On the Design of Leniency Programs

by

Zhijun Chen
ESRC Centre for Competition Policy University of East Anglia, and
School of Economics, Zhejiang University
&
Patrick Rey
University of Toulouse (IDEI, GREMAQ and IUF)

CCP Working Paper 08-18

The support of the Economic and Social Research Council is also gratefully acknowledged.

ISSN 1745-9648

On the Design of Leniency Programs[1]

Zhijun Chen[2]
The ESRC Centre for Competition Policy, University of East Anglia,
and School of Economics, Zhejiang University
E-mail: chenzj1219@gmail.com

and

Patrick Rey
University of Toulouse (IDEI, GREMAQ and IUF)
E-mail: prey@cict.fr

Version: April 2007

Abstract

We develop a simple framework for analyzing the optimal design of leniency programs, which allow cartel members to denounce their collusive agreements. We highlight a basic trade-off between two opposite forces: leniency can destabilize cartels, by encouraging firms to report and bring evidence to the antitrust authority, but it can also reduce the expected penalties that cartel members face. We characterize the optimal leniency rates, both before any investigation and once an investigation is opened, and show that these two leniency opportunities are particularly useful when random investigations are unfrequent and/or unlikely to succeed in the absence of self-reporting; we also compare the effectiveness of alternative rules for late informants and repeated offenders.

Key Words: Leniency Program, Anti-trust Law enforcement

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 after an investigation is 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.[3]

[1] We are grateful to Bruno Jullien and Michele Polo for their comments.

[2] The correspondence author

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

In spite of a great success in practice,[4] 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 program. This paper proposes a basic framework for such a normative approach. Taking for given several features of antitrust enforcement, such as the probability that a cartel would be investigated and then successfully prosecuted in the absence of reporting, it looks for the optimal amnesty rates, both before and after an investigation is started.

Gathering evidence is a challenge for antitrust agencies, which is unlikely to be successful in the absence of any tip. Thus, inducing those who are engaged in cartel activity to report it and bring adequate evidence may provide an effective enforcement tool; and granting amnesty to cartel members indeed encourages them to report this activity, and can thus contribute in this way to destabilize collusion. However, reducing the expected fine that firms have to pay if the cartel is uncovered may also make cartels more profitable and, by the same token, more robust. As we will see, the trade-off between these two conflicting forces determines the optimal level of leniency.

The main contribution of this paper is of two-fold. First, solving the trade-off just mentioned allows us to relate the optimal leniency rates (the "carrot") to the effectiveness of random investigations (the "stick"). Whenever random audits are not very effective in uncovering cartels, it is desirable to offer some amnesty, at least in the absence of any ongoing investigation; whether amnesty remains desirable once an investigation is underway depends however on both the frequency of random investigations and on the likely success of these investigations: optimal leniency rates increase as random investigations become less successful, and when success is quite unlikely, it is always optimal to offer leniency programs both pre-and post investigation, however frequent these investigations are. 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[5], when investigations are infrequent but likely to succeed once they are launched; when instead investigations are frequent but unlikely to succeed, it can however be desirable to offer more amnesty once an investigation is underway, in order to make these investigations more effective.

Second, the comparison of different variants provides several policy implications. In particular, it validates the "first informant" and "post investigation amnesty" rules introduced in the 1993 version of the US leniency program. It also shows that offering no leniency for repeated offenders may not be effective in fighting collusion, which calls for a cautious use of heavy sticks.

This paper builds on the recent literature on leniency programs. In particular, Motta and Polo (2003) analyze the impact of leniency on collusion in a framework where the antitrust agency can also launch random investigations that sometimes lead to successful prosecution. They study the most effective way to allocate antitrust resources between preliminary investigation and prosecution (the agency has a fixed budget, which it can spend on conducting more investigations or in making each investigation more likely to lead to successful prosecution); they moreover show that it can be useful to grant leniency once an investigation is underway, so as to encourage cartel members to cooperate with the antitrust authorities once a cartel is prosecuted. In contrast, we take here the likelihood of investigations and successful prosecution as given, and characterize the optimal degree of leniency; we

[4] See Hammond (2005).

[5] For 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.

also show that both pre-and post-investigation leniency can be helpful to prevent the formation of some cartels.

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, and should only reward the first informant; he also notes that, while leniency can contribute to destabilize cartels, it can also be "exploited" by the firms, which determines a maximal level of leniency. We build on his analysis by introducing heterogeneity in the stakes of collusion across industries and distinguishing pre- and post-investigation leniency. Aubert, Rey and Kovacic (2005) compare the impact of reduced fines and positive rewards and argue that rewarding individuals can deter collusion in a more effective way. Moreover, they discuss possible adverse effects of whistleblowing programs on firms' behavior and incentives to innovate and cooperate. Harrington (2005) 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 (2006) studies the impact of leniency programs on cartel desistance as well as cartel deterrence. He develops a nice 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 an elegant 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.

2. THE MODEL

2.1. The collusion game

In each industry, two identical firms play an infinitely repeated game where, in each period, they can choose to form a hard-core cartel before interacting on the product market. All firms have the same discount rate $\delta \in (0,1)$ and maximize the expected discounted sum of their profits. In each period, each firm chooses 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.

If we consider for example a standard Bertrand duopoly, in which the two firms produce perfect substitutes with the same constant unit cost c and compete in prices for a demand $D(p)$, the profits under static price competition are indeed zero while

the maximal benefit from collusion corresponds to half of the monopoly profits $\left(B = \pi^M/2 = \max_p (p-c) D(p)/2\right)$; deviating from such collusion then 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.[6]

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

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

that is, if

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

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 in "as many industries as possible", it is useful to introduce some heterogeneity among industries. For the sake of presentation we will assume that δ remains constant across industries, which however 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 that the antitrust authority can find out if it investigates the industry; however, due to budget and resource limitations, this happens only with some probability ρ $(0 < \rho < 1)$; 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. To keep the analysis simple, we assume that the evidence of collusion 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 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 $(1-q)F$, while the other

[6] For 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.

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.[7] 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[8]

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

or equivalently

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

Collusion is therefore sustainable only when its stake is sufficiently large; otherwise, each firm would find it profitable to deviate: the short-term gain from a deviation, equal to B, is then higher than the cost of foregone future collusion, equal to 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 in case of detection F increase.

Remark: Stakes versus fines. We assume here that the fine F is independent from the stakes from collusion. In practice, fines are set according to judicial principles, which vary across countries but are often related, directly or indirectly, to the nature and importance of the anticompetitive behavior, and thus, possibly, to the stakes from collusion.[9] This link between fines and the stake from collusion is however often imperfect, as the level of the fines is subject to exogenous caps,[10] and also driven by other considerations.[11]

[7] In particular, firms do not observe deviations before the end of the period, where evidence becomes obsolete; otherwise, they could threaten to punish a deviation by denouncing the cartel – which is self-sustaining here: each firm is willing to denounce if it anticipates that the other does. As discussed below, allowing for such retaliation possibilities would not qualitatively affect the analysis, although it would tend to make leniency furthermore effective in deterring collusion, by allowing deviators to avoid paying the full fine.

[8] For 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).

[9] Consider for example the previous Bertrand duopoly example with a linear demand $D(p) = d - p$, so that industries essentially vary in market size, as measured by $d - c$; collusion then reduces welfare by $\Delta W = (d-c)^2/8 = \pi^M/2 = B$ and consumer surplus by $\Delta CS = \Delta W + \pi^M = 3B$: a fine proportional to either the harm to consumers or society would thus be also proportional to the stake from collusion.

[10] In EU proceedings, fines cannot exceed 10% of the turnover of the firms. In the US, fines could not exceed $10 million until 2004, where the ceiling was pushed up to $100 million.

[11] The EU guidelines, for example, consider the following steps – see European Commision (2006). The Commission determines a first amount, based in particular (but not only) 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." Aggravating circumstances include "where an undertaking continues or repeats

3. OPTIMAL LENIENCY

3.1. Amnesty for the first informant

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 broadens the scope 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 a firm that deviates can now moreover denounce the cartel in order to benefit from leniency, and it will indeed have an incentive to do so if the amnesty rate reduces the expected fine that it faces, i.e., if:

$$q > \underline{q} \equiv 1 - \rho > 0. \tag{3}$$

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:

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

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

Alternative collusive strategies.

Firms may however try 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 may sound far-fetched, since the cartel would then be systematically denounced and yet go on forever, in practice, 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 again apply 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 extend below our framework to allow for more intense scrutiny after denunciations and show that the qualitative analysis remains the same (see subsection 3.4); we also consider later on the possibility that firms 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

the same or a similar infringement", "a refusal to cooperate with or obstruction of the Commission in carrying out its investigations", or a "role of leader in, or instigator of, the infringement". 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."

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:

$$V_R(q) \geq 2B - \left(1 - \frac{q}{2}\right) F, \tag{4}$$

that is, whenever

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

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 $q > \bar{q}$.

Optimal amnesty rates.

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

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

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. To fix ideas, we will assume that collusive behavior results in a deadweight loss of social welfare $D(B)$ and that the antitrust authority aims to minimize the total welfare loss from collusion:

$$\min_q L = \int_{\min\{B_N(q), B_R(q)\}}^{+\infty} \frac{D(B)}{1 - \delta} dG(B),$$

where $G(B)$ denotes the cumulative distribution function of the stakes from collusion. Minimizing the welfare loss from collusion boils down here to deter as many cartels as possible (later on, desistance will also become an issue); the amnesty rate q should therefore maximize the deterrence threshold

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

[12] As usual, the two firms should behave symmetrically in order to maximize the scope for collusion, and colluding in every period maximizes the value or 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.

which appears in bold in Figure 1.

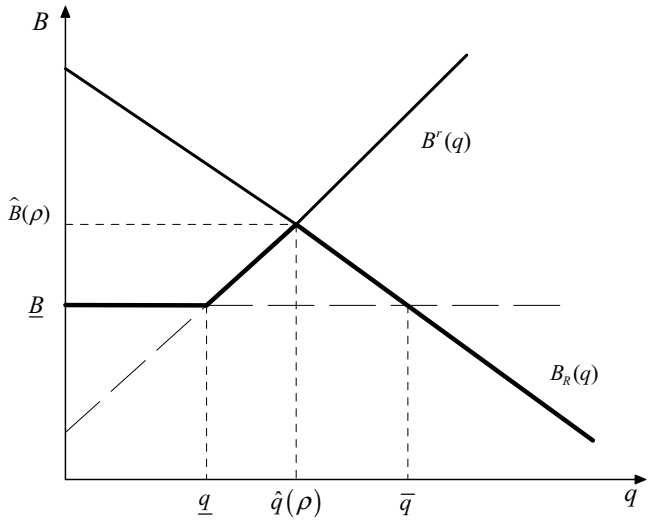

Figure 1: Optimal amnesty rate and deterrence threshold.

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^r (i.e., 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:

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

which is achieved for

$$q = \hat{q}(\rho) \equiv \frac{1-\rho}{1-\frac{\delta}{2}}. \tag{5}$$

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

$$\hat{B}(\rho) \equiv \frac{\delta(1 - \delta + \rho)}{(2\delta - 1)(2 - \delta)} F, \tag{6}$$

and is indeed higher than $\underline{B}$.

The following proposition summarizes the analysis:

PROPOSITION 1. *The optimal amnesty rate lies between $\underline{q} > 0$ and $\bar{q} > \underline{q}$ and is determined so as to deter normal collusion, without encouraging collusion with reporting: it is characterized by* (5) *and increases as the probability of prosecution, ρ, decreases.*

The above analysis highlights a "stick and carrot" logic: it is useful to complement the stick (the probability ρ of investigations) with a carrot (the amnesty rate q), and all the more so as the stick becomes weaker ($\hat{q}$ increases when ρ decreases). 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 amnesty rate thus never exceeds $\bar{q}$, in order to keep "collude and report strategies" less profitable,[13] and thus less robust, than normal collusion. The optimal leniency rate $\hat{q}$ reflects precisely the trade-off between destabilizing normal collusion and not encouraging alternative strategies and is such that, in 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: $B^r(\hat{q}) = B_R(\hat{q}) = \hat{B}(\rho)$.

