On the Design of Leniency Programs[1]

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

and

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

Version: December 2008

Abstract

Most of the existing literature on leniency programs assume that cartels are unable to exploit the generous amnesty by incorporating strategic reporting as part of collusive strategies; this assumption might be too optimistic and the effectiveness of leniency policy might be overestimated as a result. We develop a simple framework for analyzing the optimal design of leniency programs, which highlights the 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 and can thus be exploited strategically by cartels. We characterize the optimal leniency rates by solving this trade-off, both before any investigation and once an investigation is opened, and show that the optimal design of leniency programs should be related to the frequency and effectivenss of random investigations and balance the relationship between the sticks and carrots in antitrust enforcement.

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

[1]We are grateful to Bruno Jullien, Michele Polo, and Giancarlo Spagnolo for their comments, as well as siminar participants in IDEI and Centre for Competition Policy.

record-breaking corporate fines. This success has encouraged many other countries or jurisdictions to adopt their own leniency programs.[2]

In spite of a great success in practice,[3] 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. There is a growing literature on leniency programs, and most of them, including the illuminating works like Spagnolo (2000), Aubert, Rey and Kovacic (2005), and Harrington (2008), investigate the impacts on the incentives for the cartel members to deviate from collusive agreements when a leniency program is introduced. These papers assume that cartels will restrict to a collusive strategy that requires members to collude forever and never approach for amnesty in whatever circumstances; in other words the literature assumes that cartels are unable to exploit from leniency programs by organizing strategic reporting as part of collusive strategies such as "collude and report systematically". This assumption is valid under enforcement environments where cartels will be under intensified scrutiny for quite a long period after exposed and are thus unlikely to be re-organized; or under legislative environments where the repeated offenders are not eligible for leniency and will be punished harshly; in both cases cartels are prevented to "collude and report systematically".

While introducing leniency programs brings stronger incentives for cartel participants to break the collusive agreements and moreover denounce cartels for amnesty, it also broadens the scope of feasible collusive strategies; in particular cartels may find beneficial to organize collusive strategies that involve strategic denouncing of cartels for amnesty if the antitrust authority grants quite generous amnesty. Usually, an exposed cartel will be under scrutiny for some periods, but the effectiveness of such a scrutiny is indeed doubtful. First of all the antitrust agency are subject to budget constraint and is unlikely to invest much resources into ex post monitoring on cracked cartels; second, cartel activities are conducted in a secret way and it is therefore suspicious whether such a scrutiny can really deter collusive activities. Moreover, restricting leniency only to first-time offenders allows cartels to adopt the "report once and never report after" strategy which renders the leniency programs completely ineffective (see Proposition 2), and this rule should be cautiously reviewed by the antitrust authority. Finally, although repeated offenders will be punished harshly, the increased fines will be still subject to the rule of maximum penalty as cartel firms are protected by the limited liability and the antitrust authority is unlikely to push these firms into bankruptcy.

Therefore any theoretical research related to the optimal design of leniency programs will face a great risk of overestimating the effectiveness of leniency programs if the possibilities that a generous amnesty policy may be exploited strategically by cartels are excluded without reasonable justifications on enforcement and legislative environments, and the robustness of its basic results as well as policy implications are thus questionable. This concern is further supported by the recent evidence from laboratory experiments by Hinloopen and Soetevent (2008) which shows that collusive strategies like "collude and report systematically" can arise in equilibrium when the leniency programs are too generous, where reporting the cartel becomes part of the collusive agreement and participants then collude and apply for leniency in every period.

[2] 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.

[3] See Hammond (2005).

In this paper, we investigate the design of leniency programs by taking into account the possibility that cartels can exploit from strategic use of leniency, following the logic of the pioneering work by Motta and Polo (2003). Granting amnesty to cartel members encourages them to report cartel 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, more robust whenever reporting the cartel can be used strategically as part of the collusive agreement. As we will see, the trade-off between these two conflicting forces determines the optimal level of leniency.

This paper then looks for the optimal amnesty rates, both before and after an investigation is started, 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. To study the effectiveness of leniency programs, we consider an environment where 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 (that is, collude and never report), by encouraging firms to deviate and denounce the cartel; and (ii) discouraging firms from exploiting the leniency program through colluding and reporting strategies. Our simple framework allows us to relate the optimal leniency rates (the "carrot"), which is the solution of the trade-off just mentioned, to the effectiveness of random investigations (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 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[4], 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.

The revision of the US amnesty program in 1993 has two main innovations in leniency policy as mentioned above. While the effectiveness of restricting leniency only to the first informant is widely recognized in the existing literature, there is few analysis concerning about the effectiveness of the new rule that allows the first informant to remain eligible even after an investigation is underway. This paper is, to our knowledge, the first attempt to provide a normative analysis on the effect of offering leniency post-investigation, and we show that whether offering post-investigation amnesty is optimal depends on the effectiveness of investigation.

The simple model also allows us to compare the different variants of policies, and gives some interesting policy implications. The most surprising is that offering no leniency for repeated offenders will cause a countervailing effect which can make the leniency policy completely ineffective; this calls for a cautious use of heavy sticks.

[4]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.

This paper builds on the recent literature on leniency programs. In particular it is closely related to Motta and Polo (2003) which analyzes 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 take the leniency rates as given and study the deterrence as well as desistance effects of the amnesty program. Taking into account the possibility of cartels' strategic exploit from generous leniency policy, they find that the introduction of a leniency program brings two conflicting effects as mentioned above and it would be difficult to conclude that a leniency program unambiguously increase welfare. In contrast, our simple model allows us to characterize the optimal degree of leniency and show that both pre-and post-investigation leniency can be helpful to prevent the formation of some cartels and unambiguously increase welfare. On the other hand, they investigate the most effective way to allocate antitrust resources between preliminary investigation and prosecution; however we take here the likelihood of investigations and successful prosecution as given, and study the optimal design of leniency policy; but our framework also allows us to investigate the optimal allocation of enforcement resources under budget constraint, which can determine endogenously the optimal investigation rate and the relevant effectiveness of investigation.

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 (2008) characterizes the leniency program in a framework that allows the probability of discovery and successful prosecution to change over time. He points out that offering leniency can trigger a "Race-to-the-courthouse" when detection becomes likely, which in turn increases the expected penalties from engaging in cartel activity; he also shows that it is optimal to restrict eligibility to the first informant and also often optimal (assuming away positive rewards) to grant full leniency to that first informant. Harrington and Chang (2008) 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. We also characterize the optimal allocation of enforcement resource subject to the budget constraint of the antitrust agency in section 5, and finally conclude in section 6.

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.[5]

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,

[5] 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.

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.[6] To keep the analysis simple, we assume that the evidence of collusion is generated only if both firms agree on collusion and it lasts only for one period, which implies that the cartel cannot be prosecuted for its past activity.

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

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

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

[6] We assume here that the fine F is independent from the stakes from collusion; this is a modelling device that makes it easier to handle than just to assume heterogeneity in the relative benefits of deviation. 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 anti-competitive behavior, and thus, possibly, to the stakes from collusion. 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% of the turnover in EU and $100 million in the US) and also driven by other considerations. For instance, the Commision (see European Commision (2006)) determines a first amount based on the value of sales affected by the collusion and on the number of years of infringement. It may then adjust that amount "on the basis of an overall assessment which takes account of all the relevant circumstances." To ensure that fines have a sufficiently deterrent effect, the Commission may moreover "increase the fine to be imposed on undertakings which have a particularly large turnover beyond the sales of goods or services to which the infringement relates."

