty post-investigation

is desirable only if investigations are moreover sufficiently frequent. In practice, we would expect the probability of success p to be quite small, due to resource constraints and to the difficulties in uncovering hidden evidence; leniency is then also desirable once an investigation is already underway, in order to induce cartel members to bring evidence. Offering amnesty post-investigation therefore appears a valuable complement to *ex nihilo* investigations, when antitrust authorities have only limited detection tools or investigation powers.²⁶

4.4 Comparative statics

We now explore further the relation between the "stick" (measured by α and p) and the "carrot" (the amnesty rates). When random investigations are likely to uncover cartels (i.e., $p > \{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$), it is optimal to deny leniency during investigations; the optimal (pre-investigation) leniency rate is then determined as before ($q_b = \hat{q}(\alpha p)$) and decreases as the overall probability of prosecution, αp , increases. When it is instead unlikely to detect a cartel in the absence of reporting (i.e., $p < \{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$), the optimal policy is of the form $(\tilde{q}_b, \tilde{q}_a)$ and thus offers leniency both before and after investigations are launched. Under this policy, the marginal industry, characterized by $B = \tilde{B}$, is tempted to defect: (i) from normal collusion, by reporting whenever an investigation is launched (on top of deviating on the product market, when $(\tilde{q}_b, \tilde{q}_a) = (\tilde{q}_b^1, \tilde{q}_a^1)$); (ii) from the "collude and report in case of investigation" strategy, by deviating on the product market and reporting at once, before an investigation may be launched; and (iii) from the "collude and report systematically" strategy, by simply deviating on the product market.

Increasing p further destabilizes normal collusion but does not affect the alternative forms of collusion, since they all involve reporting in case of investigation (and possibly even before); in order to deter as well these forms of collusion, it is then optimal to decrease both \tilde{q}_b (otherwise, the "collude and report systematically" would remain as robust as before) and \tilde{q}_a (otherwise, the "collude and report in case of investigation" would become more robust, due to the reduction in q_b).

Similarly, increasing the frequency of investigation α destabilizes both normal collusion and the "collude and report in case of investigation" strategy, and thus enhances deterrence. And

²⁶If deviations in the product market can be quickly detected at Stage 1, it is no longer possible to "deviate and report in case of investigation", as the rival will report in the meantime; as long as a deviator can beat its rival in the race for leniency, the other sustainability conditions remain unchanged and the two relevant policies are thus $(\hat{q}(\alpha p), 0)$ and $(\tilde{q}_b^2, \tilde{q}_a^2)$; offering post-investigation leniency is then optimal if and only if $p \leq \tilde{p}_2(\alpha)$.

since this does not directly affect the "collude and report systematically" strategy, the optimal pre-investigation amnesty rate q_b necessarily decreases.

The impact of the frequency of investigation α on the post-investigation leniency rate q_a is however less clear, since both of these dimensions affect the "collude and report in case of investigation" strategy as well as normal collusion. It can be checked that, under the policy $(\tilde{q}_b^1, \tilde{q}_a^1)$, an increase in α (together with the appropriate adjustment of the pre-investigation rate q_b) destabilizes normal collusion more than the "collude and report in case of investigation" strategy; as a result, it is optimal to reduce q_a in order to destabilize further the latter form of collusion. In contrast, under the policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ the same change in α destabilizes more the "collude and report in case of investigation" strategy, and as a result it is optimal to increase q_a in order to destabilize normal collusion as well.

The following proposition formally confirms this discussion:

Proposition 5 *Increasing p or α makes the leniency program more effective, moreover:*

- *Increasing p leads to offering less leniency, both pre- and post-investigation;*
- *Increasing α leads to offering less leniency pre-investigation; the post-investigation rate \tilde{q}_a^1 also decreases as α increases, but \tilde{q}_a^2 instead increases with α .*

Proof. See Appendix C. ■

It is worth noting that offering leniency during investigations leads to reduce pre-investigation rates (that is, $\tilde{q}_b^1, \tilde{q}_b^2 < \hat{q}$). The intuition is straightforward. In the absence of post-investigation leniency, the optimal amnesty rate, \hat{q} , leaves marginal industries on the verge of sustaining "normal collusion" as well as the "collude and report" strategy. Introducing post-investigation leniency can then be used to destabilize normal collusion, by inducing cartel members to report in case of investigation, but this calls for a reduction in the pre-investigation leniency, in order to destabilize as well the other form of collusion.