The same trade-off drives the impact of random audits on the optimal amnesty rate: increasing the number of investigations or their performance destabilizes normal collusion and thus 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 ($B^r(q;\rho)$ 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.

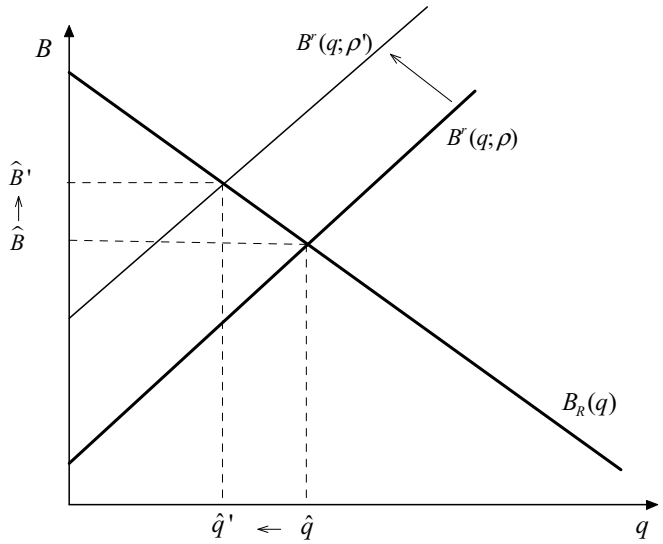

Figure 2: Impact of an increase in ρ on the optimal amnesty rate.

Remark: Observable deviations. When firms can detect deviations before the evidence of collusion becomes obsolete, they could "punish" deviations by denouncing the cartel (as already observed, this is self-sustainable here, since each firm is willing to expose the cartel when it expects the rival to do it anyway). In that case,

[13] Firms would therefore rather favor normal collusion, which is moreover weakly easier to sustain for the optimal amnesty rate: when $\rho \geq \hat{\rho}$, normal collusion is sustainable in any industry $B \geq \underline{B}$ for any rate $q \leq \bar{q}$, while offering no leniency maximally deters "collude and report" strategies; when $\rho < \hat{\rho}$ and $q = \hat{q}$, both types of collusion are sustainable whenever any one is.

the antitrust policy can theoretically be used to enhance retaliation and reinforce collusion: absent leniency, collusion is sustainable whenever

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

or

$$B \geq \underline{B}' = \frac{\rho - (1 - \delta)}{2\delta - 1} F;$$

this new threshold is lower than $\underline{B}$, and is even negative (implying that any industry could collude, whatever the stakes of collusion) when $\rho < 1 - \delta$. In such a context, leniency may become even more appealing, since it gives deviators a way to avoid paying the fine; anticipating that their rival will expose the cartel, a deviator will then always "run to the courthouse" when it plans to deviate (and it is reasonable to assume that, as the one responsible for the timing of the deviation, it will indeed be able to beat its rival in this race), even if the amnesty rate is small – even if $q < \underline{q}$, since a small reduction being always preferable to no reduction). It does not really affect the outcome of the analysis in our current framework, since the optimal amnesty rate is always higher than the minimal threshold $\underline{q}$, but it could do so in more general contexts, by providing additional motivation for leniency.

3.2. Amnesty for additional informants

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 "report and collude" more attractive and therefore more robust, which reduces the effectiveness of the leniency programme.

To see this, let q denote as before the amnesty rate when there is a single informant and suppose that, if both firms apply for leniency, each faces an expected fine equal to $(1 - \tau q)$, for some $\tau \in [1/2, 1]$: $\tau = 1/2$ corresponds to the "first informant only" rule, while $\tau = 1$ corresponds to the case when all informants benefit from the same leniency q. Then:

- the deterrence threshold for normal collusion remains unchanged: firms can sustain such collusion as long as $B \geq B_N(q) = \max\{\underline{B}, B^r(q)\}$; but
- the deterrence threshold for "report and collude" decreases; the value of such collusion becomes

  $$V_R(q;\tau) \equiv \frac{B - (1 - \tau q) F}{1 - \delta}$$

  and collusion is therefore sustainable as long as $V_R(q;\tau) \geq 2B - (1 - \tau q) F$, that is:

  $$B \geq B_R(q;\tau) \equiv \frac{\delta (1 - \tau q) F}{2\delta - 1}.$$

The threshold $B_R(q;\tau)$ obviously decreases with τ: granting leniency to additional informants both encourages and facilitates "collude and report" strategies. As a result, the amnesty rate $\hat{q}(\rho;\tau)$, characterized by

$$B^r(q) = B_R(q;\tau),$$

becomes smaller and, relatedly, the scope for leniency is reduced (i.e., the threshold $\hat{\rho}$ characterized in the previous proposition decreases with τ). In the extreme case where all informants would benefit from the same leniency, it is actually optimal to offer no leniency:

PROPOSITION 2. *Offering leniency to additional informants:*

- *makes the leniency program less effective in fighting collusion,*
- *reduces the optimal amnesty rate; in particular, if all informants must benefit from the same leniency, then it is optimal to offer no leniency.*

Proof. See Appendix A. ∎

The leniency program thus performs less well in the absence of "first-informant-wins" rule; in particular, an amnesty program that would treat all informants alike achieves worse results than antitrust enforcement without leniency. This result may explain why the original version of the US leniency program did not contribute much to defeat cartels before the 1993 revision.[14]

3.3. Leniency for repeated offenders

The above analysis supposes that firms could in principle report a cartel, benefit from leniency, and yet keep colluding in the future. This is not inconsistent with the casual observation that the same firms and the same industries (e.g., the cement industry) are regular "customers" of cartel offices. However, one would expect that in practice, once a cartel has been exposed, the industry will be kept under closer scrutiny for at least some time;[15] in the same vein, fines can be larger for repeated offenders, which further contributes to reduce the appeal of "collude and report" strategies. In addition, in many countries like the USA and the EU, amnesty is never offered to a repeated offender. This prevents cartels from adopting "collude and report" strategies, but may also lead to other forms of collusion, such as "report once and never after that". 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 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 the cartel from colluding and reporting systematically. 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 further report; collusion is then sustainable as long as it resists deviations in the product market, i.e. whenever:

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

[14] This also confirms previous insights along the same line; for other formal explorations, see for example Spagnolo (2004) or Harrington (2005).

[15] More generally, we have restricted attention here to "stationary" antitrust policies. Frezal (2006) however 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.

In the first period of this collusion path, firms report the cartel anyway; collusion is thus again sustainable as long as it simply resists 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 brings a trade-off: it prevents cartels from colluding and reporting systematically, but also 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 in the first period. As a result, "collude and report once" is sustainable whenever normal collusion is sustainable absent leniency; in other words, ruling out leniency for repeated offenders renders the leniency program completely ineffective:

PROPOSITION 3. *Restricting leniency to first-time offenders makes it ineffective in deterring collusion.*

The analysis suggests that the antitrust authority should be cautious when refusing to grant leniency to repeated offenders, unless it can deter exposed cartels from returning to collusion, e.g., by intensified monitoring; we now further explore this latter possibility.

3.4. Intensified scrutiny for exposed cartels

The scope for collusion can be reduced if the antitrust authority can monitor exposed cartels and prevent them from colluding again. Suppose for example that once a cartel is revealed, either by self-reporting or by random investigations, firms remain under close scrutiny – and thus cannot collude – for T periods. This reduces the value of "collude and report" strategies to

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

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

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

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

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

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

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

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

and

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

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

Similarly, "collude and report" strategies are deterred if

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

Obviously, an increase in the duration T of monitoring reduces the profitability of (any form of) collusion and thus also makes it more fragile. In particular, if $\gamma(T) < 1/2$, no collusion can be sustained. However, this monitoring hurts the alternative forms of collusion, where cartels are denounced and thus exposed to scrutiny, even more than it harms normal collusion. This, in turn, makes higher amnesty rates more appealing; in particular, if

$$\gamma(T) > \frac{1}{2} > \beta(T),$$

then firms cannot sustain "collude and report" and can only hope to sustain normal collusion; in that case, any increase in the amnesty rate further contributes to deter normal collusion, without having no countervailing perverse impact on alternative forms of collusion.

Consider now the case where $\beta(T) > 1/2$, so that firms can "collude and report" for large enough collusion benefits. To prevent as many cartels as possible, the optimal amnesty rate q should as before maximize

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

The threshold $B_R(q;\beta)$ decreases again as q increases; excessive leniency does not have perverse effects (i.e., $B_R(q;T) \geq \underline{B}(T)$) as long as:

$$q \leq \bar{q}(T) \equiv 2(1-\rho) + \frac{2(\gamma - \beta)\rho}{(2\gamma - 1)\beta},$$

where $\bar{q}(.)$ increases with T: keeping a close eye on exposed cartels discourages "collude and report" strategies and makes them also less robust to deviations, which makes it possible to offer higher leniency rates without triggering perverse effects.

The optimal amnesty rate lies again between $\underline{q}$ and $\bar{q}(T)$ and is now determined by

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

which yields

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

and

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

Since increasing the duration T of scrutiny reduces β and the value of collusion, both $B^r(q;T)$ and $B_R(q;T)$, and thus the optimal deterrence threshold $\hat{B}(T)$,

increases (see Figure 3 below). 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:

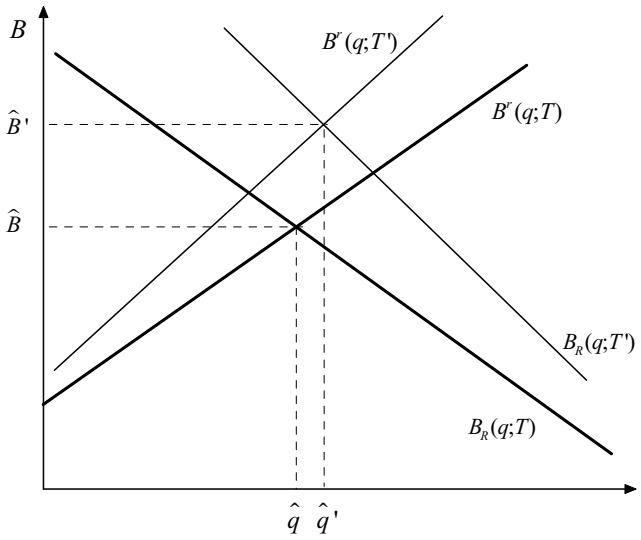

Figure 3: Optimal leniency under intensified scrutiny.

PROPOSITION 4. *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 B. ∎

Remark: repeated offenders. 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. As before, prohibiting leniency for repeated offenders does not affect normal collusion, but facilitates "collude and report once" strategies, although less so than in the absence of closer scrutiny for exposed cartels. "Collude and report once" is sustainable when:

$$V_O(q,T) \equiv B - \left(1 - \frac{q}{2}\right) F + \frac{\beta(B - \rho F)}{1-\gamma} \geq 2B - (1 - \frac{q}{2})F,$$

which now boils down to:

$$B \geq \underline{B}_O(T) \equiv \frac{\beta\rho F}{\beta + \gamma - 1}.$$

"Collude and report once" is however less profitable than normal collusion (because of the closer scrutiny that follows the report) and is thus more difficult to sustain than normal collusion without leniency:

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

However, "Collude and report once", which reverts to normal collusion after the first period, is however more profitable and therefore easier to sustain than "Collude

and report systematically" when leniency is extended to repeated offenders:

$$\underline{B}_O(T) > \hat{B}(T) = B_R(\hat{q}(T); T).$$

As a result, prohibiting leniency for repeated offenders:

- increases the scope for collusion, since "Collude and report once" becomes sustainable in industries with $\hat{B}(T) > B \geq \underline{B}_O(T)$;
- reduces the optimal amnesty rate, which is now such that $B^r(q;T) = \underline{B}_O(T) < \hat{B}(T) = B^r(\hat{q}(T;T))$.