[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).

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.

3. OPTIMAL LENIENCY UNDER SECRET INVESTIGATIONS

3.1. Collusive Strategies

We now introduce a leniency program, which allows the *first informant* (and only the first one) to benefit from a reduced fine $(1-q)F$ (or even from a positive reward, if $q > 1$). As we will see, leniency makes "normal" collusion more difficult, but also 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:

$$B \geq B_N^r(q) \equiv \frac{\rho - (1-\delta)(1-q)}{2\delta - 1} F, \tag{4}$$

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

Alternative collusive strategies.

Firms may however 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 corresponds to a polar case of enforcement environments that cartels can reorganize after exposition without any delay. In reality, one would expect the antitrust agency to keep such an industry under close scrutiny, making it difficult to collude for at least some time. Yet firms could start colluding later on and 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 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 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:

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

that is, whenever

$$B \geq B_R(q) \equiv \frac{\delta \left(1 - \frac{q}{2}\right) F}{2\delta - 1}. \tag{5}$$

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 $q > \bar{q}$.

3.2. Optimal amnesty rate

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

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

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.[9] We now seek to characterize the optimal degree of leniency. The antitrust authority aims to as many cartels as possible; the amnesty rate q should therefore maximize the deterrence threshold

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

which appears in bold in Figure 1.

[9] 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.

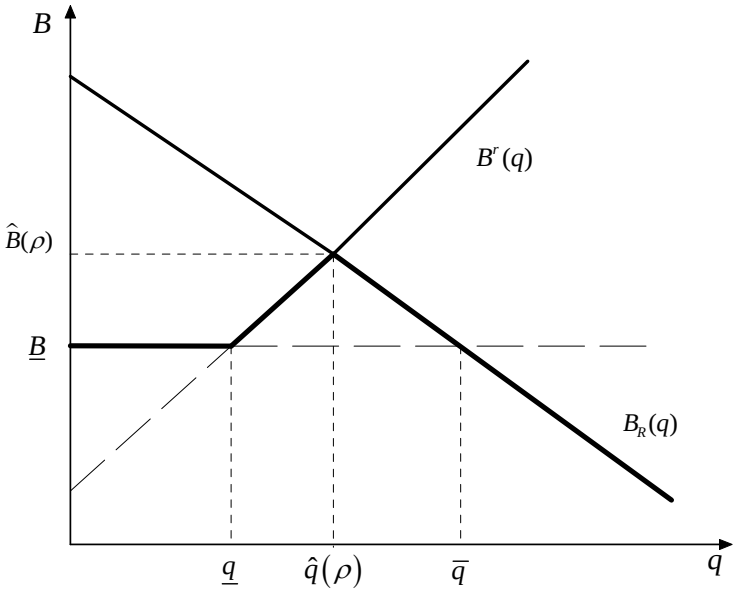

FIG. 1

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:

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

which is achieved for

$$q = \hat{q}(\rho) \equiv \frac{1-\rho}{1-\frac{\delta}{2}}. \tag{6}$$

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

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

$$\hat{B}(\rho) \equiv \frac{\delta(1 - \delta + \rho)}{(2\delta - 1)(2 - \delta)} F, \tag{7}$$

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

The following proposition summarizes the analysis:

PROPOSITION 1. *The optimal amnesty rate is determined so as to deter normal collusion, without encouraging collusion with reporting: it is characterized by* (6) *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,[10] 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: $B_N^r(\hat{q}) = B_R(\hat{q}) = \hat{B}(\rho)$.

The same trade-off drives the impact of random audits on the optimal amnesty rate: increasing the 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 ($B_N^r(q;\rho)$ moves up) in the marginal industry and neighboring ones (that is, for B slightly larger than $\hat{B}(\rho)$). A small reduction in the leniency rate q then deters also "collude and report" strategies, while still deterring normal collusion, in these additional industries.

Remark 1: 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 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). It does not really affect the outcome of the analysis in our current framework, but it could do so in more general contexts, by providing additional motivation for leniency.

Remark 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 "collude and report" more attractive and therefore more robust, which reduces the effectiveness of the leniency programme. The threshold of facilitating the "collude and report" strategy decreases when both informants can benefit from leniency:

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

[10] 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.

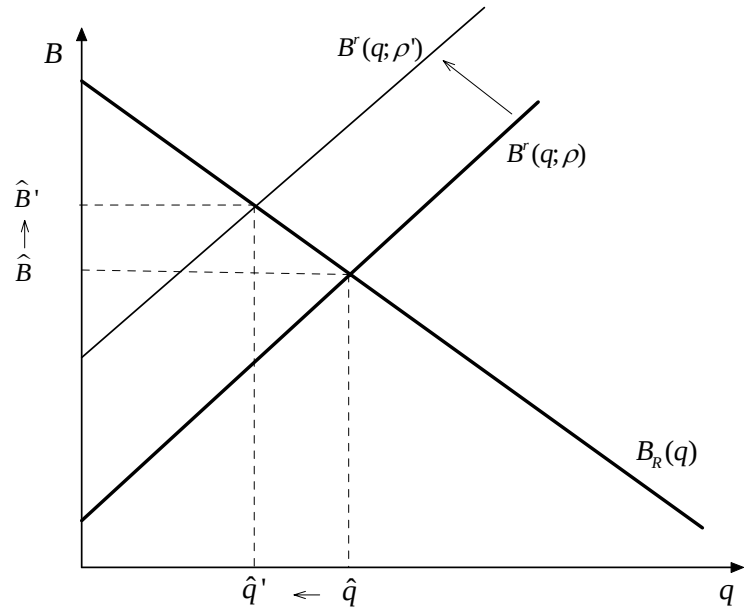

FIG. 2

which leads to a lower leniency rate $q = \underline{q}$ in equilibrium. As a result the equilibrium threshold decreases to $\underline{B}$, so it is actually optimal to offer no leniency in this case. The leniency program thus performs less well in the absence of "first-informant-wins" rule; this result may explain why the original version of the US leniency program did not contribute much to defeat cartels before the 1993 revision.[11]

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;[12] 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

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

[12] 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.

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:

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

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 2. *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. Exploitable or Non-exploitable Leniency Programs? A Discussion

The existing literature on leniency programs diversifies along two directions. Most of the literature assume that cartels cannot exploit from a generous leniency program by organizing strategic reporting as part of collusive strategies, and thus cartels can restrict only to "collude and never report" strategy even when the leniency programs are quite generous. Quite intuitively, the existing literature that excludes the possibility of strategic exploit of generous amnesty would conclude that

granting more generous leniency and even rewards to the informant can contribute more to deterrence, and it is thus optimal to give a fines-financed bounty to the first reporting firm; this is, for instance, argued by Spagnolo (2000). As we have addressed in the introduction that this assumption can be validated only under very restrictive enforcement environments which are unrealistic, and/or legislative environments that can make the leniency policy completely ineffective; therefore such conclusions might overestimate the effectiveness of leniency policy and as a result offering too generous leniency will make leniency program less effective. In contrast, we take into account the possibility that leniency programs might be exploited strategically, and appeal for a balanced use of carrots and sticks.