When leniency is also offered in case of investigation, the comparison between the pre- and post-investigation rates depends on the effectiveness of random investigation in the absence of cooperation from the firms:

Proposition 6 *The leniency policies $(\tilde{q}_b^1, \tilde{q}_a^1)$ and $(\tilde{q}_b^2, \tilde{q}_a^2)$ are such that more leniency is offered*

during investigations when these are unlikely to succeed in uncovering cartels:

$$\begin{aligned} \tilde{q}_b^1 &< \tilde{q}_a^1 \text{ if and only if } p < \hat{p}_1(\alpha) \equiv (1-\delta) \frac{(\alpha^2 + \delta(1-\alpha))}{\alpha(2-\delta-\alpha)}, \\ \tilde{q}_b^2 &< \tilde{q}_a^2 \text{ if and only if } p < \hat{p}_2(\alpha) \equiv \frac{1-\delta}{2\delta-1} \frac{\delta(1-\delta) + \alpha(2\delta-1)}{(1-\delta+\alpha\delta)(2-\delta-\alpha)}, \end{aligned}$$

Proof. See Appendix D. ■

Offering more generous leniency during investigations is therefore useful when these investigations are not too effective, as illustrated by Figure 6.

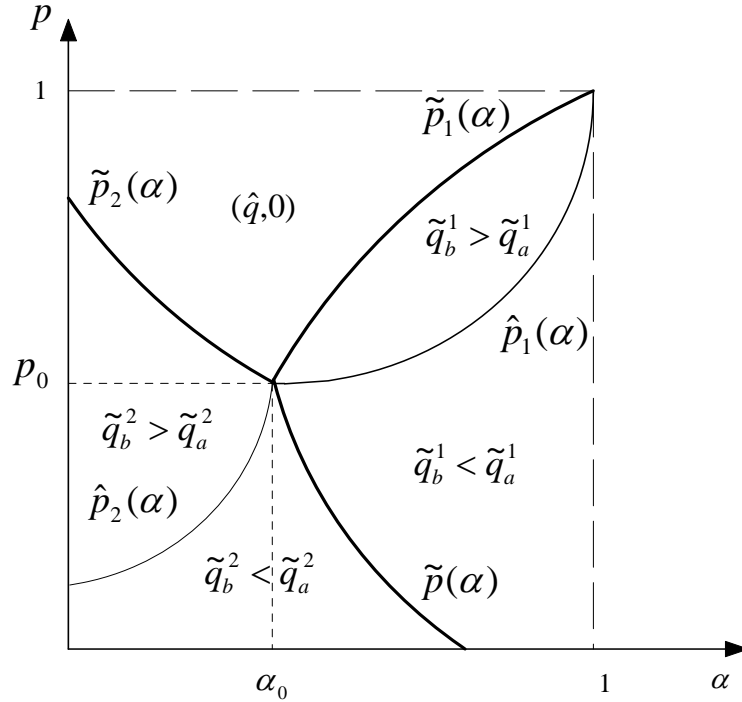


Figure 6: Comparative Statics

In practice, leniency programs do not offer more amnesty post-investigation than pre-investigation. For instance, the U.S. leniency program offers complete amnesty to the first informant, whether an investigation is underway or not; and the EU leniency program offers to the first informant a 75%-100% reduction of fines before investigation, but only a 50%-75% reduction once an investigation is started. Our analysis suggests that such policies may not be optimal when investigations are relatively unlikely to succeed in the absence of self-reporting.

5 Concluding remarks

We develop a simple normative framework for the design of leniency programs which highlights a basic trade-off between deterring the formation of cartels, by encouraging firms to report

illegal activities, and encouraging them by reducing the expected fines when cartels are caught. The framework relies on a standard model of tacit collusion and focuses on stationary antitrust policies based on random investigations and fines. We show that offering leniency can indeed help fighting collusion.

Our simple framework allows us to relate the optimal leniency policy to the frequency and effectiveness of investigations. In particular, it is always optimal to offer some leniency before investigations; it is moreover optimal to keep offering leniency once an investigation is underway, if its probability of success is small in the absence of cooperation from the firms. Our analysis also confirms the usefulness of restricting leniency to the first informant only. In contrast, it does not support prohibiting leniency for repeated offenders.