3.5. Deterrence versus desistance

Until now, we assumed that the antitrust authority focuses on deterring collusion as much as possible which, as noted, appears justified when exposed cartel members can keep colluding; however, when discovered cartels are destabilized for some time, e.g. because of intensified scrutiny, leniency programs can also contribute to foster this desistance effect.

Consider the context just analyzed, where exposed cartels are subject to close scrutiny, preventing them from colluding, for T periods. Normal collusion is uncovered with probability ρ, in which case it is monitored for T periods; the associated welfare loss is thus (for simplicity we will drop the argument throughout this section):

$$L_N(B) = D(B) + \gamma L_N(B) = \frac{D(B)}{1-\gamma}.$$

In contrast, when firms pursue "collude and report" strategies, collusion occurs only every $T+1$ periods, and the welfare loss will thus be equal to

$$L_R(B) = \frac{D(B)}{1-\beta}.$$

Obviously, "collude and report" strategies generate less welfare loss than normal collusion:[16] $\gamma > \beta$ implies

$$L_R(B) < L_N(B).$$

It would therefore be welfare-improving to induce cartels to favor "collude and report" strategies over normal collusion, which calls for higher amnesty rates. If both types of collusion are sustainable, cartel members prefer to stick to normal collusion if

$$V_N \geq V_R,$$

that is, if

$$B \geq B^N(q;T) \equiv \frac{(1-\gamma)\frac{q}{2} - (1-\delta)(1-\rho)}{\gamma - \beta} F.$$

This threshold is lower than $B^r(q)$ for $q = \hat{q}$ (by construction, normal collusion is more profitable whenever it is sustainable),[17] but also increases with q, since

[16] This however neglects the policy costs associated with intensified monitoring, which are incurred more often when firms systematically report.

[17] For $q = \hat{q}$ and $B = \hat{B}$, firms are indifferent between sticking to collusion or reporting it, since $B = B^r(q)$, and between sticking to "collude or report" or deviating from it, since $B = B_R(q)$; therefore (using $\hat{q} > \underline{q} = 1 - \rho$):

$$V_R = 2B - \rho F < 2B - (1-q)F = V_N.$$

"collude and report" becomes more attractive, and its slope is actually higher than that of $B^r(q)$.[18] Therefore, for $q < \hat{q}$, normal collusion is more profitable whenever it is sustainable; assuming that firms coordinate on the most profitable collusion, reducing q below $\hat{q}$ thus reduces deterrence (more industries can collude in a standard fashion) without improving desistance (cartel members favor normal collusion over alternatives forms). However, increasing q above $\hat{q}$ involves a trade-off between deterrence and desistance. More precisely, letting q^N denote the amnesty rate for which:

$$B^r\left(q^N\right) = B^N\left(q^N\right),$$

two cases can be distinguished (see Figure 4):

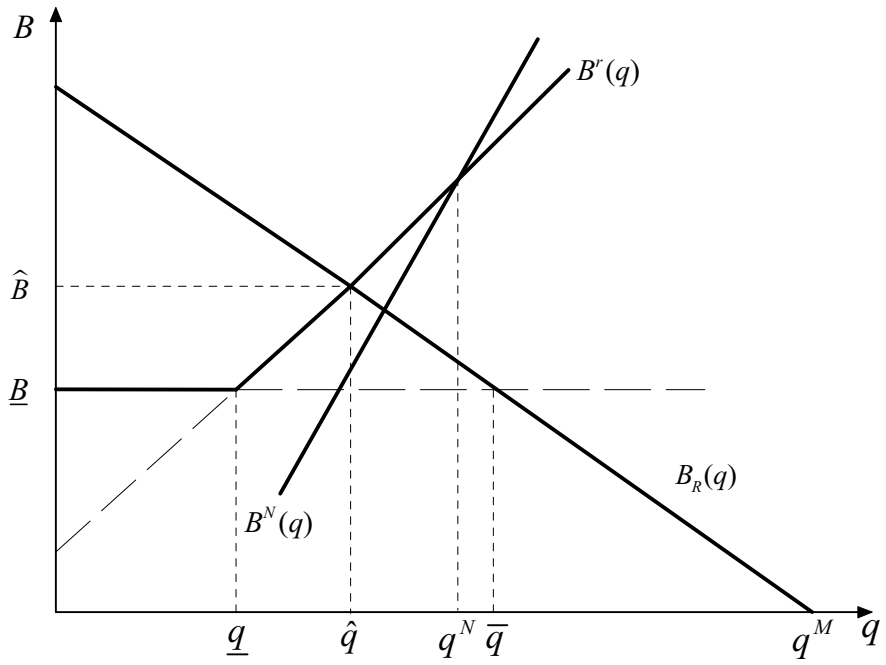

Figure 4: Deterrence versus desistance

1) As long as $(\hat{q}(T) \leq) q \leq q^N(T)$, cartels still favor normal collusion whenever it is sustainable, i.e., when $B \geq B^r$, but can only "collude and report" when $B_R \leq B < B^r$. The total welfare loss is thus equal to

$$\begin{aligned} L(q) &= \int_{B_R(q)}^{B^r(q)} L_R(B)\, dG(B) + \int_{B^r(q)}^{+\infty} L_N(B)\, dG(B) \\ &= \int_{B_R(q)}^{+\infty} L_R(B)\, dG(B) + \int_{B^r(q)}^{+\infty} \left(L_N(B) - L_R(B)\right) dG(B). \end{aligned}$$

Increasing q thus reduces deterrence but enhances desistance. Indeed, industries with $B \geq B_R(q)$ collude but, among these, those with $B_R(q) \leq B \leq B^r(q)$ are exposed with probability 1 while those with $B \geq B^r(q)$ are exposed only with probability ρ. Increasing the amnesty rate by dq decreases deterrence: it increases the number of cartels by $g(B_R) B'_R dq$, causing a further loss of welfare

[18] Indeed, $\beta > 1/2$ implies $\partial B^r/\partial q = (1-\gamma)/(2\gamma - 1) < \partial B^N/\partial q = (1-\gamma)/(2\gamma - 2\beta)$.

$L_R(B_R)\, g(B_R)\, B_R' dq$, but enhances desistance by reducing the number of cartels who resort to normal collusion by $g(B^r)(B^r)' dq$, which *reduces* the deadweight-loss by $(L_N(B) - L_R(B))\, g(B^r)(B^r)' dq$.

2) When $q > q^N(T)$, in addition to the previous insights, cartels with $B^r(q) \leq B \leq B^N(q)$ also prefer to "collude and report" over normal collusion; the total welfare loss is then given by

$$L(q) = \int_{B_R(q)}^{+\infty} L_R(B)\, dG(B) + \int_{B^N(q)}^{+\infty} (L_N(B) - L_R(B))\, dG(B).$$

A similar trade-off again arises: increasing q induces more cartels to "collude and report", which leads to an additional loss of welfare, but leads some already colluding industries to favor alternative, "reporting" strategies rather than normal collusion, which limits the deadweight-loss. In this case, however, the impact on desistance is – other things being equal – more important, as the slope of B^N is steeper than that of B^r.[19]

We explore in Appendix C the particular case where the collusion stake B is uniformly distributed over a range $\left[0, \bar{B}\right]$, where $\hat{B} < \bar{B} < \max\left\{B^r(2), B^N(2)\right\}$, so as to ensure that the optimal amnesty rate lies between $\hat{q}$ (for which $B_R = B^r$) and $q^M = 2$ (for which $B_R = 0$); when the deadweight-loss is moreover constant across industries ($D(B) = D$), it is optimal to focus on deterrence and thus to set the amnesty rate to $q = \hat{q}$, in order to prevent the formation of cartels in as many industries as possible. When instead the deadweight-loss is positively related to the stakes from collusion ($D'(B) > 0$), a trade-off arises between deterring "small", not-so-harmful cartels, and desisting "big", harmful cartels, which can tilt the balance in favor of desistance and, thus, of higher amnesty rates. And indeed, we provide an example where the optimal leniency policy consists in focusing on desistance and, through a high enough amnesty rate, eliminates normal collusion in all industries. The trade-off between deterrence and desistance is also obviously affected by the distribution of cartel characteristics. For example, if collusion benefits are all higher than $\hat{B}$, all industries can sustain some form of collusion anyway; it is then optimal to focus on desistance and to set the amnesty rate as high as possible, since increasing q beyond $\hat{q}$ enhances desistance (since B^r and B^N both increase with q) without any adverse impact on deterrence.

4. AMNESTY BEFORE AND AFTER INVESTIGATIONS

So far we simply assumed that cartels could be detected with some probability ρ. We now refine the analysis by distinguishing the probability that the antitrust agency launches an investigation from the probability of "success" of such an investigation. More precisely, we drop for simplicity any close scrutiny for exposed cartels, and thus come back to the initial framework in that respect, but suppose that:

- the antitrust authority can launch an investigation with some probability α, where $0 < \alpha < 1$;
- when an investigation is launched, in the absence of reporting by cartel members it succeeds in uncovering the cartel with probability p, where $0 < p < 1$.

[19] Note that $B_R(q)$ becomes negative for $q > q^M = 2$; any further increase in q then keeps enhancing desistance, without any additional adverse impact on deterrence.

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

When the antitrust authority launches an investigation, it can do so openly or try to keep it secret. We first consider the latter possibility (secret investigations), before turning to the case where cartel members are alerted whenever an investigation gets started (open investigations).

4.1. Secret investigations

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[20] $\hat{q}(\alpha p)$ characterized by Proposition 1; 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

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

4.2. Open investigations

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 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.

[20] The amnesty rate might differ when an investigation is already underway; in that case, the relevant amnesty rate is the expected one, $q = \alpha q_a + (1 - \alpha) q_b$.

4.3. Open or secret investigations?

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.

4.4. Optimal amnesty rates

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 consits in reporting only After an investigation is launched. We now characterize the conditions under which firms can sustain these forms of collusion.

Normal collusion (N). The value of normal collusion is now equal to

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

To be sustainable, this collusion must resist three types of defection, which we successively consider: deviating in the product market and reporting at Stage 4 in case of investigation, deviating and reporting at Stage 2 before there may be an investigation, and deviating without reporting.

Cartel members have no incentives to defect and report once an investigation is underway if:

$$B - pF + \delta V_N \geq 2B - (1 - q_a) F,$$

or equivalently:

$$B \geq B_N^a(q_a) \equiv \frac{\delta \alpha p + (1 - \delta)(p - (1 - q_a))}{2\delta - 1} F. \tag{7}$$

Second, deviating and reporting at stage 2 is not attractive if:

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

or:

$$B \geq B_N^b(q_b) \equiv \frac{\alpha p - (1 - \delta)(1 - q_b)}{2\delta - 1} F. \tag{8}$$

Last, deviating in the product market at stage 1 is not profitable (in which case it can be check that it is not profitable to deviate when an investigation is already underway)[21] if:

$$V_N = \frac{B - \alpha pF}{1-\delta} \geq 2B - \alpha pF,$$

that is, when:

$$B \geq B_N^d \equiv \frac{\delta \alpha pF}{2\delta - 1}. \tag{9}$$

Hence, normal collusion is sustainable if and only if:

$$B \geq B_N(q_b, q_a) \equiv \max\left\{B_N^d, B_N^b(q_b), B_N^a(q_a)\right\}. \tag{10}$$

Collude and report After an investigation is launched (A). Reporting once an investigation is underway is self-sustainable at stage 4, irrespective of whether a firm deviates in the product market or not: since the others will report, reporting is profitable since it reduces the expected fine by $q_a/2$. To be sustainable, "collude and report After" must therefore resist only two types of deviations: reporting before an investigation, and deviating in the product market without reporting.