One may argue that the "collude and report" strategy sounds far-fetched, since the cartel would then be systematically denounced and yet go on forever and one never find empirically relevant evidence on such a collusive strategy. This is indeed a misunderstanding that matches the theoretical framework into the real world cases in an incorrect way. First of all, "collude and report systematically" will never arise on the equilibrium path in our model under optimal leniency programs. To see this, note that for firms with collusive stakes less than the threshold $\widehat{B}$, collusion cannot be sustainable and one would expect that these firms are deterred from collusion; for those firms with $B \geq \widehat{B}$, however, both collusive strategies are sustainable but cartels strictly prefer "normal collusion" as it brings more stakes:

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

When leniency is too generous, that is, $q > \widehat{q}$, then cartels with $B_R(q) < B < B_N^r(q)$ prefer "collude and report" strategy. This theoretical result is strongly supported by the evidence from recent laboratory experiments by Hinloopen and Soetevent (2008). They compare two treatments with $\rho = 0.4$ and $\delta = 0.8$, in which case the optimal leniency rate should be set $\widehat{q} = 1$ according to our model. In the first treatment the amnesty rate is $q = 1$ (that is $q = \widehat{q}$) whilst the equivalent amnesty rate is $q = 1.8 (> \widehat{q})$ in the second treatment.[13] They find that players never adopt the "collude and report" strategy in the first treatment, while more than 70% of subjects play "collude and report" in every period in the second treatment! The theoretical prediction of our model coincides exactly with the experimental results.

Secondly, empirical observations on leniency applications are far from complete due to relatively short history of leniency programs. In addition, most antitrust agencies grant leniency that gives a reduction of fines up to 100%, and few leniency program grants reward to the firms approaches leniency successfully. In other words, the current leniency programs are not abused according to our analysis, so one would expect that the risk that leniency program will be strategically exploited is relatively low.

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

[13] In the second treatment, the leniency program grants each player 90% reduction of fine when both players report, but gives 100% reduction of fine when only one player reports (which is the case of deviation). This rule encourages the player to adopt "collude and report" rather than collude and never report.

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 $0 < p < 1$.

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

4.1. Open or secret investigations?

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).

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[14] $\hat{q}(\alpha p)$ 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

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

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:

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

- 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.

Making investigations public creates additional forms of collusion, since firms can try to abuse the program by reporting for example only when an investigation is launched. However, the antitrust agency can also adjust the amnesty rate once it launches an investigation, and this actually allows antitrust enforcement to remain as effective as with secret investigations.

To see this, suppose that the agency grants no leniency once an investigation is started (i.e., $q_a = 0$). Then, firms cannot benefit from reporting the cartel once an investigation is underway, since doing so would increase the probability of prosecution (from p to 1), without any reduction in the fine. Thus, cartel members' only relevant choice is between "never reporting" and "reporting before an investigation is launched". But this choice is essentially the same as the one they face (between "normal collusion" and "collude and report") when investigations are launched secretly, and thus the antitrust agency can still perform as well as with open investigations as it can with secret ones.

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

4.2. Feasible Collusive Strategies

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

4.2.1. 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 four types of defection, which we successively consider: deviating in the product market and reporting at Stage 4 in case of an investigation, deviating and reporting at Stage 2 before there may be an investigation, colluding but reporting in case of an investigation, and deviating without reporting.

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

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

or equivalently:

$$B \geq B_N^a(q_a) \equiv \frac{\alpha p - \alpha\left(1-\delta\right)\left(1-q_a\right)}{2\delta - 1}F. \tag{8}$$

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

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

or:

$$B \geq B_N^b(q_b) \equiv \frac{\alpha p - \left(1-\delta\right)\left(1-q_b\right)}{2\delta - 1}F. \tag{9}$$

Another relevant defection from the normal collusive strategy is that the cartel firms do not deviate in the market but denounce the cartel when an investigation is launched. The "collude-but-report-after-investigation" strategy brings less present benefit of defection but will generate more value in the future since cartel firm can still go on collusion if there is no investigation, which occurs with probability $1-\alpha$. So normal collusion can be immune to such kind of defection only if

$$V_N \geq B - \alpha\left(1 - q_a\right)F + (1-\alpha)\delta V_N,$$

or equivalently

$$B \geq B_N^c(q_a) \equiv \frac{p(1-\delta+\alpha\delta) - \left(1-\delta\right)\left(1-q_a\right)}{\delta}F. \tag{10}$$

The "collude-but-report-after-investigation" strategy is more attractive only if the investigation rate is low, in which case the cartel members would expect that collusion can sustain with high likelihood, or the post-investigation leniency rate is high so that the cartel members have incentives to report when investigation is underway. To see this, comparing $B_N^c(q_a)$ with $B_N^a(q_a)$ we get

$$B_N^a(q_a) > B_N^c(q_a) \text{ if } \alpha \geq \alpha_0 \equiv \frac{2\delta - 1}{\delta},$$

and

$$B_N^c(q_a) \geq B_N^a(q_a) \text{ if } \alpha < \alpha_0 \text{ and } q_a \geq 1 - \frac{(2\delta-1) - 2\alpha\delta}{(2\delta-1) - \alpha\delta}p.$$

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)[15] if:

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

that is, when:

$$B \geq B_N^d \equiv \frac{\delta\alpha pF}{2\delta - 1}. \tag{11}$$

Hence, normal collusion is sustainable if and only if[16]:

$$B \geq B_N(q_b, q_a) \equiv \max\left\{B_N^a(q_a), B_N^b(q_b), B_N^c(q_a), B_N^d\right\}. \tag{12}$$

[15] This is the case when

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

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

[16] There is also one possible deviation strategy, "collude but report before investigation" which is dominated by "deviate and report before investigation" as easy to check.

4.2.2. 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 and reporting after investigation.[17]

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:

$$B \geq B_A^b(q_b, q_a) \equiv \frac{\alpha\left(1 - \frac{q_a}{2}\right) - (1-\delta)(1-q_b)}{2\delta - 1}F. \tag{13}$$

Similarly, deviating and reporting after investigation is not profitable if:

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

or:

$$B \geq B_A^a(q_a) \equiv \frac{\delta\alpha\left(1 - \frac{q_a}{2}\right)}{2\delta - 1}F. \tag{14}$$

Hence, this collusion is sustainable if and only if:

$$B \geq B_A(q_b, q_a) \equiv \max\left\{B_A^a(q_a), B_A^b(q_b, q_a)\right\}. \tag{15}$$

4.2.3. 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:

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

that is, when:

$$B \geq B_B(q_b) \equiv \frac{\delta\left(1 - \frac{q_b}{2}\right)F}{2\delta - 1}. \tag{16}$$

[17] Other possible defections include the "collude-but-report-before-investigation" strategy which is dominated by "deviate-and-report-before-investigation" strategy, as well as "deviate-but-never-report" strategy which is dominated by "deviate-and-report" strategy.