Although we have assumed that the enforcement measures, as characterized by the investigation rate α and likelihood of detection p , are given exogenously, the framework can also allow to consider further the impact of leniency programs on the optimal allocation of enforcement resources under budget constraint;²⁷ we leave these topics for future research.

²⁷We could for example apply modelling approach developed by Motta and Polo (2003) to incorporate budget constraints and characterize the optimal allocation of enforcement resources that maximizes the deterrence threshold.

Appendices

Appendix A: Proof of Proposition 3

We have:

$$\frac{\partial B_N^r(q; T)}{\partial T} = \frac{\partial B_N^r(q; T)}{\partial \gamma} \gamma'(T) = \frac{(1-q-2\rho)\gamma'(T)}{(2\gamma-1)^2} F > 0,$$

as $q > \underline{q} = 1 - \rho$ and $\gamma' < 0$; and by analogy one can check that both $\underline{B}(q; T)$ and $B_R(q; T)$ are increasing in T .

To check the impact of T on \hat{q} , write \hat{q} as (using $\gamma = \gamma(\beta) = \beta + (1 - \rho)(\delta - \beta)$)

$$\hat{q} = \frac{2(1-\rho)(\beta + \delta - 1)}{(2\beta - 1) + (2\gamma - 1)(1 - \beta)} = \frac{2(1-\rho)(\beta + \delta - 1)}{(2\beta - 1)(2 - \beta) + 2(1-\rho)(1 - \beta)(\delta - \beta)}$$

Then:

$$\begin{aligned} \frac{\partial \hat{q}}{\partial \beta} &= \frac{2(1-\rho)}{D^2} ((2\beta - 1)(2 - \beta) + 2(1-\rho)(1 - \beta)(\delta - \beta) - (2\rho - 2\delta - 4\beta\rho + 2\delta\rho + 3)(\beta + \delta - 1)) \\ &\equiv \frac{2(1-\rho)}{D^2} \phi(\beta), \end{aligned}$$

where D is the denominator of \hat{q} and the function ϕ satisfies

$$\phi'(\beta) = 4\rho(\beta + \delta - 1) > 0$$

and:

$$\phi(\delta) = -(1 + 2\rho)(2\delta - 1)(1 - \delta) < 0.$$

Therefore, starting from $T = 0$ (and thus $\gamma = \beta = \delta$), increasing T reduces β below δ (which implies $\phi(\beta) < \phi(\delta) < 0$) and thus leads to increase \hat{q} . [When T is large enough, $\beta(T) \leq 1/2$ and "collude and report" is no longer sustainable; it would then be optimal to offer "maximal leniency".]

Appendix B: Proof of Proposition 4

To deter collusion in as many industries as possible, the amnesty rates q_b and q_a should maximize the deterrence threshold:

$$B(q_b, q_a) = \min \{B_N(q_b, q_a), B_I(q_b, q_a), B_R(q_b)\},$$

where

$$\begin{aligned} B_N(q_b, q_a) &= \max \{B_N^i(q_a), B_N^r(q_b), B_N^c(q_a), \underline{B}\}, \\ B_I(q_b, q_a) &= \max \{B_I^i(q_a), B_I^r(q_b, q_a)\}. \end{aligned}$$

Since the antitrust authority can always do as well as with secret investigations, *some leniency* is optimal; in particular, refusing leniency during investigations ($q_a = 0$) and setting the pre-investigation amnesty rate to the same level as with secret investigation ($q_b = \hat{q} > 0$) still deters collusion in industries $B < \hat{B}$, where \hat{q} and \hat{B} are such that:

$$B_N^r(\hat{q}) = B_R(\hat{q}) = \hat{B} = \frac{1 - \delta + \alpha p}{2 - \delta} \frac{\delta F}{2\delta - 1} > \underline{B} = \frac{\delta \alpha p F}{2\delta - 1}.$$

Increasing $B(q_b, q_a)$ above \hat{B} requires $B_R(q_b) > \hat{B}$, which implies $B_N^r(q_b) < \hat{B}$ since B_R and B_N^r respectively decreases and increases with q_b . This can therefore be achieved only by adopting q_a such that

$$B_N(q_b, q_a) = \max \{B_N^i(q_a), B_N^c(q_a)\} > \hat{B} > B_N^r(q_b), \underline{B}.$$