Firms have no incentives to deviate and report before an investigation may be launched if

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

that is:

$$B \geq B_A^b(q_b, q_a) \equiv \frac{\alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)(1-q_b)}{2\delta - 1}F. \tag{11}$$

Similarly, deviating without reporting is not profitable if:

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

or:

$$B \geq B_A^d(q_a) \equiv \frac{\delta\alpha\left(1 - \frac{q_a}{2}\right)F}{2\delta - 1}F. \tag{12}$$

Hence, this collusion is sustainable if and only if:

$$B \geq B_A(q_b, q_a) \equiv \max\left\{B_A^d(q_a), B_A^b(q_b, q_a)\right\}. \tag{13}$$

Collude and report Before an investigation is launched (B). This strategy is self-sustainable at stages 2 and 4, since it is again a best response to report when the others will report anyway. This strategy is therefore sustainable when it resists deviations in the product market:

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

[21] This is the case when

$$B - pF + \delta V_N \geq 2B - pF,$$

which boils down to $\delta V_N \geq B$ or $B \geq B_N^d$.

that is, when:

$$B \geq B_B(q_b) \equiv \frac{\delta\left(1-\frac{q_b}{2}\right)F}{2\delta-1}. \tag{14}$$

Optimal lenciency.

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

$$B(q_b, q_a) \equiv \min\left\{B_N(q_b, q_a), B_A(q_b, q_a), B_B(q_b)\right\}.$$

As already noted, it is still possible to deter collusion in industries $B < \hat{B}$ by refusing leniency once an investigation is underway ($q_a = 0$) and setting the pre-investigation amnesty rate to $\hat{q}_b \equiv \hat{q}(\alpha p)$. Offering some leniency once an investigation is already ongoing ($q_a > 0$) however provides another way to destabilize collusion and, since $B_N^a(q_a)$ increases with q_a, this alternative way is moreover more effective in fighting normal collusion when q_a is large enough. This however encourages an additional form of collusion: $B_A(q_b, q_a) = \max\left\{B_A^d(q_a), B_A^b(q_b, q_a)\right\}$ decreases as q_a increases, since both B_A^b and B_A^d do so, which limits the usefulness of post-investigation leniency. In particular, it is never optimal to rely solely on post-investigation leniency. To see this, suppose that the antitrust authority:

- relies on post-investigation amnesty to deter normal collusion, i.e., $B_N(q_b, q_a) = B_N^a(q_a) > B_N^d, B_N^b(q_b)$;
- and offers little leniency pre-investigation:

$$q_b \leq 1-\alpha\left(1-\frac{q_a}{2}\right),$$

implying:

$$B_A(q_b, q_a) = B_A^d(q_a) = \frac{\alpha\left(1-\frac{q_a}{2}\right)}{2\delta-1} > B_A^b(q_b, q_a).$$

Then, increasing the pre-investigation amnesty rate to $q_b' > 0$ such that

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

yields:

$$B_A(q_b', q_a) = B_A^b(q_b', q_a) = \frac{\delta\alpha\left(1-\frac{q_a}{2}\right)+(1-\delta)\left[\alpha\left(1-\frac{q_a}{2}\right)-(1-q_b')\right]}{2\delta-1} > B_A(q_b, q_a),$$

as well as:

$$B_B(q_b') = \frac{\delta\left(1-\frac{q_b'}{2}\right)}{2\delta-1} > B_A(q_b, q_a);$$

in other words, offering more generous pre-investigation leniency contributes to deter futher "collude and report in case of investigations" strategies, without excessively triggering "collude and report systematically" ones. It may also further deter normal collusion (if $B_N^b(q_b') > B_N^a(q_a)$), otherwise a slight reduction in the

post-investigation amnesty rate q_a also increases $B_N\left(q_b', q_a\right) = B_N^a\left(q_a\right)$ while maintaining B_A and B_B above the initial levels; in both cases, the new policy improves all deterrence thresholds and thus makes the leniency programme more effective.

Given this, when post-investigation leniency is used to deter normal collusion, and since B_A^b decreases while B_N^a increases with q_a, the best amnesty rate q_a is such that these two thresholds coincide. Similarly, since B_B decreases while B_A^b increases with q_b, the best amnesty rate q_b is such that these two thresholds also coincide, as illustrated by Figure 5.

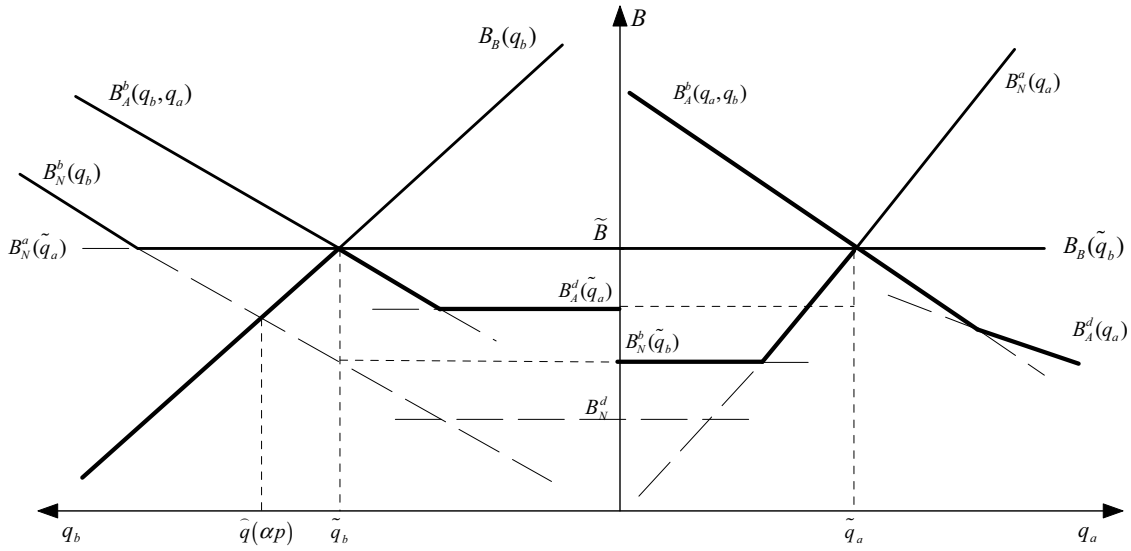

Figure 5: Optimal amnesty rates and deterrence threshold.

The candidate optimal amnesty rates $(\tilde{q}_b, \tilde{q}_a)$ are then such that

$$\tilde{B} = B_N^a\left(\tilde{q}_a\right) = B_A^b\left(\tilde{q}_b, \tilde{q}_a\right) = B_B\left(\tilde{q}_b\right),$$

and thus equal to:

$$\tilde{q}_b\left(\alpha, p\right) = 2\frac{\left(2\delta-1\right)\alpha + 2\left(1-\delta\right) - \left(1-\left(1-\alpha\right)\delta\right)\alpha p}{2\left(1-\delta\right)\left(2-\delta\right) + \delta\alpha}, \tag{15}$$

$$\tilde{q}_a\left(\alpha, p\right) = 2\frac{\left(1-\alpha\right)\delta\left(1-\delta\right) + \left(2-\delta\right)\left(1-\left(1-\alpha\right)\delta\right)\left(1-p\right)}{2\left(1-\delta\right)\left(2-\delta\right) + \delta\alpha}, \tag{16}$$

while the resulting deterrence threshold is:

$$\tilde{B}\left(\alpha, p\right) = \frac{\delta}{2\delta-1}\frac{\left(1-\delta\right)\left(2\left(1-\delta\right)+\alpha\right) + \left(1-\left(1-\alpha\right)\delta\right)\alpha p}{2\left(1-\delta\right)\left(2-\delta\right)+\delta\alpha}F.$$

To be optimal, pre-investigation leniency must however be more effective in deterring normal collusion, namely:

$$B_N^a\left(q_a\right) = \frac{\delta\alpha p + \left(1-\delta\right)\left(p-\left(1-q_a\right)\right)}{2\delta-1}F > \hat{B} = B_N^b\left(\hat{q}_b\right) = \frac{\alpha p - \left(1-\delta\right)\left(1-\hat{q}_b\right)}{2\delta-1}F,$$

or:

$$q_a > \underline{q}_a\left(\alpha, p\right) \equiv \hat{q}_b - \left(1-\alpha\right)p.$$

The policy should however also avoid triggering additional alternative collusive strategies, i.e.:

$$B_A^b\left(q_b, q_a\right), B_B\left(q_b\right) > \hat{B} = B_B\left(\hat{q}_b\right),$$

which, since B_B decreases as q_b increases, implies $q_b > \hat{q}_b$; since B_A^b increases with q_b, we must therefore have:

$$B_A^b\left(\hat{q}_b, q_a\right) = \frac{\alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)\left(1-\hat{q}_b\right)}{2\delta - 1} F > \hat{B} = B_N^b\left(\hat{q}_b\right) = \frac{\alpha p - (1-\delta)\left(1-\hat{q}_b\right)}{2\delta - 1} F,$$

that is:

$$1 - \frac{q_a}{2} > p,$$

or:

$$q_a < \bar{q}_a\left(p\right) \equiv 2\left(1-p\right).$$

As a result, post-investigation leniency can be useful only when $\underline{q}_a\left(\alpha, p\right) < \bar{q}_a\left(p\right)$, which amounts to:

$$p < \tilde{p}\left(\alpha\right) \equiv \frac{2\left(1-\delta\right)}{2 - \left(1+\alpha\right)\delta} \left(< 1\right),$$

where the threshold $\tilde{p}\left(\alpha\right)$ increases with α but remains lower than 1 for $\alpha < 1$, as illustrated in Figure 6.

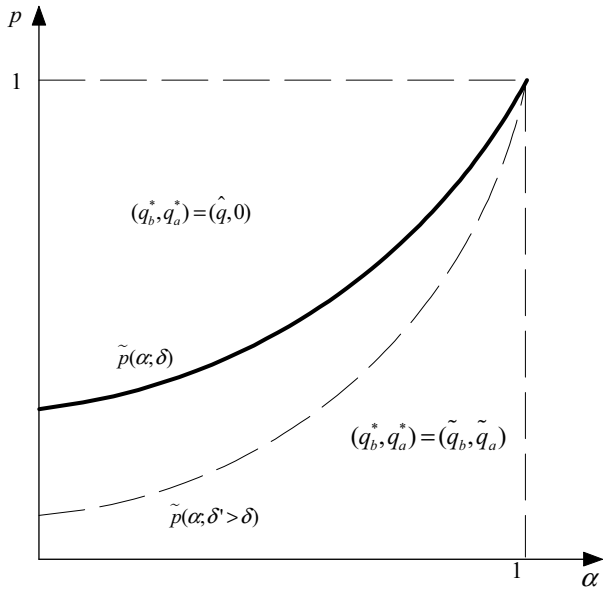

Figure 6: Optimal leniency pre- and post-investigation.