4.3. Optimal leniency policy

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

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

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 $B_N^a(q_a)$ and $B_N^c(q_a)$ increase 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: $B_A(q_b, q_a) = \max\left\{B_A^a(q_a), B_A^b(q_b, q_a)\right\}$ decreases as q_a increases, since both B_A^b and B_A^a 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., $B_N(q_b, q_a) = \max\{B_N^a(q_a), B_N^c(q_a)\} > B_N^d, B_N^b(q_b)$;
- and offers little leniency pre-investigation:

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

  implying:

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

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 further "collude and report in case of investigations" strategies, without excessively triggering "collude and report systematically" ones. It may also further deter normal collusion (if $B_N^b(q_b') > \max\{B_N^a(q_a), B_N^c(q_a)\}$, otherwise a slight reduction in the post-investigation amnesty rate q_a also increases $B_N(q_b', q_a) = \max\{B_N^a(q_a), B_N^c(q_a)\}$ 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.

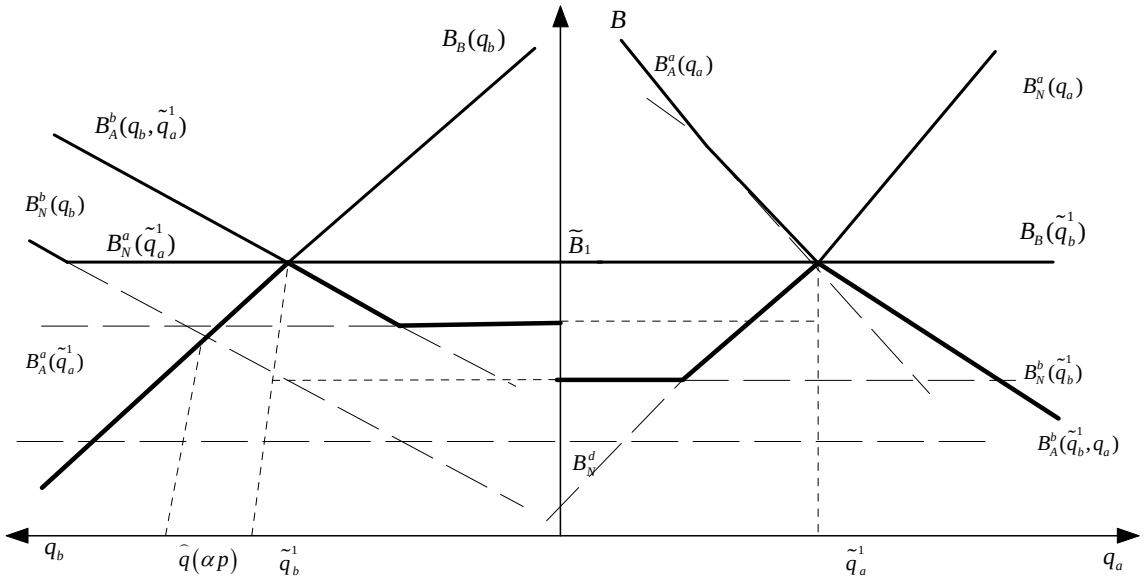

FIG. 3

Given this, when post-investigation leniency is used to deter normal collusion, and since B_A^b decreases while $\max\{B_N^a(q_a), B_N^c(q_a)\}$ increases with q_a, the best amnesty rate q_a is such that these two thresholds coincide: $B_A^b(q_b, q_a) = \max\{B_N^a(q_a), B_N^c(q_a)\}$. Similarly, since $B_B(q_b)$ decreases while $B_A^b(q_b, q_a)$ increases with q_b, the best amnesty rate q_b is such that these two thresholds also coincide, as illustrated by Figure 3.

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

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

which leads to

$$\begin{aligned}\tilde{B}_1 &= B_N^a\left(\tilde{q}_a^1\right) = B_A^b\left(\tilde{q}_b^1, \tilde{q}_a^1\right) = B_B\left(\tilde{q}_b^1\right), \text{ if } B_N^a(\tilde{q}_a) > B_N^c(\tilde{q}_a),\\ \tilde{B}_2 &= B_N^c\left(\tilde{q}_a^2\right) = B_A^b\left(\tilde{q}_b^2, \tilde{q}_a^2\right) = B_B\left(\tilde{q}_b^2\right), \text{ otherwise.}\end{aligned}$$

So they are given by

$$\tilde{q}_b^1(\alpha, p) = \frac{2\left[(1-\alpha)(2-\delta) + \alpha(1-p)\right]}{2(1-\delta)^2 + (2-\delta)}, \tag{17}$$

$$\tilde{q}_a^1(\alpha, p) = \frac{2\left[\delta(1-\delta)(1-\alpha) + \alpha(1-p)(2-\delta)\right]}{\alpha\left[2(1-\delta)^2 + (2-\delta)\right]}, \tag{18}$$

while the resulting deterrence threshold is:

$$\tilde{B}_1(\alpha, p) = \frac{\delta}{(2\delta - 1)} \frac{\left[2(1-\delta)^2 + \alpha(1-\delta+p)\right]}{\left[2(1-\delta)^2 + (2-\delta)\right]} F; \tag{19}$$

and

$$\tilde{q}_b^2(\alpha,p) = \frac{2\left[2(1-\delta)(2\delta-1)+\alpha(3\delta^2-3\delta+1)-(2\delta-1)(1-\delta+\alpha\delta)\alpha p\right]}{2(2\delta-1)(1-\delta)(2-\delta)+\alpha\delta^2} \quad (20)$$

$$\tilde{q}_a^2(\alpha,p) = \frac{2\left[(1-\delta)(5\delta-\delta^2-2)+\alpha\delta^2-(2\delta-1)(2-\delta)(1-\delta+\alpha\delta)p\right]}{2(2\delta-1)(1-\delta)(2-\delta)+\alpha\delta^2}, \quad (21)$$

with the resulting threshold

$$\tilde{B}_2(\alpha,p) = \frac{\delta\left[2(1-\delta)^2+\alpha(1-\delta)+(1-\delta+\alpha\delta)\alpha p\right]}{2(2\delta-1)(1-\delta)(2-\delta)+\alpha\delta^2}F. \quad (22)$$

Note that, in the absence of post-investigation leniency, the deterrence threshold is given by

$$\hat{B}(\alpha p) \equiv \frac{\delta(1-\delta+\alpha p)}{(2\delta-1)(2-\delta)}F. \quad (23)$$

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

$$p \leq \tilde{p}_1(\alpha) \equiv \frac{\alpha(2-\delta)-\delta}{2\alpha(1-\delta)}, \quad (24)$$

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

$$p \leq \tilde{p}_2(\alpha) \equiv \frac{(1-\delta)(3\delta-2)}{(2\delta-1)(2-\delta)+2\alpha\delta(1-\delta)}, \quad (25)$$

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

$$p \leq \tilde{p}_3(\alpha) \equiv \frac{2(1-\delta)(2\delta-1)+\alpha\left(2\delta^2-1\right)-\alpha^2\delta}{\alpha\left[(2\delta-1)(3-2\delta)+4\alpha(1-\delta)^2\right]}. \quad (26)$$

As $\tilde{p}_2(\alpha) > 0$ if and only if $\delta > 2/3$, two cases should be taken into account:

Case (1): $1/2 < \delta \leq 2/3$. The cartel firms are impatient and so that the future stakes of collusion is less valuable. In this case, leniency policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ is dominated by policy $(\widehat{q}(\alpha p), 0)$, and the later is dominated by $(\tilde{q}_b^1, \tilde{q}_a^1)$ if and only if $p \leq \tilde{p}_1(\alpha)$, as illustrated in Figure 4.