In addition, since decreasing q_b enhances the threshold B_R , q_b should be set "as low as possible". However, since

$$B_I(0, q_a) = \max \left\{ \frac{\delta F - \left(1 - \alpha \left(1 - \frac{q_a}{2}\right)\right) F}{2\delta - 1}, \frac{\delta \alpha \left(1 - \frac{q_a}{2}\right) F}{2\delta - 1} \right\} < B_R(0) = \frac{\delta F}{2\delta - 1},$$

the relevant constraint will be to deter the "collude and report in case of investigation" strategy as well (that is, $B_R(q_b) \leq B_I(q_b, q_a) = \max \{B_I^i(q_a), B_I^r(q_b, q_a)\}$), and this constraint will moreover be binding. This, in turn, implies that $B_I^r(q_b, q_a)$ is the relevant term for $B_I(q_b, q_a)$. To see this, note that both B_I^r and B_I^i are decreasing functions of q_a , that B_I^r decreases faster than B_I^i ($|\partial B_I^i / \partial q_a| = \delta |\partial B_I^r / \partial q_a|$) and that they cross at a point where they lie below $B_R(q_b)$: indeed,

$$B_I^r(q_b, q_a) = \frac{\alpha \left(1 - \frac{q_a}{2}\right) - (1 - \delta)(1 - q_b)}{2\delta - 1} = B_I^i(q_a) = \frac{\delta \alpha \left(1 - \frac{q_a}{2}\right)}{2\delta - 1}$$

implies

$$\alpha \left(1 - \frac{q_a}{2}\right) = (1 - q_b) < 1 - \frac{q_b}{2}$$

and thus

$$B_I^r(q_b, q_a) = B_I^i(q_a) = \frac{\delta \alpha \left(1 - \frac{q_a}{2}\right)}{2\delta - 1} < B_R(q_b) = \frac{\delta \left(1 - \frac{q_b}{2}\right)}{2\delta - 1} F.$$

Therefore, $B_I(q_b, q_a) = \max \{B_I^r(q_b, q_a), B_I^i(q_a)\} = B_R(q_b)$ is equivalent to $B_I^r(q_b, q_a) = B_R(q_b)$, which implies $B_I^r(q_b, q_a) > B_I^i(q_a)$, and the equality is achieved by setting q_b to

$$\tilde{q}_b(q_a) \equiv \frac{2(1 - \alpha) + \alpha q_a}{2 - \delta} > 0,$$

which yields:

$$B_I^r(q_b, q_a) = B_R(q_b) = \tilde{B}(q_a) \equiv \left(1 - \frac{1 - \alpha \left(1 - \frac{q_a}{2}\right)}{2 - \delta}\right) \frac{\delta F}{2\delta - 1}.$$

To recap, increasing the deterrence threshold above \hat{B} requires: (i) setting q_a such that $B_N(q_b, q_a) = \max\{B_N^i(q_a), B_N^c(q_a)\} > B_N^r(q_b), \underline{B}$ and (ii) setting q_b such that $B_I^r(q_b, q_a) = B_R(q_b) = \tilde{B}(q_a)$. Since $\max\{B_N^i(q_a), B_N^c(q_a)\}$ and $\tilde{B}(q_a)$ respectively increase and decrease with q_a , the candidate optimal policy must satisfy

$$\max\{B_N^i(q_a), B_N^c(q_a)\} = \tilde{B}(q_a).$$

The equality $\tilde{B}(q_a) = B_N^i(q_a)$ yields $q_a = \tilde{q}_a^1$ and $q_b = \tilde{q}_b(\tilde{q}_a^1) = \tilde{q}_b^1$, while the equality $\tilde{B}(q_a) = B_N^c(q_a)$ yields $q_a = \tilde{q}_a^2$ and $q_b = \tilde{q}_b(\tilde{q}_a^2) = \tilde{q}_b^2$, where the amnesty rates $\tilde{q}_b^1, \tilde{q}_a^1, \tilde{q}_b^2$ and \tilde{q}_a^2 are given by (10) and (13). The corresponding deterrence thresholds, \tilde{B}_1 and \tilde{B}_2 are respectively given by (11) and (14).