Conversely, Appendix D shows that post-investigation leniency is indeed desirable whenever $p < \tilde{p}\left(\alpha\right)$. We thus have:

PROPOSITION 5. *It is always optimal to offer leniency before investigations; moreover:*

- *for $p < \tilde{p}\left(\alpha\right)$, it is optimal to offer amnesty also when an investigation is already underway: the optimal policy is then $\left(q_b^*, q_a^*\right) = \left(\tilde{q}_b\left(\alpha, p\right), \tilde{q}_a\left(\alpha, p\right)\right)$ and the deterrence threshold is $B^* = \tilde{B}\left(\alpha, p\right)$.*

- *for $p > \tilde{p}(\alpha)$, it is optimal to restrict amnesty to the pre-investigation phase: the optimal policy is then $(q_b^*, q_a^*) = (\hat{q}(\alpha p), 0)$ and the deterrence threshold is $B^* = \hat{B}(\alpha p)$.*

Proof. See Appendix D. ∎

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). Obviously, an increase in either α or p furthers deters collusion: all deterrence thresholds increase with either α or p. However, α and p have different impacts on the desirability of post-investigation leniency: it is optimal to offer no leniency once an investigation is launched when random investigations are quite effective (i.e., when p is sufficiently *high*) or infrequent (i.e., when α, and thus $\tilde{p}(\alpha)$, is *low*). In practice, we would expect the probability 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. Moreover, since

$$\tilde{p}(0) = \frac{2(1-\delta)}{2-\delta} > 0,$$

our analysis suggests that offering amnesty post-investigation is indeed a valuable complement to *ex nihilo* investigations, whatever their frequency, when antitrust authorities have only limited detection tools or investigation powers.

4.5. Comparative statics

We now explore further the relation between the "stick" (measured by α and p) and the "carrot" (the amnesty rates). When cartels are likely to be uncovered even without any reporting (i.e., $p > \tilde{p}(\alpha)$), it is optimal to restrict leniency to pre-investigation phases, and the optimal amnesty rate is then determined as before: $q_b = \hat{q}(\alpha p)$, which decreases as the overall probability of prosecution, αp, increases. When it is instead unlikely to detect a cartel absent reporting (i.e., $p < \tilde{p}(\alpha)$), it is optimal to offer leniency both before and after investigations are launched. By construction, the marginal industry, characterized by $B = \tilde{B}$, is tempted to deviate from normal collusion by reporting whenever an investigation is launched, to deviate from "collude and report After an investigation" by reporting even before an investigation is launched, and to deviate from "collude and report Before investigations" by cheating on the product market.

Increasing p contributes to destabilize normal collusion and therefore overall enhances deterrence; since this does not directly affect the alternative forms of collusion that involve some reporting, these are deterred by decreasing both $\tilde{q}_b$ (otherwise, "collude and systematically report" would remain as robust as before) and $\tilde{q}_a$ (otherwise, "collude and report in case of investigation" would become more robust, due to the reduction in q_b). More precisely, an increase in p makes normal collusion more fragile and thus moves up the deterrence threshold $B_N^a(q_a)$, as illustrated in Figure 7; reducing the post-investigation amnesty rate to q_a^1 then prevents the marginal industry from colluding and reporting in case of investigations, while still keeping this industry away from normal collusion; this, in turn, makes it possible to deter this industry from adopting "collude and systematically report" strategies, by decreasing the amnesty rate before investigation to q_b^1, which in turn

calls for a further reduction in q_a, and so forth. As a result, an increase in p leads to a decrease in both amnesty rates, before and after investigations.

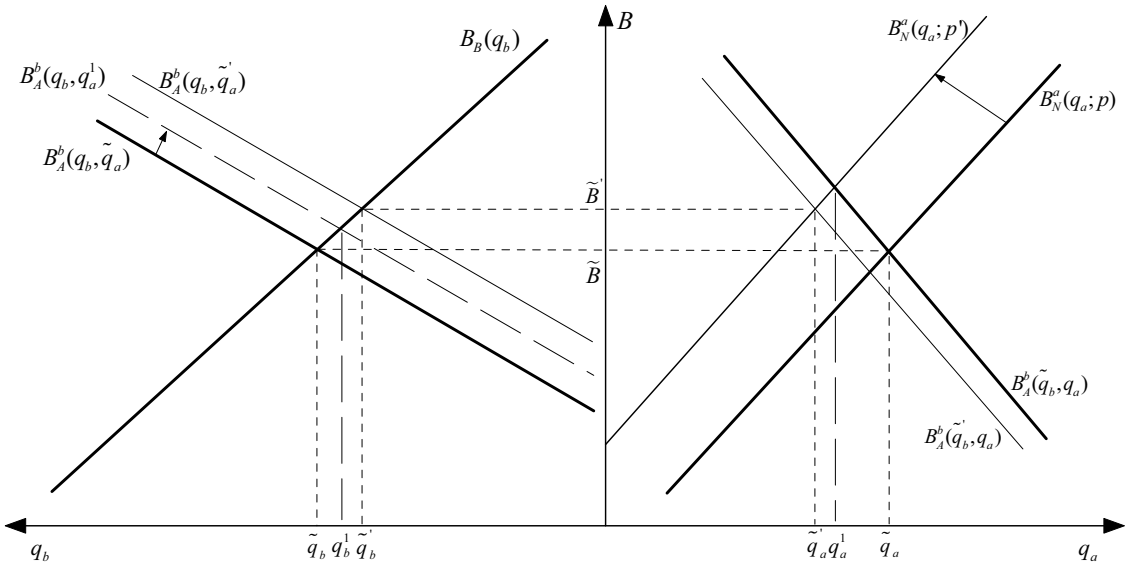

Figure 7: An increase in p leads to a decrease in $\tilde{q}_b$ and $\tilde{q}_a$.

Similarly, increasing the frequency of investigation α destabilizes both normal collusion and "collude and report under investigation" strategies, and thus enhances deterrence. And since this does not directly affect "collude and systematically report" strategies, the optimal pre-investigation amnesty rate q_b necessarily decreases. The impact on post-investigation leniency is however ambiguous, due to the fact that decreasing q_a, say, weakens "collude and report After investigation" but strengthens normal collusion. Whether the optimal rate goes up or down then depends on whether the increase in the frequency α and the concommitant decrease in $\tilde{q}_b$ have an overall relatively larger effect on normal collusion, in which case $\tilde{q}_a$ would also decrease, as in Figure 8a, or on "collude and report After an

investigation", in which case $\tilde{q}_a$ would instead increase, as illustrated in Figure 8b.

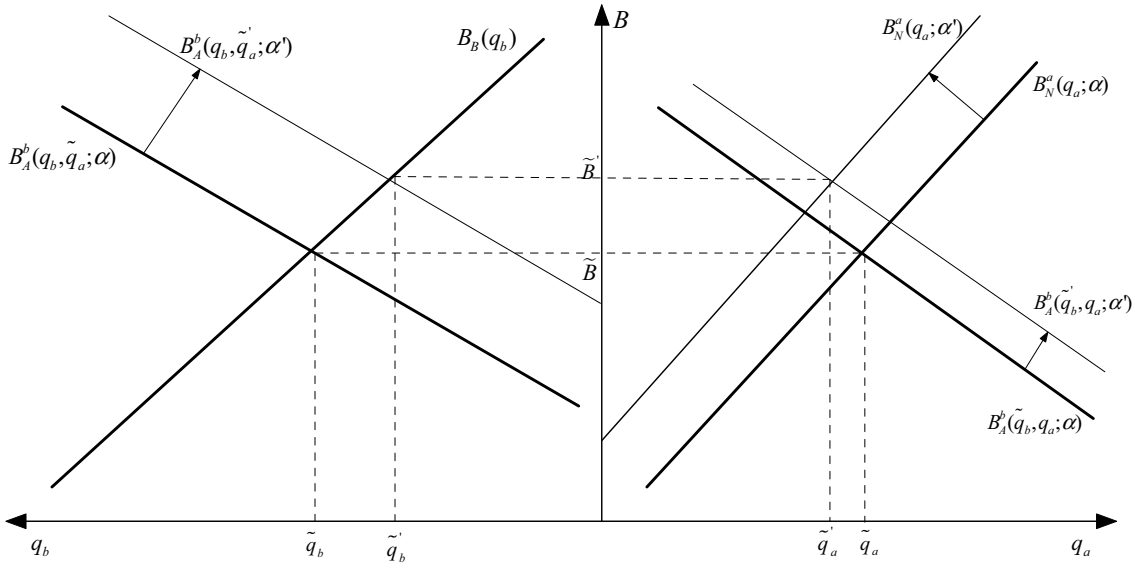

Figure 8a: An increase in α leads to a decrease in $\tilde{q}_b$ and $\tilde{q}_a$.

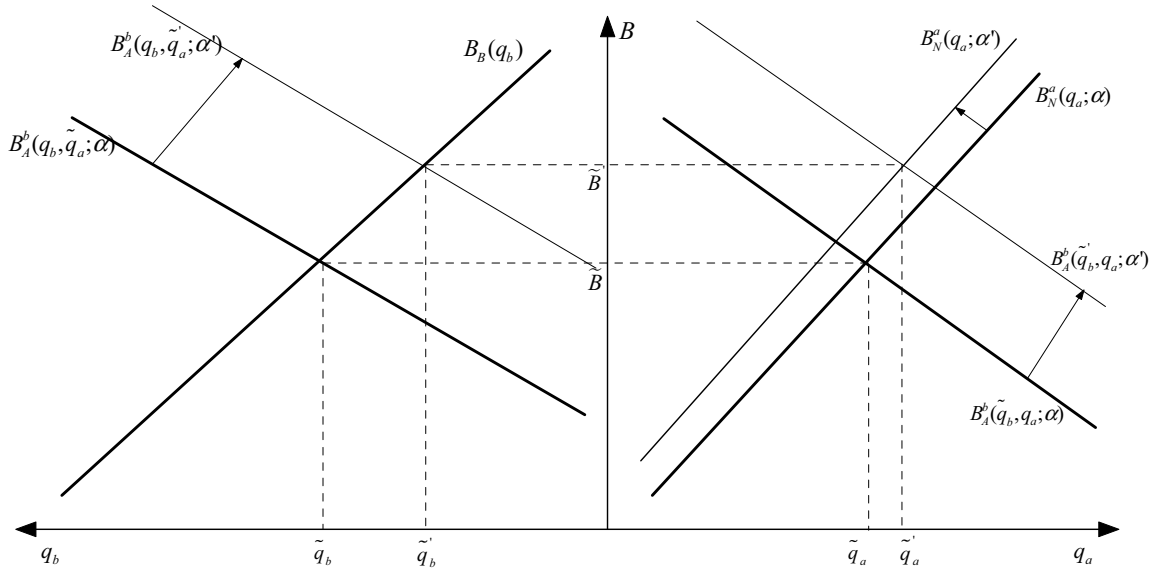

Figure 8b: An increase in α leads to a decrease in $\tilde{q}_b$ and an increase in $\tilde{q}_a$.

The latter case is likely to occur when increasing α has a relatively small impact on $\tilde{q}_b$, since this impact mitigates the direct positive impact of α on "collude and report After investigations" (indeed, keeping q_b constant, an increase in α would call for an increase in q_a),[22] as well as on normal collusion, which is the case when p is small. The detailed analysis, presented in Appendix E, shows that the latter

[22] Given $\tilde{q}_b$, the amnesty rate $\tilde{q}_a$ is determined by:

$$B_N^a = \delta\alpha p - (1-\delta)(1-p-q_a) = B_A^b = \alpha\left(A - \frac{q_a}{2}\right) - (1-\delta)(1-\tilde{q}_b),$$

case indeed occurs when $p > p_a$, where p_a is given by:

$$p_a \equiv \frac{2(1-\delta)}{(2-\delta)(3-2\delta)},$$

which does not depend on α and is positive, but lower than 1; we thus have:

PROPOSITION 6. *Increasing p or α makes the leniency program more effective. Moreover:*

- *$\hat{q}_b$, $\tilde{q}_b$ and $\tilde{q}_a$ decrease as p increases.*
- *$\hat{q}_b$ and $\tilde{q}_b$ decrease as α increases.*
- *There exists $p_a \in [0,1)$ such that $\widetilde{q}_a$ decreases as α increases when $p > p_a$, and $\widetilde{q}_a$ increases as α increases when $p < p_a$.*

Proof. See appendix E. ∎

It is interesting to see whether launching an investigation should lead to offer more or less leniency. The above analysis shows that when p is low, increasing α calls for less leniency before investigations, but more leniency once investigations are already underway. As a result, for low values of p and/or large values of α, it may become optimal to offer *more* leniency once an investigation is underway. The detailed analysis is presented in Appendix F and shows that this is indeed the case when (see Figure 9):

$$p < p_b(\alpha) = \frac{\alpha(1-\delta)}{(1-\delta+\alpha\delta)(2-\delta-\alpha)}.$$

which, keeping q_b constant, yields:

$$\frac{\partial q_a}{\partial \alpha} = \frac{1 - \frac{q_a}{2} - \delta p}{1 - \frac{\delta}{2}},$$

which is positive since, by construction, $\tilde{q}_a < \bar{q}_a$ implies $1 - \frac{q_a}{2} > p > \delta p$.