Case (2): $\delta > 2/3$. In this case, "collude and report after investigation" becomes a relevant defection from normal collusion, and the policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ is optimal if $p \leq \tilde{p}_3(\alpha)$ and $p \leq \tilde{p}_2(\alpha)$, as depicted by Region C in Figure 5. This will be the case when the investigation rate is relatively low ($\alpha < \alpha_0$), and cartels would prefer "collude and report after investigation" rather than "deviate and report after investigation" when they intend to denounce the cartel, as the later strategy will result to the collapse of cartel definitely even when the investigation happens with very low probability. On the other hand, when the investigation rate is relatively high, then the "deviate and report after investigation" becomes a dominant strategy when cartel firms intend to defect, by which they can acquire full stakes of collusion which overweighs the future benefit of ongoing collusion as they expect that investigations will occur quite likely. In this case, policy $(\tilde{q}_b^1, \tilde{q}_a^1)$ is optimal if $\tilde{p}_3(\alpha) \leq p \leq \tilde{p}_1(\alpha)$, which is depicted by Region B in Figure 5. Finally, when the investigation is likely to be successful, they it will be optimal to restrict leniency

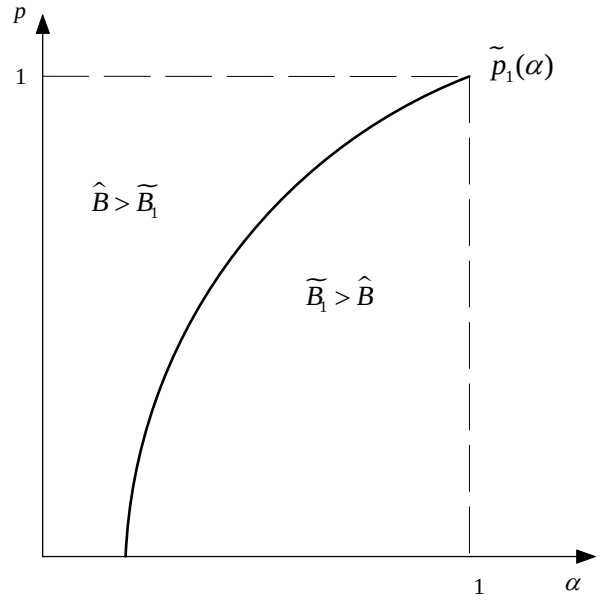
p
1
$\tilde{p}_1(\alpha)$
$\hat{B} > \tilde{B}_1$
$\tilde{B}_1 > \hat{B}$
1
α

FIG. 4

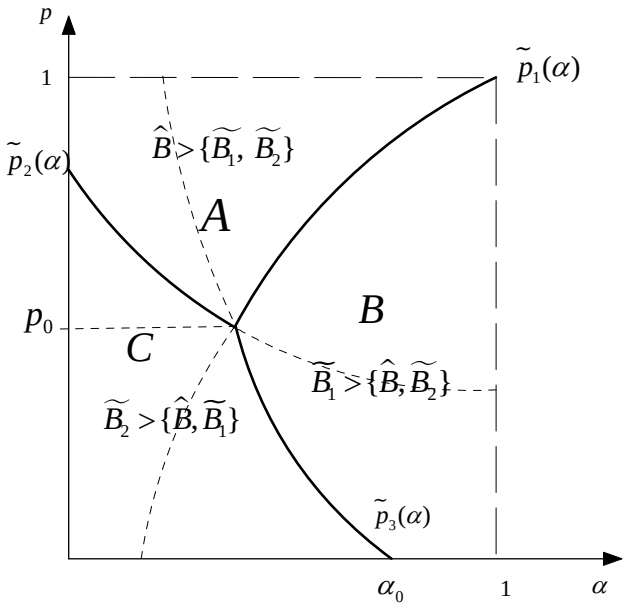
p
1
$\tilde{p}_1(\alpha)$
$\tilde{p}_2(\alpha)$
$\hat{B} > \{\tilde{B}_1, \tilde{B}_2\}$
A
B
p_0
C
$\tilde{B}_1 > \{\hat{B}, \tilde{B}_2\}$
$\tilde{B}_2 > \{\hat{B}, \tilde{B}_1\}$
$\tilde{p}_3(\alpha)$
α_0
1
α

FIG. 5

only before investigation, which is the case either $p \geq \tilde{p}_2(\alpha)$ when the investigation rate is low, or $p \geq \tilde{p}_1(\alpha)$ when investigations are launched frequently; the optimal policy is depicted in Region A.

Summarizing the above analysis results to the main conclusions:

PROPOSITION 3. *It is always optimal to offer leniency before investigations; moreover:*

- *When cartels are less patient ($\delta \leq 2/3$), it is optimal to offer amnesty also when an investigation is already underway if $p \leq \tilde{p}_1(\alpha)$ and the optimal policy is then $(q_b^*, q_a^*) = \left(\tilde{q}_b^1(\alpha, p), \tilde{q}_a^1(\alpha, p)\right)$;*
- *Whenever cartels are more patient ($\delta > 2/3$), it is optimal to offer post-investigation leniency if $p \leq \tilde{p}_1(\alpha)$ or $p \leq \tilde{p}_2(\alpha)$; the optimal policy is then given by $(q_b^*, q_a^*) = \left(\tilde{q}_b^1(\alpha, p), \tilde{q}_a^1(\alpha, p)\right)$ if $\tilde{p}_1(\alpha) \geq p \geq \tilde{p}_3(\alpha)$ and $(q_b^*, q_a^*) = \left(\tilde{q}_b^2(\alpha, p), \tilde{q}_a^2(\alpha, p)\right)$ otherwise;*
- *The optimal deterrence thresholds in these two cases are given by $B^* = \{\tilde{B}_1(\alpha, p), \tilde{B}_2(\alpha, p)\}$ respectively.*

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*). 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, let p_0 denote the value of p such that $\tilde{p}_1(\alpha) = \tilde{p}_2(\alpha)$, then it is always optimal to grant some amnesty post-investigation if $p \leq p_0$. 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.

4.4. 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., $p > \{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$), 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., $p < \{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$), 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.