By construction, $\tilde{B}_i \geq \hat{B}$ ($i = 1, 2$) implies $\tilde{q}_b^i \leq \hat{q}$ (since $\tilde{B}_i = B_R(\tilde{q}_b^i)$, $\hat{B} = B_R(\hat{q})$ and $B_R(q_b)$ decreases with q_b) and thus $B_N^r(\tilde{q}_b^i) \leq \hat{B}(\alpha p) \leq \tilde{B}_i$; it also implies $\tilde{B}_i > \underline{B}$ (since $\hat{B} > \underline{B}$). Likewise, $\tilde{B}_2 \geq \tilde{B}_1$ implies $\tilde{q}_a^2 \leq \tilde{q}_a^1$ (since $\tilde{B}_i = \tilde{B}(\tilde{q}_a^i)$ and $\tilde{B}(q_a)$ decreases with q_a) and thus $B_N^c(\tilde{q}_a^2) = \tilde{B}_2 \geq \tilde{B}_1 = B_N^i(\tilde{q}_a^1) \geq B_N^i(\tilde{q}_a^2)$ (since B_N^i increases with q_a); conversely, $\tilde{B}_1 \geq \tilde{B}_2$ implies $B_N^i(\tilde{q}_a^1) \geq B_N^c(\tilde{q}_a^1)$. We thus have three candidate optimal leniency policies: $(\hat{q}_b, 0)$, $(\tilde{q}_b^1(\alpha, p), \tilde{q}_a^1(\alpha, p))$ and $(\tilde{q}_b^2(\alpha, p), \tilde{q}_a^2(\alpha, p))$.

Straightforward computations moreover show that:

- $\tilde{B}_1(\alpha, p) \geq \hat{B}(\alpha p)$ if and only if

$$p \leq \tilde{p}_1(\alpha) = \frac{\alpha(2 - \delta) - \delta}{2\alpha(1 - \delta)} (\leq 1);$$

- $\tilde{B}_2(\alpha, p) \geq \hat{B}(\alpha p)$ if and only if

$$p \leq \tilde{p}_2(\alpha) = \frac{(1 - \delta)(3\delta - 2)}{(2\delta - 1)(2 - \delta) + 2\alpha\delta(1 - \delta)};$$

- and $\tilde{B}_2(\alpha, p) \geq \tilde{B}_1(\alpha, p)$ if and only if

$$p \leq \tilde{p}(\alpha) = \frac{2(1 - \delta)(2\delta - 1) + \alpha(2\delta^2 - 1) - \alpha^2\delta}{\alpha[(2\delta - 1)(3 - 2\delta) + 4\alpha(1 - \delta)^2]}.$$

Note that $\tilde{p}_2(\alpha) \leq 0$ when $\delta \leq 2/3$, in which case the policy $(\hat{q}, 0)$ always dominates the policy $(\tilde{q}_b^2, \tilde{q}_a^2)$. In that case, the optimal policy is thus either $(\tilde{q}_b^1, \tilde{q}_a^1)$ or $(\hat{q}, 0)$, depending on

whether p is lower or higher than $\tilde{p}_1(\alpha)$. When instead $\delta > 2/3$, then $\tilde{p}_2(\alpha)$, which decreases with α , is positive and satisfies:

$$1 > \tilde{p}_2(0) = \frac{(1-\delta)(3\delta-2)}{(2\delta-1)(2-\delta)} > \tilde{p}_2(1) = \frac{(1-\delta)(3\delta-2)}{(2\delta-1)(2-\delta) + 2\delta(1-\delta)} > 0.$$

In contrast, $\tilde{p}_1(\alpha)$ increases with α and satisfies:

$$\tilde{p}_1(0) = -\infty < \tilde{p}_1(1) = 1.$$

These two thresholds $\tilde{p}_1(\alpha)$ and $\tilde{p}_2(\alpha)$ therefore coincide for some $\alpha_0 \in (0, 1)$, in which by construction they also coincide with $\tilde{p}(\alpha)$. Finally:

$$\begin{aligned} \frac{d\tilde{p}}{d\alpha} = & -\frac{\delta(2\delta-1)(2\delta^2-4\delta+3) + 2(1-\delta)(3\delta-2)}{[(2\delta-1)(3-2\delta) + 4\alpha(1-\delta)]^2} \\ & - \frac{16(1-\delta)^3(2\delta-1)\alpha + 2(1-\delta)(2\delta-1)^2(3-2\delta)}{\alpha^2[(2\delta-1)(3-2\delta) + 4\alpha(1-\delta)]^2}, \end{aligned}$$

it follows that $\tilde{p}(\alpha)$ decreases with α whenever $\delta > \frac{2}{3}$ (since all terms are then negative in the expression of $d\tilde{p}/d\alpha$); we also have:

$$\tilde{p}(1) = \frac{(1-\delta)(2\delta-3)}{(2\delta-1)(3-2\delta) + 4(1-\delta)^2} < 0,$$

which validates the illustration provided by Figure 4. The policy $(\hat{q}, 0)$ is then optimal when $p \geq \max\{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$, otherwise the optimal policy is either $(\tilde{q}_b^1, \tilde{q}_a^2)$ or $(\tilde{q}_b^2, \tilde{q}_a^2)$, depending on whether p is higher or lower than $\tilde{p}(\alpha)$.