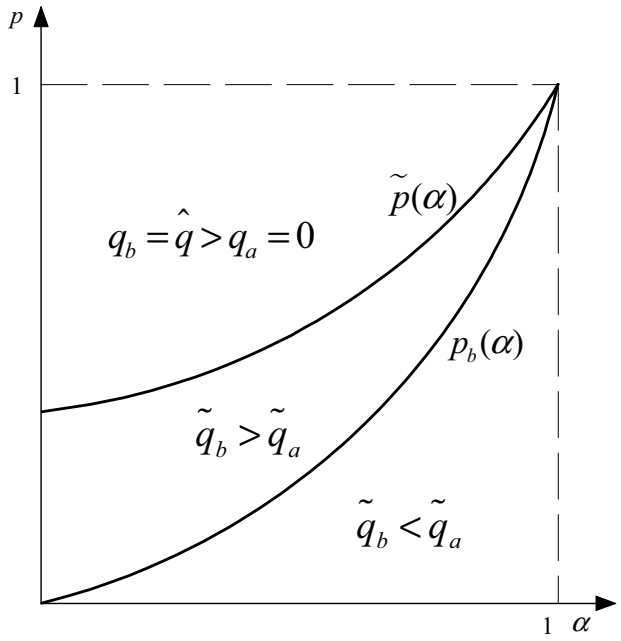

Figure 9: Impact of p and α on the relative importance of pre- and post-investigation leniency.

We have:

PROPOSITION 7. *It is optimal to offer more amnesty post-investigation than pre-investigation if and only if $p < p_b(\alpha)$, where $p_b(\alpha)$ increases with α and is positive for $\alpha > 0$, and moreover satisfies $p_b(\alpha) < \tilde{p}(\alpha)$.*

Proof. See Appendix F. ∎

Leniency programs usually 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 shows that such policies may not be optimal when investigations are quite frequent or (more realistically) 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 basic trade-offs between destabilizing collusion and deterring cartel formation. We use a standard model of tacit collusion in a repeated competition game and focus on stationary antitrust policies which rely on random investigations and fines for exposed cartels. In this context, we show that offering leniency, before or after an investigation is launched, can help fight collusion; we also relate the scope for leniency to the frequency of investigations and their likelihood of success.

To study the effectiveness of leniency programs, we suppose that industries differ in their benefits from collusion. Deterring collusion "as much as possible" then amounts to maximize the threshold on collusion benefits below which collusion

is deterred. The optimal leniency programs balances two effects: (i) destabilizing usual collusion, by encouraging firms to deviate and denounce the cartel; and (ii) discouraging firms from abusing the system, by colluding as well on reporting strategies. Our simple framework allows us to relate the optimal solution to this trade-off to the frequency and likely success of investigations. In particular, it is optimal to offer more leniency before investigations whenever random investigations are insufficiently frequent or successful; 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 appear to support limiting leniency for repeated offenders (increasing the monitoring of such industries appears desirable, though).

The framework can also allow to consider further the impact of leniency programs on desistance, which becomes relevant when exposed cartels are prevented from colluding for at least some time. Our first exploration suggests that favoring deterrence may be optimal when cartels are "uniformly bad", but that desistance may become interesting when the harm to consumers or society increases with the private benefits from collusion. Extending the framework in the spirit of Harrington (2006), would allow a further analysis of the impact of pre- and post-investigation leniency on cartel duration as well as cartel formation.

Appendices

Appendix A: Proof of Proposition 2

The first part of the proposition follows directly from the previous considerations: $\hat{q}$ is characterized as the threshold where the increasing function $B^r(.)$ intersects the decreasing function $B_R(.;\tau)$; thus an increase in τ, which moves $B_R(.;\tau)$ down, results in a reduction in $\hat{q}$:

$$\frac{d\hat{q}}{d\tau} = \frac{-\partial_\tau B_R}{\partial_q B_R - \partial_q B^r} < 0.$$

Consider now the case where all informants would benefit from the same amnesty rate q (that is, $\tau = 1$). In that case,

$$B_R(q,1) = \frac{\delta(1-q)F}{2\delta - 1},$$

which is lower than $\underline{B}$ whenever $q > \bar{q}(\tau = 1)\, 1 - \rho = \underline{q}$; therefore, any $q \leq \underline{q}$, including $q = 0$, is optimal, while any higher q would foster alternative forms of collusion.

Appendix B: Proof of Proposition 4

We have:

$$\begin{aligned}\frac{\partial B^r(q;T)}{\partial T} &= \frac{\partial B^r(q;T)}{\partial \gamma}\gamma'(T) \\ &= \frac{\delta(1-q-2\rho)}{(2\gamma-1)^2}F\rho\beta'(T) > 0,\end{aligned}$$

as $q > \underline{q} = 1 - \rho$ and $\beta' < 0$, and similarly:

$$\begin{aligned}\frac{\partial \underline{B}(q;T)}{\partial T} &= \frac{\partial B^r(q;T)}{\partial \gamma}\gamma'(T) \\ &= \frac{-\rho F}{(2\gamma-1)^2}\gamma'(T) > 0,\end{aligned}$$

as $\gamma' < 0$.

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

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

Then:

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

where the function ϕ satisfies

$$\begin{aligned}\phi'(\beta) &= \frac{\partial\left(((2\beta-1)(2-\beta)+2(1-\rho)(1-\beta)(\delta-\beta)) - ((2\rho-2\delta-4\beta\rho+2\delta\rho+3)(\beta+\delta-1))\right)}{\partial\beta} \\ &= 4\rho(\beta+\delta-1) > 0\end{aligned}$$

and:

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

Therefore, as T increases, β decreases below δ and $\hat{q}$ increases.

Appendix C: Deterrence vs Desistance

Suppose that the collusion benefit, B, is uniformly distributed over $\left[0, \overline{B}\right]$, where $\hat{B}(T) < \overline{B} < \max\left\{B^r\left(q^M\right), B^N\left(q^M\right)\right\}$, where $q^M = 2$ is the amnesty rate for which $B_R(q) = 0$. This ensures that the relevant range for q in $[\hat{q}(T), \check{q}(T)]$, where $\check{q}(T) < q^M$ is such that $\max\left\{B^r(q;T), B^N(q;T)\right\} = \overline{B}$: setting q lower than $\hat{q}(T)$ (in which case all industries $B \geq \max\{\underline{B}, B^r(q)\}$ engage in normal collusion, which they can sustain and favor over alternative forms of collusion) is dominated by $q = \hat{q}(T)$, since this improves deterrence without affecting desistance. And setting q higher than $\check{q}$ (in which case all industries $B \geq B_R(q)$ collude and report, since they can sustain this form of collusion and prefer it to normal collusion) is dominated by $q = \check{q}(T)$, since this improves deterrence (the above assumption implies $\check{q}(T) < q^M$, and thus $B_R(\check{q}(T)) > 0$) without affecting desistance.

For $\hat{q}(T) \leq q \leq q^N(T)$, $B^r(q) > B^N(q), B_R(q)$ and the total welfare loss is thus equal to

$$\begin{aligned} L(q;T) &= L_1(q;T) \equiv \int_{B_R(q;T)}^{\overline{B}} L_R(B)\frac{dB}{\overline{B}} + \int_{B^r(q;T)}^{\overline{B}} (L_N(B) - L_R(B))\frac{dB}{\overline{B}} \\ &= \int_{B_R(q;T)}^{\overline{B}} \frac{D(B)}{1-\beta}\frac{dB}{\overline{B}} + \int_{B^r(q;T)}^{\overline{B}} \frac{(\gamma-\beta)D(B)}{(1-\gamma)(1-\beta)}\frac{dB}{\overline{B}}. \end{aligned}$$

The first-order derivative is equal to:

$$\begin{aligned} \frac{\partial L_1}{\partial q} &= -\frac{D(B_R)}{\overline{B}(1-\beta)}\frac{\partial B_R}{\partial q} - \frac{(\gamma-\beta)D(B^r)}{\overline{B}(1-\beta)(1-\gamma)}\frac{\partial B^r}{\partial q} \\ &= \frac{F}{\overline{B}(1-\beta)}\left\{\frac{\beta D(B_R)}{2(2\beta-1)} - \frac{(\gamma-\beta)D(B^r)}{2\gamma-1}\right\}. \end{aligned}$$

Similarly, for $q \geq q^N(T)$, $B^N(q) > B^r(q) > B_R(q)$ and:

$$L(q;T) = L_2(q;T) \equiv \int_{B_R(q;T)}^{\overline{B}} \frac{D(B)}{1-\beta}\frac{dB}{\overline{B}} + \int_{B^N(q;T)}^{\overline{B}} \frac{(\gamma-\beta)D(B)}{(1-\gamma)(1-\beta)}\frac{dB}{\overline{B}},$$

$$\frac{\partial L_2}{\partial q} = -\frac{D(B_R)}{\overline{B}(1-\beta)}\frac{\partial B_R}{\partial q} - \frac{(\gamma-\beta)D\left(B^N\right)}{\overline{B}(1-\beta)(1-\gamma)}\frac{\partial B^N}{\partial q}.$$

Suppose first that $D(.)$ is constant across industries ($D(B) = D$, and thus $D'(B) = 0$); then, since $0 < \partial B^r/\partial q < \partial B^N/\partial q$:

$$\begin{aligned} \frac{\partial L}{\partial q} &\geq \frac{\partial L_2}{\partial q} = -\frac{D}{\overline{B}(1-\beta)}\frac{\partial B_R}{\partial q} - \frac{(\gamma-\beta)D}{\overline{B}(1-\beta)(1-\gamma)}\frac{\partial B^N}{\partial q} \\ &= \frac{DF}{\overline{B}(1-\beta)}\left(\frac{\beta}{2(2\beta-1)} - \frac{(\gamma-\beta)}{(1-\gamma)}\frac{(1-\gamma)}{2(\gamma-\beta)}\right) \\ &= \frac{DF}{2\overline{B}(2\beta-1)} > 0. \end{aligned}$$

It is therefore optimal to set $q = \hat{q}(T)$.