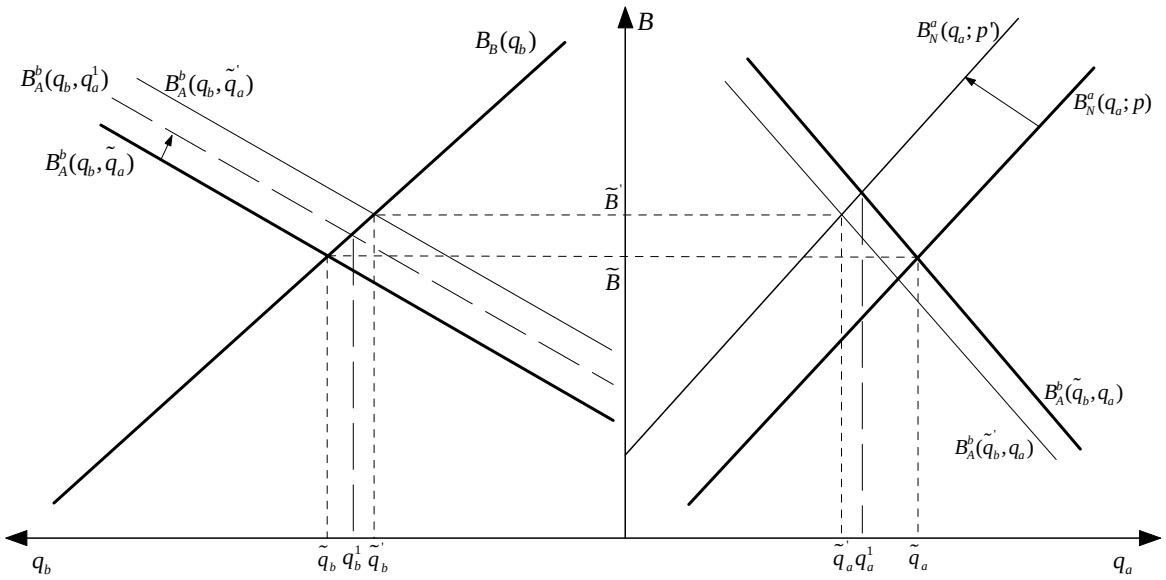

FIG. 6

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 $B_N^a(q_a)$ and $B_N^c(q_a)$, as illustrated in Figure 6; 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.

(Insert Figure 6 here.)

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 composite, due to the fact that decreasing q_a, say, weakens "collude and report After investigation" but strengthens normal collusion. When α is relatively high and thus "deviating and reporting after investigation" becomes a relevant defection to normal collusion, the increase in the frequency α and the successively decrease in $\tilde{q}_b$ have an overall relatively larger effect on normal collusion, and so $\tilde{q}_a$ also decreases; when instead the investigation rate is relatively low, the increase in α and the successively decrease in $\tilde{q}_b$ have an overall smaller impact on normal collusion, in this case it is optimal

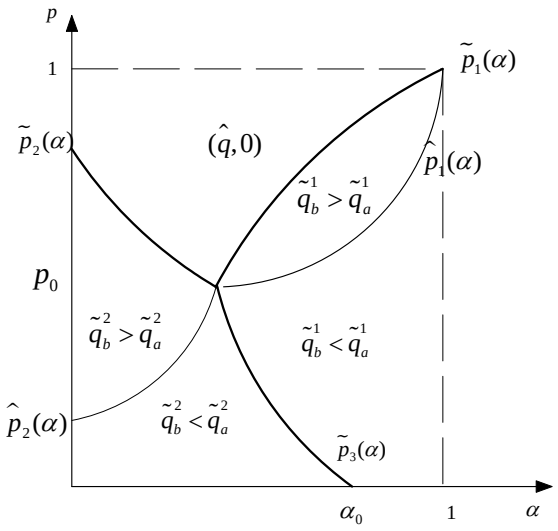

FIG. 7

to increase $\tilde{q}_a$ such that cartel members are induced to "collude and report after investigation" which destabilizes normal collusion.

The above comparative statics is concluded as follows:

PROPOSITION 4. *Increasing p or α makes the leniency program more effective, moreover:*

- *Increasing p allows to decrease the leniency rates pre- and post-investigation;*
- *Increasing α allows to decrease the leniency rates pre-investigation and post-investigation $\tilde{q}_a^1$ but will increase $\tilde{q}_a^2$.*

Proof. See appendix B. ∎

It is interesting to see whether launching an investigation should lead to offer more or less leniency. To this end, we need to compare the leniency rates pre- and post-investigation in different regimes.

The detailed analysis is presented in Appendix C, which shows that it is optimal to offer more leniency post-investigation when p is relatively lower:

$$\tilde{q}_b^1 \geq \tilde{q}_a^1 \text{ if and only if } p \geq \widehat{p}_1(\alpha) \equiv \frac{\left(\alpha^2 + \delta(1-\alpha)\right)(1-\delta)}{\alpha\,(2-\delta-\alpha)},$$

$$\tilde{q}_b^2 \geq \tilde{q}_a^2 \text{ if and only if } p \geq \widehat{p}_2(\alpha) \equiv \frac{(1-\delta)\left[\delta\,(1-\delta) + \alpha\,(2\delta-1)\right]}{(2\delta-1)\,(1-\delta+\alpha\delta)(2-\delta-\alpha)},$$

as illustrated in Figure 7 and concluded by the following proposition:

(Insert Figure 7 here.)

PROPOSITION 5. *It is optimal to offer more amnesty post-investigation than pre-investigation if and only if $p < \widehat{p}_1(\alpha)$ or $p < \widehat{p}_2(\alpha)$.*

Proof. See Appendix C. ∎

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 relatively unlikely to succeed in the absence of self-reporting.

5. OPTIMAL ALLOCATION OF ENFORCEMENT RESOURCES

So far as we have assumed that the enforcement measures, as characterized by the investigation rate α and likelihood of detection p, are given exogenously; this is due to the budget and technology constraints in antitrust enforcement. Taking any efforts to increase the probability of investigation which requires to investigate targeted industries more frequently would shorten the duration of investigation and thus decrease the likelihood of uncovering cartel activities, given that the total amount of labors invested in investigation are subject to budget constraint. However, as the comparative statics shows that increasing α and/or p brings different effects on deterrence, it is therefore necessary to examine the optimal allocation of enforcement resources which aims to maximize the deterrence thresholds under budget constraints.

To fix ideas, let L denote the total amount of labors engaged in cartel investigations that is fixed due to budget constraint. There are N targeted industries for investigations, and the labors contributed to each investigation l is inversely related to the frequency of investigation as given by[18]

$$l = \frac{L}{\alpha N}.$$

A minimum scale of labors l_0 is needed in order to launch an investigation and the likelihood of success p increases with the investment of labors in an investigation l, but there is decreasing returns to scale for labors. This relationship is characterized by the following function:

$$\begin{aligned} p &= (l - l_0)^{\frac{1}{b}}, \text{ with } b > 1, \text{ if } l > l_0; \\ p &= 0, \text{ if } l \leq l_0. \end{aligned}$$

or equivalently

$$l = l_0 + p^b;$$

and we can now rewrite the budget constraint as

$$\frac{L}{N} = \alpha l = \alpha \left(l_0 + p^b \right).$$

The optimal enforcement policy (α, p) can be derived by maximizing the thresholds of deterrence subject to the budget constraint (29); and using the first-order-conditions yield the following equation:

$$R(B^*) \equiv \frac{\partial B^* / \partial \alpha}{\partial B^* / \partial p} = \frac{l_0 + p^b}{\alpha b p^{b-1}}, \tag{27}$$

[18] The logic of modelling budget constraint is exactly the same as Motta and Polo (2003).

which reflects the balance between the ratio of marginal contribution to deterrence (as represented in LHS) and the marginal costs (denoted by RHS) for α and p at the optimum.