Appendix C: Proof of Proposition 5

An increase in p enhances deterrence and thus increases \tilde{B} and \hat{B} , which in turn calls for less leniency both before and after investigation; indeed it is straightforward to check that:

$$\begin{aligned} \frac{\partial \tilde{q}_a^1}{\partial p} &< 0, \quad \frac{\partial \tilde{q}_b^1}{\partial p} < 0; \\ \frac{\partial \tilde{q}_a^2}{\partial p} &< 0, \quad \frac{\partial \tilde{q}_b^2}{\partial p} < 0; \end{aligned}$$

and $\frac{\partial \hat{q}}{\partial p} = \alpha \hat{q}'(\alpha p) < 0$. Similarly, an increase in α also increases \tilde{B} and \hat{B} and both \tilde{q}_b^1 and \tilde{q}_a^1 decrease with α :

$$\begin{aligned} \frac{\partial \tilde{q}_b^1}{\partial \alpha} &= -2 \frac{1-\delta+p}{2(1-\delta)^2 + (2-\delta)} < 0, \\ \frac{\partial \tilde{q}_a^1}{\partial \alpha} &= -\frac{2\delta}{\alpha^2} \frac{1-\delta}{2(1-\delta)^2 + (2-\delta)} < 0. \end{aligned}$$

Moreover, we have

$$\frac{\partial \tilde{q}_b^2(\alpha, p)}{\partial \alpha} = -2(2\delta - 1) \frac{2(1 - \delta)^3(3\delta - 2) + \alpha^2\delta^3 p + 2(2\delta - 1)(1 - \delta)(2 - \delta)(1 - \delta + 2\alpha\delta)p}{(2(2\delta - 1)(1 - \delta)(2 - \delta) + \alpha\delta^2)^2} < 0,$$

and

$$\frac{\partial \tilde{q}_a^2(\alpha, p)}{\partial \alpha} = 2\delta(1 - \delta) \frac{\delta(1 - \delta)(3\delta - 2) - p(2\delta - 1)(2 - \delta)(\delta - 4(1 - \delta)^2)}{[2(2\delta - 1)(1 - \delta)(2 - \delta) + \alpha\delta^2]^2}.$$

The policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ is optimal when $p < \tilde{p}_2(\alpha)$ (which implies $\tilde{p}_2(\alpha) > 0$), in which case we have:

$$\begin{aligned} \frac{\partial \tilde{q}_a^2(\alpha, p)}{\partial \alpha} &\geq 2\delta(1 - \delta) \frac{\delta(1 - \delta)(3\delta - 2) - \tilde{p}_2(\alpha)(2\delta - 1)(2 - \delta)(\delta - 4(1 - \delta)^2)}{[2(2\delta - 1)(1 - \delta)(2 - \delta) + \alpha\delta^2]^2} \\ &= 2\delta(1 - \delta)^2 \tilde{p}_2(\alpha) \frac{2\alpha\delta^2 + 4(2\delta - 1)(1 - \delta)(2 - \delta)}{[2(2\delta - 1)(1 - \delta)(2 - \delta) + \alpha\delta^2]^2} \\ &> 0. \end{aligned}$$

Appendix D: Proof of Proposition 6

We have $\tilde{q}_b^1 \geq \tilde{q}_a^1$ if and only if

$$\alpha[(1 - \alpha)(2 - \delta) + \alpha(1 - p)] \geq \delta(1 - \delta)(1 - \alpha) + \alpha(1 - p)(2 - \delta),$$

which is equivalent to

$$p \geq \hat{p}_1(\alpha) \equiv \frac{(\alpha^2 + \delta(1 - \alpha))(1 - \delta)}{\alpha(2 - \delta - \alpha)}.$$