Suppose now that social damages are positively related to the collusion benefit: $D'(B) > 0$. Then:

$$\begin{aligned}
\frac{\partial^2 L_1}{\partial q^2} &= -\frac{D'(B_R)}{\overline{B}(1-\beta)}\left(\frac{\partial B_R}{\partial q}\right)^2 - \frac{(\gamma-\beta)D'(B^r)}{(1-\beta)(1-\gamma)}\left(\frac{\partial B^r}{\partial q}\right)^2 > 0, \\
\frac{\partial^2 L_2}{\partial q^2} &= -\frac{D'(B_R)}{\overline{B}(1-\beta)}\left(\frac{\partial B_R}{\partial q}\right)^2 - \frac{(\gamma-\beta)D'\left(B^N\right)}{(1-\beta)(1-\gamma)}\left(\frac{\partial B^N}{\partial q}\right)^2 > 0.
\end{aligned}$$

Moreover, for $q = q^N(T)$, $B^r(q;T) = B^N(q;T)$ and thus:

$$\begin{aligned}
\frac{\partial L_2}{\partial q} &= -\frac{D(B_R)}{\overline{B}(1-\beta)}\frac{\partial B_R}{\partial q} - \frac{(\gamma-\beta)D\left(B^N\right)}{(1-\beta)(1-\gamma)}\frac{\partial B^N}{\partial q} \\
&= -\frac{D(B_R)}{\overline{B}(1-\beta)}\frac{\partial B_R}{\partial q} - \frac{(\gamma-\beta)D(B^r)}{(1-\beta)(1-\gamma)}\frac{\partial B^N}{\partial q} \\
&< \frac{\partial L_1}{\partial q}\left(q^N(T);T\right) = -\frac{D(B_R)}{\overline{B}(1-\beta)}\frac{\partial B_R}{\partial q} - \frac{(\gamma-\beta)D(B^r)}{(1-\beta)(1-\gamma)}\frac{\partial B^r}{\partial q}.
\end{aligned}$$

Therefore, $L(q;T)$ is globally concave over the relevant range and the optimal amnesty rate is thus either $q = \hat{q}(T)$ or $q = \check{q}(T)$.

In particular, for $D(B) = kB$ and $q > q^N(T)$:

$$\begin{aligned}
\Delta(q;T) &\equiv L_1\left(\hat{q}(T);T\right) - L_2(q;T) \\
&= \int_{\hat{B}(T)}^{B^N(q;T)} \left(L_N(B) - L_R(B)\right)\frac{dB}{\overline{B}} - \int_{B_R(q;T)}^{\hat{B}(T)} L_R(B)\frac{dB}{\overline{B}} \\
&= \frac{k}{(1-\beta)\overline{B}}\left\{\int_{\hat{B}(T)}^{B^N(q;T)} \frac{\gamma-\beta}{1-\gamma} BdB - \int_{B_R(q;T)}^{\hat{B}(T)} BdB\right\} \\
&= \frac{k}{2(1-\beta)\overline{B}}\left\{\frac{\gamma-\beta}{1-\gamma}\left[\left(B^N(q;T)\right)^2 - \left(\hat{B}(T)\right)^2\right] - \left[\left(\hat{B}(T)\right)^2 - \left(B_R(q;T)\right)^2\right]\right\} \\
&= \frac{k}{2(1-\beta)\overline{B}}\left\{\frac{\gamma-\beta}{1-\gamma}\left(B^N(q;T)\right)^2 + \left(B_R(q;T)\right)^2 - \frac{1-\beta}{1-\gamma}\left(\hat{B}(T)\right)^2\right\} \\
&= \frac{k}{2(1-\gamma)\overline{B}}\left\{\frac{\gamma-\beta}{1-\beta}\left(B^N(q;T)\right)^2 + \frac{1-\beta}{1-\gamma}\left(B_R(q;T)\right)^2 - \left(\hat{B}(T)\right)^2\right\}
\end{aligned}$$

In the particular case where $\delta = 0.85$, $\beta = 0.8$ and $\rho = 0.3$, $q^N(T) \simeq 1.5$ and, in the limit case where $\bar{B} = B^N(2;T)\,(> B^r(2;T))$, so that $\check{q}(T) = 2$ and thus $B_R\left(\check{q}(T);T\right) = 0$, the sign of $\Delta\left(\check{q}(T);T\right)$ is the same as that of

$$\psi \equiv \frac{\gamma-\beta}{1-\beta}\left(B^N(2;T)\right)^2 - \left(\hat{B}(T)\right)^2 \simeq 0.26 > 0.$$

Therefore, for $\bar{B}$ close to $B^N(2;T)$, the optimal leniency policy will be to focus on desistance and thus to set the amnesty rate to the maximal relevant level, $q = \check{q}(T)$, in order to induce alternative, "reporting" strategies in all collusive industries.

Appendix D: Proof of Proposition 5

We first note that, since the antitrust authority can always do as well as with secret investigations, *some leniency* is optimal; in particular, the optimal deterrence threshold, B^*, is necessarily such that $B^* > B_N^d$, which in turn implies that the constraint $B \geq B_N^d$ is not relevant.

Now, let $B_N^* \equiv B_N(q_b^*, q_a^*)$, $B_A^* \equiv B_A(q_b^*, q_a^*)$ and $B_B^* \equiv B_B(q_b^*)$ denote the deterrence thresholds for the three types of collusion strategies, under the optimal leniency program. The following lemma shows that, without loss of generality, we can restrict attention to the situation where these three thresholds coincide:

LEMMA 1. *There exists an optimal policy such that* $B_N^* = B_A^* = B_B^* = B^*$.

Proof. Several cases can arise, which we study in turn.

(1) Suppose $B^* = B_i^* < B_j^*, B_k^*$, for $i \neq j \neq k = N, A, B$. Then:

- If $i = N$, slightly increasing either q_b^* or q_a^* would increase $B^* = B_N(q_b^*, q_a^*) = \max\{B_N^a(q_a^*), B_N^b(q_b^*)\}$ ($> B_N^d$ from the previous Lemma), a contradiction.
- If $i = A$, slightly decreasing q_a^* would increase $B^* = B_A(q_b^*, q_a^*) = \max\left\{B_A^d(q_a^*), B_A^b(q_b^*, q_a^*)\right\}$, a contradiction.
- If $i = B$, slightly decreasing q_b^* would increase $B^* = B_B(q_b^*) = B_B(q_b^*)$, a contradiction.

(2) Suppose $B_N^* > B_A^* = B_B^*$. Then $B_N^b(q_b^*, q_a^*) > B_A(q_b^*, q_a^*) = \max\{B_A^d(q_a^*), B_A^b(q_b^*, q_a^*)\} = B_B(q_b^*) = B^*$, where B_A decreases in q_a and may increase in q_b (if $B_A = B_A^b$), while $B_B(q_b)$ decreases in q_b. But then, slightly decreasing q_a^* would increase B_A^*, which in turn would allow increasing B_B^* (by decreasing q_b), a contradiction.

(3) Suppose $B_A^* > B_N^* = B_B^*$. There are two cases to consider:

1) $B_A(q_b^*, q_a^*) > B^* = B_B(q_b^*) = B_N^a(q_a^*) \geq B_N^b(q_b^*)$. Then decreasing q_b and increasing q_a would increase B_B and $B_N = B_N^a$, a contradiction.

2) $B_A(q_b^*, q_a^*) > B^* = B_B(q_b^*) = B_N^b(q_b^*) \geq B_N^a(q_a^*)$. Since $B_N^a(q_a)$ and $B_A(q_b^*, q_a)$ respectively increase and decrease with q_a, there thus exists $q_a^0 > q_a^*$ such that $B_N^a(q_a) = B_A(q_b^*, q_a)$. Two subcases need to be considered:

a) If $B_N^a(q_a^0) = B_A(q_b^*, q_a^0) > B^*$, then increasing q_a^* to q_a^0 yields $B_A = B_A(q_b^*, q_a^0) > B^*$ and $B_N = B_N^a(q_a^0) > B_N^b(q_b^*) = B^*$, and a slight decrease in q_b^* would then increase $B_B = B_B(q_b^*)$ as well, a contradiction.

b) If $B_N^a(q_a^0) = B_A(q_b^*, q_a^0) \leq B^*$, there exists q_a^1 satisfying $q_a^* < q_a^1 < q_a^0$ and such that $B_A(q_b^*, q_a^1) = B^* > B_N^a(q_a^1)$; then, increasing q_a^* to q_a^1: (i) does not affect $B_B = B_B(q_b^*)$; (ii) leaves $B_N = B_N^b(q_b^*) = B^* > B_N^a(q_a^1)$ unchanged; and (iii) reduces B_A to $B_A(q_b^*, q_a^1) = B^*$. Thus, $\left(q_b^*, q_a^1\right)$ is also optimal and moreover satisfies $B_N\left(q_b^*, q_a^1\right) = B_A\left(q_b^*, q_a^1\right) = B_B\left(q_b^*\right) = B^*$.

(4) Suppose $B_B^* > B_N^* = B_A^* = B^*$. Then $B_B(q_b^*) > \max\{B_A^d(q_a^*), B_A^b(q_b^*, q_a^*)\} = \max\{B_N^a(q_a^*), B_N^b(q_b^*)\}$. There are three cases to consider:

1) $B_N^* = B_N^b(q_b^*) \geq B_N^a(q_a^*)$. Then increasing q_b would increase $B_N(q_b, q_a^*) = B_N^b(q_b) > B_N^a(q_a^*)$, which would allow to increase B_A as well (by slightly decreasing q_a), a contradiction.

2) $B_A^* = B_A^b(q_b^*, q_a^*) \geq B_A^d(q_a^*)$. Then increasing q_b would increase $B_A(q_b, q_a^*) = B_A^b(q_b, q_a^*) > B_A^d(q_a^*)$, which would allow to increase B_N as well (by slightly increasing q_a), a contradiction.

3) $B_N^* = B_N^a(q_a^*) > B_N^b(q_b^*)$ and $B_A^* = B_A^d(q_a^*) > B_A^b(q_b^*, q_a^*)$. Then $B_B(q_b^*) > B^* > \max\{B_N^b(q_b^*), B_A^b(q_b^*, q_a^*)\}$. Define $B_{N,A}^b(q_b, q_a) \equiv \max\{B_N^b(q_b), B_A^b(q_b, q_a)\}$ and $q_b^0 > q_b^*$ such that $B_B^*(q_b^0) = B_{N,A}^b(q_b^0, q_a^*)$. There are two subcases:

a) If $B_B(q_b^0) = B_{N,A}^b(q_b^0, q_a^*) > B^*$, then increasing q_b to q_b^0:

- either leads to $B_N^b(q_b) > B^* = B_N^a(q_a^*)$, and thus $B_N, B_B > B^*$; B_A could then be also increased by decreasing q_a, a contradiction.
- or leads to $B_A^b(q_b^0, q_a^*) > B^* = B_A^d(q_a^*)$, and thus $B_A, B_B > B^*$; B_N could then be also increased by increasing q_a, a contradiction

b) If $B_B(q_b^0) = B_{N,A}^b(q_b^0, q_a^*) \leq B^*$, there exists q_b^1 satisfying $q_b^* < q_b^1 < q_b^0$ and such that $B_B^*(q_b^1) = B^* \geq B_{N,A}^b(q_b^1, q_a^*)$. Hence $\left(q_b^1, q_a^*\right)$ is also optimal and moreover satisfies $B_N\left(q_b^1, q_a^*\right) = B_A\left(q_b^1, q_a^*\right) = B_B\left(q_b^1\right) = B^*$.

Q.E.D. ∎

Thanks to Lemma 1, to characterize the optimum, we only need to consider three situations.

(1) Situation 1: $B\left(q_b, q_a\right) = B_N^b\left(q_b\right) = B_A\left(q_b, q_a\right) = B_B\left(q_b\right) \geq B_N^a\left(q_a\right), B_N^d$. The pre-investigation amnesty rate q_b is thus characterized by

$$B_N^b(q_b) = B_B(q_b),$$

and is therefore equal to

$$\hat{q}_b \equiv \hat{q}\left(\alpha p\right) = \frac{1 - \alpha p}{1 - \dfrac{\delta}{2}}.$$

The resulting deterrence threshold is equal to

$$\hat{B}\left(\alpha p\right) = \frac{\delta\left(1 - \delta + \alpha p\right)}{\left(2\delta - 1\right)\left(2 - \delta\right)} F,$$

which, as already noted, is indeed higher than B_N^d.