In particular when $\delta \leq 2/3$, the Regime C that corresponding to leniency policy $(\tilde{q}_b^2, \tilde{q}_a^2)$ is irrelevant, and we can figure out the optimal (α, p) in simple expressions. Whenever $p > \widehat{p}_1(\alpha)$ and thus $B^* = \widehat{B}$, the first-order condition can be expressed by

$$R(\widehat{B}) = \frac{p}{\alpha} = \frac{l_0 + p^b}{\alpha b p^{b-1}},$$

and the optimal policy is thus given by

$$p = \left(\frac{l_0}{b-1}\right)^{\frac{1}{b}}, \ \alpha = \frac{L(b-1)}{bNl_0}. \tag{28}$$

Whenever $p < \widehat{p}_1(\alpha)$ such that $B^* = \widetilde{B}_1$, the optimal allocation of resources is determined by

$$R(\widetilde{B}_1) = \frac{1-\delta+p}{\alpha} = \frac{l_0 + p^b}{\alpha b p^{b-1}},$$

and the optimal enforcement policy is given by

$$\begin{aligned} (b-1)p^b + (1-\delta)bp^{b-1} &= l_0, \\ \alpha b p^{b-1}(1-\delta+p) &= \frac{L}{N}. \end{aligned} \tag{29}$$

The above analysis is summarized as follows:

PROPOSITION 6. *The optimal allocation of enforcement resources is governed by (27) ; in particular the optimal enforcement policies are given by (28) to (29) corresponding to different leniency policies when $\delta \leq 2/3$.*

The analysis shows that the optimal allocation of enforcement resources can also be analyzed under this framework. However, this paper aims to the optimal design of leniency programs and therefore will not go through in details to resolve the optimal allocation of enforcement resources, we leave this as a research work in the future.

6. 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.

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.

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, as well as the optimal allocation of enforcement resources under budget constraint; and these topics are left for the future research works.

Appendices

Appendix A: Proof of Proposition 3

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

(2) Suppose $B_N^* > B_A^* = B_B^*$. Then $B_N^b(q_b^*, q_a^*) > B_A(q_b^*, q_a^*) = \max\{B_A^a(q_a^*), B_A^b(q_b^*, q_a^*)\} = B_B(q_b^*) = B^*$, 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) $B_A(q_b^*, q_a^*) > B^* = B_B(q_b^*) = \max\{B_N^a(q_a^*), B_N^c(q_a^*)\} \geq B_N^b(q_b^*)$. Then decreasing q_b and increasing q_a would increase B_B and $B_N = \max\{B_N^a, B_N^c\}$, a contradiction.
2) $B_A(q_b^*, q_a^*) > B^* = B_B(q_b^*) = B_N^b(q_b^*) \geq \max\{B_N^a(q_a^*), B_N^c(q_a^*)\}$. Since $\max\{B_N^a(q_a^*), B_N^c(q_a^*)\}$ and $B_A(q_b^*, q_a)$ respectively increase and decrease with q_a, there thus exists $q_a^0 > q_a^*$ such that $\max\{B_N^a(q_a^*), B_N^c(q_a^*)\} = B_A(q_b^*, q_a)$. Two subcases need to be considered:

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

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

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

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

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

3) $B_N^* = \max\{B_N^a(q_a^*), B_N^c(q_a^*)\} > B_N^b(q_b^*)$ and $B_A^* = B_A^a(q_a^*) > B_A^b(q_b^*, q_a^*)$. 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 $B_N^b(q_b) > B^* = \max\{B_N^a(q_a^*), B_N^c(q_a^*)\}$, and thus $B_N, B_B > B^*$; B_A could then be also increased by decreasing q_a, a contradiction.
- or leads to $B_A^b(q_b^0, q_a^*) > B^* = B_A^a(q_a^*)$, 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 (q_b^1, q_a^*) is also optimal and moreover satisfies $B_N(q_b^1, q_a^*) = B_A(q_b^1, q_a^*) = B_B(q_b^1) = B^*$.

Q.E.D. ∎

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

(1) Situation 1: $B(q_b, q_a) = B_N^b(q_b) = B_A(q_b, q_a) = B_B(q_b) \geq B_N^a(q_a), B_N^c(q_a), B_N^d$. 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

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

The resulting deterrence threshold is equal to

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

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 $B_N^a(q_a)$ and $B_N^c(q_a)$, but has no impact on $B_N(\hat{q}_b, q_a) = B_N^b(\hat{q}_b) \geq \max\{B_N^a(q_a), B_N^c(q_a)\}$, and increase $B_A(\hat{q}_b, q_a)$. Furthermore, for $q_a = 0$ and $q_b = \hat{q}_b$, we have:

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

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

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

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: $\max\{B_N^a(q_a), B_N^c(q_a)\} = B_A^a(q_a) = B_B(q_b) \geq B_N^b(q_b), B_A^b(q_b, q_a), B_N^d$. The rate q_b therefore satisfies:

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

and thus:

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

but this implies

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

a contradiction.

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

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

that is:

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

and they are thus equal to:

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

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

$$\tilde{B}_1(\alpha, p) = \frac{\delta}{(2\delta - 1)} \frac{\left[2(1-\delta)^2 + \alpha(1 - \delta + p)\right]}{\left[2(1-\delta)^2 + (2-\delta)\right]} F.$$

Moreover we have $B_N^a(q_a) \geq B_A^a(q_a)$ if and only if $q_a \geq \frac{2(1-p)}{2-\delta}$, so $B_N^a\left(\tilde{q}_a^1\right) > B_A^a\left(\tilde{q}_a^1\right)$ as

$$\tilde{q}_a^1(\alpha, p) = \frac{2\delta(1-\delta)(1-\alpha)}{\alpha\left[2(1-\delta)^2 + (2-\delta)\right]} + \frac{2(1-p)(2-\delta)}{2(1-\delta)^2 + (2-\delta)} > \frac{2(1-p)(2-\delta)}{(2-\delta)^2}.$$

However, $B_N^b\left(\tilde{q}_b^1\right) < \tilde{B} = B_A^b\left(\tilde{q}_b^1, \tilde{q}_a^1\right)$ only when

$$\begin{aligned} p &< 1 - \frac{\tilde{q}_a^1}{2} \\ &= \frac{2\alpha(1-\delta)^2 - \delta(1-\delta)(1-\alpha) + \alpha p(2-\delta)}{\alpha\left[2(1-\delta)^2 + (2-\delta)\right]}, \end{aligned}$$

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

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

Conversely, when this condition is satisfied, we have:

$$B_N^b\left(\tilde{q}_b^1\right) < \tilde{B} = B_B\left(\tilde{q}_b^1\right),$$

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 $\tilde{q}_b^1 < \hat{q}(\alpha p)$ and thus:

$$\tilde{B}_1 = B_B\left(\tilde{q}_b^1\right) > B_B(\hat{q}(\alpha p)) = \hat{B}(\alpha p).$$

(4) Situation 4: $B_N^c(q_a) = B_A^b(q_b, q_a) = B_B(q_b) = \tilde{B} \geq B_N^b(q_b), B_A^a(q_a), B_N^d$. The optimal amnesty rates are then such that

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

and they are equal to:

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

with the resulting threshold

$$\tilde{B}_2(\alpha,p)=\frac{\delta\left[2(1-\delta)^2+\alpha(1-\delta)+(1-\delta+\alpha\delta)\alpha p\right]}{2(2\delta-1)(1-\delta)(2-\delta)+\alpha\delta^2}F.$$

By analogy, $\tilde{B}_2(\alpha,p)=B_B\left(\tilde{q}_b^2\right)>B_B\left(\hat{q}(\alpha p)\right)=\hat{B}(\alpha p)$ if and only if

$$\frac{\left[2(1-\delta)^2+\alpha(1-\delta)+(1-\delta+\alpha\delta)\alpha p\right]}{2(2\delta-1)(1-\delta)(2-\delta)+\alpha\delta^2}\geq\frac{(1-\delta+\alpha p)}{(2\delta-1)(2-\delta)},$$

or

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

and one may check that, when this condition holds, we have $B_N^b\left(\tilde{q}_b^2\right)<B_B\left(\tilde{q}_b^2\right)$ and $B_A^a\left(\tilde{q}_a^2\right)<B_A^b\left(\tilde{q}_b^2,\tilde{q}_a^2\right)$.