Similarly $\tilde{q}_b^2 \geq \tilde{q}_a^2$ if and only if

$$\begin{aligned} &2(1 - \delta)(2\delta - 1) + \alpha(3\delta^2 - 3\delta + 1) - (2\delta - 1)(1 - \delta + \alpha\delta)\alpha p \\ &> (1 - \delta)(5\delta - \delta^2 - 2) + \alpha\delta^2 - (2\delta - 1)(2 - \delta)(1 - \delta + \alpha\delta)p, \end{aligned}$$

or

$$p \geq \hat{p}_2(\alpha) \equiv \frac{(1 - \delta)[\delta(1 - \delta) + \alpha(2\delta - 1)]}{(2\delta - 1)(1 - \delta + \alpha\delta)(2 - \delta - \alpha)}.$$

Note that $\hat{p}_1(1) = 1$ and

$$\hat{p}_2(0) = \frac{(1 - \delta)\delta}{(2\delta - 1)(2 - \delta)} < 1.$$

In addition, for $p = \hat{p}_1(\alpha) = \hat{p}_2(\alpha)$, we have $\tilde{q}_b^1 = \tilde{q}_a^1 = \tilde{q}^1$ and $\tilde{q}_b^2 = \tilde{q}_a^2 = \tilde{q}^2$, where \tilde{q}^1 and \tilde{q}^2 moreover satisfy $\varphi(\tilde{q}^1) = \varphi(\tilde{q}^2) = 0$, where:

$$\varphi(q) \equiv B_I^r(q, q; \alpha) - B_R(q) = \frac{\alpha\left(1 - \frac{q}{2}\right) - (1 - \delta)(1 - q)}{2\delta - 1} F - \frac{\delta\left(1 - \frac{q}{2}\right) F}{2\delta - 1}$$

is monotonic ($\varphi' = (1 - \frac{\delta + \alpha}{2}) / (2\delta - 1) > 0$), it follows that $\tilde{q}^1 = \tilde{q}^2$; in other words, $\hat{p}_1(\alpha)$ and $\hat{p}_2(\alpha)$ intersect each other at the same point (α_0, p_0) as $\tilde{p}(\alpha)$, $\tilde{p}_1(\alpha)$ and $\tilde{p}_2(\alpha)$.

References

- Aubert, C., P. Rey and W. Kovacic (2006), "The Impact of Leniency and Whistleblowing Program on Cartels", *International Journal of Industrial Organization*, 24:1241-1266.
- European Commission (2006), "Guidelines on the method of setting fines imposed pursuant to Article 23(2)(a) of Regulation No 1/2003", *Official Journal of the European Union*, C 210, 49:2-5.
- Frezal, S. (2006), "On Optimal Cartel Deterrence Policies", *International Journal of Industrial Organization*, 24(6):1231-1240.
- Hammond, S. (2005), "An update of the Antitrust Division's Criminal Enforcement Program", speech before the ABA Section of antitrust law, available at http://www.usdoj.gov/atr/public/speeches/speech_criminal.htm
- Harrington, J. (2008), "Optimal Corporate Leniency Programs", *Journal of Industrial Economics*, 56(2):215-246.
- Harrington, J., and M. Chang (2009), "Modelling the Birth and Death of Cartels with an Application to Evaluating Antitrust Policy", *Journal of European Economic Association*, 7:1400-1435.
- Hinlopen, J., and A. Soetevent (2008), "From Overt to Tacit Collusion: Experimental Evidence on the Adverse Effects of Corporate Leniency Programs", *working paper*.
- Motta, M., and M. Polo (2003), "Leniency Programs and Cartel Prosecution", *International Journal of Industrial Organization* 21:347-379.
- Polinsky, M, and S. Shavell (2000), "The Economic Theory of Public Enforcement of Law", *Journal of Economic Literature*, March: 45-76
- Rey, P. (2003), "Toward a theory of Competition Policy", in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, M. Dewatripont, L. P. Hansen, S. J. Turnovsky eds, Cambridge University Press.
- Spagnolo, G (2000), "Optimal Leniency Programs", *working paper*.
- Spagnolo, G. (2004), "*Divide et Impera*: Optimal Leniency Programmes", CEPR Discussion Paper N°4840, available at: www.cepr.org/pubs/dps/DP4840.asp.
- Wils, W. (2008), *Efficiency and Justice in European Antitrust Enforcement*, Hart Publishing.