In this situation, without loss of generality, one can moreover set $q_a = 0$, i.e., by grant amnesty only before an investigation is opened: reducing q_a reduces B_N^a, but has no impact on $B_N\left(\hat{q}_b, q_a\right) = B_N^b\left(\hat{q}_b\right) \geq B_N^a\left(q_a\right)$, and increase $B_A\left(\hat{q}_b, q_a\right)$. Furthermore, for $q_a = 0$ and $q_b = \hat{q}_b$, we have:

$$B_N^a\left(q_a = 0\right) = \frac{\delta\alpha p - \left(1 - \delta\right)\left(1 - p\right)}{2\delta - 1} F < \frac{\delta\alpha p}{2\delta - 1} F = B_N^d < \hat{B}\left(\alpha p\right),$$

so that $B_N\left(q_b, q_a\right) = B_N^b(q_b) = \hat{B}\left(\alpha p\right) > B_N^d > B_N^a\left(q_a\right)$, and:

$$\begin{aligned} B_A\left(q_b, 0\right) &\geq B_A^b\left(q_b, 0\right) \\ &= \frac{\alpha - \left(1 - \delta\right)\left(1 - q_b\right)}{2\delta - 1} F \\ &> \frac{\alpha p - \left(1 - \delta\right)\left(1 - q_b\right)}{2\delta - 1} F \\ &= B_N^b\left(q_b\right) \\ &= \hat{B}\left(\alpha p\right), \end{aligned}$$

which thus ensures $B(q_b, q_a) = B_N(q_b, q_a) = B_B(q_b) = \hat{B}(\alpha p) \leq B_A(q_b, q_a)$.

(2) Situation 2: $B_N^a(q_a) = B_A^d(q_a) = B_B(q_b) \geq B_N^b(q_b), B_A^b(q_b, q_a), B_N^d$. The rate q_b therefore satisfies:

$$B_B(q_b) = \frac{\delta\left(1 - \frac{q_b}{2}\right)}{2\delta - 1}F = B_A^d(q_a) \equiv \frac{\delta\alpha\left(1 - \frac{q_a}{2}\right)}{2\delta - 1}F,$$

and thus:

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

but this implies

$$B_A^b(q_b, q_a) = B_A^d(q_a) + \left[\alpha\left(1 - \frac{q_a}{2}\right) - (1 - q_b)\right]\frac{(1-\delta)F}{2\delta - 1} > B_A^d(q_a),$$

a contradiction.

(3) Situation 3: $B_N^a(q_a) = B_A^b(q_b, q_a) = B_B(q_b) = \tilde{B} \geq B_N^b(q_b), B_A^d(q_a), B_N^d$. The optimal amnesty rates are then such that

$$\begin{aligned} B_N^a(q_a) &= \frac{\delta\alpha p + (1-\delta)(p - (1-q_a))}{2\delta - 1}F = B_A^b(q_b, q_a) = \frac{\alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)(1-q_b)}{2\delta - 1}F, \\ B_A^b(q_b, q_a) &= \frac{\alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)(1-q_b)}{2\delta - 1}F = B_B(q_b) = \frac{\delta\left(1 - \frac{q_b}{2}\right)}{2\delta - 1}F, \end{aligned}$$

that is:

$$\begin{gathered} \delta\alpha p + (1-\delta)(p - (1-q_a)) = \alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)(1-q_b) \\ \alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)(1-q_b) = \delta\left(1 - \frac{q_b}{2}\right) \end{gathered}$$

and they are thus equal to:

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

It is straightforward to check that $\tilde{q}_b$ and $\tilde{q}_a$ both decrease as p increases (but remain positive even for $p = 1$). The resulting deterrence threshold is equal to

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

We have moreover $B_A^q(\tilde{q}_a) < \tilde{B} = B_A^b(\tilde{q}_b, \tilde{q}_a)$ when

$$\varphi \equiv \alpha\left(1 - \frac{\tilde{q}_a}{2}\right) - (1 - \tilde{q}_b) > 0,$$

which is satisfied since $p \leq 1$ and:

$$\begin{aligned} \frac{\partial\varphi}{\partial p} &= -\alpha\delta\frac{1 - \delta + \alpha\delta}{2(1-\delta)(2-\delta) + \delta\alpha} < 0, \\ [\varphi]_{p=1} &= \delta(1-\alpha)\frac{2(1-\delta) + \alpha\delta}{2(1-\delta)(2-\delta) + \delta\alpha} > 0. \end{aligned}$$

However, $B_N^b(\tilde{q}_b) < \tilde{B} = B_A^b(\tilde{q}_b, \tilde{q}_a)$ only when

$$\begin{aligned} 0 &< 1 - \frac{q_a}{2} - p \\ &= (1-\delta)\frac{2(1-\delta) - (2-(1+\alpha)\delta)p}{2(1-\delta)(2-\delta) + \delta\alpha}, \end{aligned}$$

which puts a ceiling on admissible values of p, which must satisfy:

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

Conversely, when this condition is satisfied, we have:

$$B_N^b(\tilde{q}_b) < \tilde{B} = B_B(\tilde{q}_b),$$

where $B_N^b(q_b)$ increases while $B_B(q_b)$ decreases with q_b, and intersects for $q_b = \hat{q}(\alpha p)$,which in turn implies $\tilde{q}_b < \hat{q}(\alpha p)$ and thus:

$$\tilde{B} = B_B(\tilde{q}_b) > B_B(\hat{q}(\alpha p)) = \hat{B}(\alpha p).$$

Therefore:

- when $p < \tilde{p}$ the optimal policy involves post-investigation leniency; it is given by $(q_b^*, q_a^*) = (\tilde{q}_b, \tilde{q}_a)$, where $\tilde{q}_b < \hat{q}(\alpha p)$, and yields a deterrence threshold $\tilde{B} > \hat{B}(\alpha p)$.
- when instead $p > \tilde{p}$ the optimal policy involves no post-investigation leniency; it then given as before by $(q_b^*, q_a^*) = (\hat{q}(\alpha p))$ and its deterrence threshold is $\hat{B}(\alpha p)$.

Appendix E: Proof of Proposition 6

As we have seen, an increase in p makes the leniency program more robust and thus increases $\tilde{B}$ and $\hat{B}$, which in turn calls for less leniency both before and after investigation:$\frac{\partial \tilde{q}_a}{\partial p} < 0, \frac{\partial \tilde{q}_b}{\partial p} < 0$ and $\frac{\partial \hat{q}}{\partial p} = \alpha \hat{q}'(\alpha p) < 0$. Similarly, an increase in α also makes the antitrust policy more effective and thus increases $\tilde{B}$ and $\hat{B}$. We have moreover:

$$\begin{aligned} \frac{\partial \tilde{q}_b}{\partial \alpha} &= -2\frac{4p - 12\delta - 10p\delta + 12\delta^2 - 4\delta^3 + 8p\delta^2 - 2p\delta^3 + p\alpha^2\delta^2 + 8p\alpha\delta - 12p\alpha\delta^2 + 4p\alpha\delta^3 + 4}{(2(1-\delta)(2-\delta) + \delta\alpha)^2} \\ &\equiv -2\frac{\varphi_b}{(2(1-\delta)(2-\delta) + \delta\alpha)^2}, \end{aligned}$$

where

$$\frac{\partial \varphi_b}{\partial \alpha} = 2p\delta(2(1-\delta)(2-\delta) + \delta\alpha) > 0 \text{ and } \varphi|_{\alpha=0} = \left(2(\delta-1)^2(2p - 2\delta - p\delta + 2)\right) > 0.$$

Therefore, $\tilde{q}_b$ decreases as α increases.

Finally,

$$\frac{\partial \tilde{q}_a}{\partial \alpha} = 2\delta\left(1-\delta\right)\frac{2\left(1-\delta\right)-\left(2-\delta\right)\left(3-2\delta\right)p}{\left(2\left(1-\delta\right)\left(2-\delta\right)+\delta\alpha\right)^2},$$

and thus $\tilde{q}_a$ increases with α as long as

$$p < p_a \equiv \frac{2\left(1-\delta\right)}{\left(2-\delta\right)\left(3-2\delta\right)},$$

where $p_a \in [0,1)$, as illustrated by the following figure:

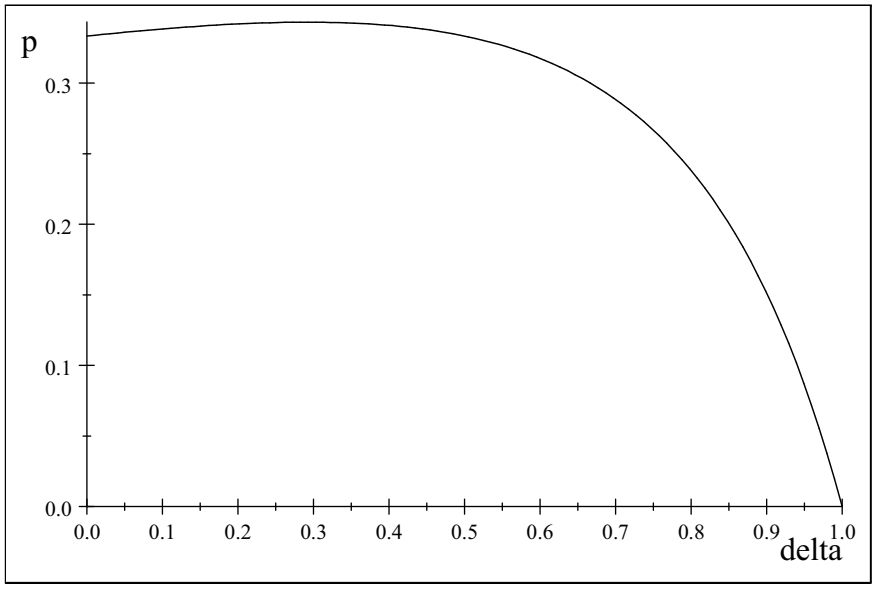

Figure A.1: p_a as a function of the discount factor δ

Appendix F: Proof of Proposition 7

We now compare $\tilde{q}_b$ and $\tilde{q}_a$:

$$\tilde{q}_b - \tilde{q}_a = 2\frac{\left(1-\delta+\alpha\delta\right)\left(2-\delta-\alpha\right)p-\alpha\left(1-\delta\right)}{2\left(1-\delta\right)\left(2-\delta\right)+\delta\alpha},$$

and thus $\tilde{q}_b \geq \tilde{q}_a$ if and only if

$$p \geq p_b(\alpha) = \frac{\alpha\left(1-\delta\right)}{\left(1-\delta+\alpha\delta\right)\left(2-\delta-\alpha\right)},$$

where

$$\tilde{p}\left(\alpha\right) - p_b(\alpha) = \left(1-\delta\right)\left(1-\alpha\right)\frac{2\left(1-\delta\right)\left(2-\delta\right)+\delta\alpha}{\left(2-\left(1+\alpha\right)\delta\right)\left(1-\delta+\alpha\delta\right)\left(2-\delta-\alpha\right)} > 0,$$

and:

$$\frac{dp_b}{d\alpha} = \left(1-\delta\right)\frac{\left(1-\delta\right)\left(2-\delta\right)+\alpha^2\delta}{\left(1-\delta+\alpha\delta\right)^2\left(2-\delta-\alpha\right)^2} > 0.$$

References

Aubert, C., P. Rey and W. Kovacic (2005), "The Impact of Leniency and Whistleblowing Program on Cartels", *International Journal of Industrial Organization*, forthcoming.

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*, forthcoming.

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. (2005), "Optimal Corporate Leniency Programs", *mimeo*, Johns Hopkins University.

Harrington, J. (2006), "Modelling the Birth and Death of Cartels with an Application to Evaluating Antitrust Policy", *mimeo*, Johns Hopkins University.

Motta, M., and M. Polo (2003), "Leniency Programs and Cartel Prosecution", *International Journal of Industrial Organization* 21:347-379.

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. (2004), "*Divide et Impera*: Optimal Leniency Programmes", CEPR Discussion Paper N°4840, available at: www.cepr.org/pubs/dps/DP4840.asp.