Note that, $\tilde{p}_2(\alpha)>0$ if and only if $\delta>2/3$, and it decreases with α in this case; and moreover we have

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

When $1/2<\delta\leq 2/3$, however, we have $\hat{B}(\alpha p)\geq\tilde{B}_2(\alpha,p)$ for any α and p and this implies that leniency policy $(\tilde{q}_b^2,\tilde{q}_a^2)$ is dominated by the policy $\hat{q}(\alpha p)$.

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

$$p\leq\tilde{p}_3(\alpha)\equiv\frac{2(1-\delta)(2\delta-1)+\alpha\left(2\delta^2-1\right)-\alpha^2\delta}{\alpha\left[(2\delta-1)(3-2\delta)+4\alpha(1-\delta)^2\right]};$$

We have $\tilde{p}_3(\alpha)$ decreases with α:

$$\begin{aligned}&\frac{d\tilde{p}_3}{d\alpha}\\ =\;&\frac{-\left[\delta(2\delta-1)+2\delta(1-\delta)^2(2\delta-1)+2(1-\delta)(3\delta-2)\right]\alpha^2-16(1-\delta)^3(2\delta-1)\alpha-2(1-\delta)(2\delta-1)^2(3-2\delta)}{\alpha^2\left[(2\delta-1)(3-2\delta)+4\alpha(1-\delta)^2\right]^2}\\ <\;&0,\quad\text{if }\delta>\frac{2}{3},\end{aligned}$$

and

$$\begin{aligned}\tilde{p}_3(0)&=+\infty,\ \tilde{p}_3(\alpha_0)=0,\\ \tilde{p}_3(1)&=\frac{(1-\delta)(2\delta-3)}{(2\delta-1)(3-2\delta)+4(1-\delta)^2}<0.\end{aligned}$$

As a result, post-investigation leniency can be useful when

$$p<\{\tilde{p}_1(\alpha),\tilde{p}_2(\alpha)\}.$$

Therefore:

- when $p<\{\tilde{p}_1(\alpha),\tilde{p}_2(\alpha)\}$ the optimal policy involves post-investigation leniency; it is given by $(q_b^*,q_a^*)=\left(\tilde{q}_b^1,\tilde{q}_a^1\right)$ when $p\geq\tilde{p}_3(\alpha)$ and $(q_b^*,q_a^*)=\left(\tilde{q}_b^2,\tilde{q}_a^2\right)$ when $p<\tilde{p}_3(\alpha)$, and the threshold of deterrence is given by $\tilde{B}_1(\alpha,p)$ and $\tilde{B}_2(\alpha,p)$ respectively.

- when instead $p > \{\tilde{p}_1(\alpha), \tilde{p}_2(\alpha)\}$ the optimal policy involves no post-investigation leniency; it is then given as before by $(q_b^*, q_a^*) = (\hat{q}(\alpha p))$ and its deterrence threshold is $\hat{B}(\alpha p)$.

Appendix B: Proof of Proposition 4

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:

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

and $\frac{\partial \hat{q}}{\partial p} = \alpha \hat{q}'(\alpha p) < 0$. Similarly, an increase in α also makes the antitrust policy more effective and thus increases $\tilde{B}$ and $\hat{B}$. We have moreover that both $\tilde{q}_b^1$ and $\tilde{q}_a^1$ decrease with α :

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

Moreover, we have

$$\frac{\partial \tilde{q}_b^2(\alpha,p)}{\partial \alpha} = \frac{-2(2\delta-1)\left[2(1-\delta)^3(3\delta-2)+\alpha\delta p\left(\alpha\delta^2+4(2\delta-1)(1-\delta)(2-\delta)\right)\right]}{\left[2(2\delta-1)(1-\delta)(2-\delta)+\alpha\delta^2\right]^2} < 0,$$

and

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

As $p \leq \tilde{p}_2(\alpha)$, then

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

Appendix C: Proof of Proposition 5

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

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

which is equivalent to

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

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

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

or

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

To characterize further properties of $\widehat{p}_1(\alpha)$ and $\widehat{p}_2(\alpha)$, note that $\widehat{p}_1(1)=1$ and

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

Moreover, as $\widehat{p}_1(\alpha)=\widehat{p}_2(\alpha)$ implies $\tilde{q}_b^1=\tilde{q}_a^1=\tilde{q}_b^2=\tilde{q}_a^2$, so $\widehat{p}_1(\alpha)$ and $\widehat{p}_2(\alpha)$ intersect at the point where $\tilde{p}_2(\alpha)$ and $\tilde{p}_1(\alpha)$ coincide: $\widetilde{B}_1=\widetilde{B}_2$ implies $\tilde{q}_b^1=\tilde{q}_b^2$ and $\tilde{q}_a^1=\tilde{q}_a^2$.

In addition, we have

$$\frac{d\widehat{p}_1(\alpha)}{d\alpha} = \frac{(1-\delta)\left[2\alpha^2(1-\delta)+2\alpha\delta-(2-\delta)\delta\right]}{\alpha^2(2-\delta-\alpha)^2},$$

and $\frac{d\widehat{p}_1(\alpha)}{d\alpha} \geq 0$ if and only if

$$2\alpha^2(1-\delta)+2\alpha\delta-(2-\delta)\delta \geq 0;$$

meanwhile

$$\begin{aligned} \frac{d\widehat{p}_2(\alpha)}{d\alpha} \quad &= \quad \frac{(1-\delta)\left[\alpha^2\delta(2\delta-1)+(1-\delta)^2\left((3\delta-2)+(1-\delta)\delta\right)\right]}{(2\delta-1)(1-\delta+\alpha\delta)^2(2-\delta-\alpha)^2} \\ &> \quad 0. \end{aligned}$$

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. (2008), "Optimal Corporate Leniency Programs", *Journal of Industrial Economics, (2): 215-246*

Harrington, J and M. Chang (2008), "Modelling the Birth and Death of Cartels with an Application to Evaluating Antitrust Policy", *Journal of European Economic Association, forthcoming.*

Hinloopen, J and A. Soetevent (2008) "From Overt to Tacit Collusion: Experimental Evidence on the Adverse Effects of Corporate Leniency Programs", *working paper.*

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

Rey, P. (2003), "Toward a theory of Competition Policy", in *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, M. Dewatripont, L. P. Hansen, S. J. Turnovsky eds, Cambridge University Press.

Spagnolo, G (2000), "Optimal Leniency Programs", working paper.

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