Cost Asymmetry and Market-Dividing Cartels: Implications for Leniency Programs*

Kebin Ma[†] Lucy White[‡]

Abstract

In a duopoly setting, this paper studies how cost asymmetry influences cartel stability and the effectiveness of leniency programs. It shows that when two firms collude to divide a market, the low-cost firm gains more from that collusion and has a stronger incentive to form a cartel than its high-cost conspirator does. When the cartel remains stable under leniency programs, the inefficient firm can gain a greater market share by threatening to self-report, which reduces both productive and allocative efficiencies. However, the efficient firm will foresee the hold-up problem and hence become reluctant to collude. The introduction of leniency programs therefore present a trade-off between ex-ante deterrence and ex-post efficiency. By contrast, traditional antitrust investigations are shown to both deter cartels and improve allocation. Leniency programs should be viewed as a second-best solution for budget-constrained antitrust authorities.

Keywords: Corporate leniency; Cost asymmetry; Nash bargaining

JEL Classification: K21; K42; L41

*The author thanks Cedric Argenton, Jan Boone, Martijn Han and Igor Mouraviev for their insightful discussion. Special thanks go to Eric van Damme whose advices and suggestions greatly helped improve the paper. The helpful comments of seminar participants at Tilburg University, University of York, Biannual Conference of Hong Kong Economic Association, CLEEN and ENTER Conferences are gratefully acknowledged.

†The University of Warwick, E-mail: *kebin.ma@wbs.ac.uk*

‡Boston University, Email: *lwhite81@bu.edu*.

1 Introduction

Cartels are illegal per se, but to detect and convict them is intrinsically difficult. To induce information from colluding firms, leniency programs grant fine reductions to those firms that participate in cartels but come forward to antitrust authorities to self-report their offenses.[1] By creating a prisoners' dilemma and fostering distrust among cartel members, leniency programs aim to prevent and destabilize cartels.

Since its major revision in the U.S. in 1993, the leniency program has become a major tool in fighting cartels. Having brought about a "more than ten-fold increase in amnesty applications" and record-breaking antitrust fines, leniency program is highly praised by antitrust officials. They claim that corporate leniency program is "unquestionably, the single greatest investigative tool available to anti-cartel enforcers", "has cracked more cartels than all other tools at our disposal combined", and "led to the detection and dismantling of the largest global cartels ever prosecuted", Hammond (2001, 2005).[2] This success has also inspired the implementation of similar programs in Canada, the EU, the UK and Korea, OECD (2002).

In light of the policy's achievement, a burgeoning game theoretical literature studies firms' incentives under leniency programs, and how the programs should be accordingly designed.[3] Since the pioneering work of Motta and Polo (2003), a sequence of papers, such as Spagnolo (2004), Aubert, Rey, and Kovacic (2006) and Harrington (2008), study the impact of leniency programs. While it is generally acknowledged that leniency programs help fight cartels, various disadvantages are also pointed out. For example, leniency programs are suspected to have encouraged cartel formations by granting reduced fines (Motta and Polo (2003)), or to have increased firms' ability to punish deviation and therefore stabilize cartels (Ellis and Wilson (2001), Harrington and Chen (2005) and Motchenkova and Leliefeld (2010)).

Most of the literature assumes that colluding firms are symmetric. Asymmetry, however, is common in real world scenarios and is well documented for cartels. For example, the recently convicted beer cartel in the Netherlands featured small firms conspiring with large ones. Ganslandt, Persson, and Vasconcelos (2008) record that "in approximately 38% of the 43 European cartel cases between 2002 and 2007, the size of the largest firm exceeds the size of

[1]When cartels are criminalized, criminal charge are also waived for executives. To isolate the effect of cost asymmetry, the current paper focuses on corporate leniency programs, setting aside the individual criminal charge for price-fixing and the associated agency problems.

[2]Prominent cases include the international cartels of Vitamins, DRAM, and Graphite Electrodes.

[3]See Spagnolo (2006) for an excellent survey.

second largest firm by at least 50%." While the asymmetry may result from different causes, cost differences can be a prominent driver.[4]

Assuming cost symmetry simplifies economic modeling, but relaxing this assumption can lend important new insights to both economic analyses and policy practices. First of all, it has been well recognized that cost differences affect cartel formation and firms' collusive agreements. For example, Harrington (2006) records that in the international cartel of Lysine "...it was important for ADM to reveal the extent of its capacity and its low-cost of production; both towards assuring that the cartel form and, in that event, that ADM would be allocated a sufficiently large sales quota".[5] Second, from a policy point of view, productive efficiency should be taken into account in the design of anti-cartel policies. While the paramount goal of anti-cartel policies is to deter cartel formation, it is desirable to minimize the distortion in production and allocation for any cartel that is formed and remains undetected.[6] Ignoring productive efficiency can even defeat the objective of promoting consumer surplus. To the extent that production costs affect prices, consumer surplus cannot be analyzed independently of productive efficiency: Any productive inefficiency would simply lead to high production costs and prices, and therefore reduce consumer welfare. An effective anti-cartel policy, whether it aims to maximize total welfare or consumer surplus, should pay proper attention to productive efficiency. Cost saving is desirable not only for its own sake but also because it contributes to consumer surplus through lower prices.

The recognition of cost asymmetry raises several questions. How do cost differences affect firms' negotiation and their collusive agreements? Why do cartels tend to involve large and efficient firms? How does cost asymmetry affect the effectiveness of anti-cartel policies, and in particular, leniency programs? The current paper aims to answer these questions.

In a duopoly setup with asymmetric costs, I emphasize firms' relative bargaining positions due to their cost differences. As a low-cost firm earns a higher competitive profit, this better outside option gives the firm a strong bargaining position, which enables it to claim a big market

[4]Alternative sources for asymmetry include different capacities, as in Compte, Jenny, and Rey (2002) and Bos and Harrington (2010) or different product varieties, as in Kuhn and Motta (1999). Concerning cost asymmetry, it may rise because of differences in productive and organizational efficiency, but also naturally rises in the context of international cartels: exchange rate variation and tariff can directly create cost differences and affect collusive outcomes. See Fung (1992) and Veugelers and Vandenbussche (1999).

[5]For more theoretical discussions, see Schmalensee (1987), Harrington (1991), Rothschild (1999), and Vasconcelos (2005).

[6]When the two objectives contradict with each other, a policy trade-off rises between cost saving and preventing monopolisitic distortion. Antitrust authorities may face similar trade-off in other contexts such as horizontal mergers, where efficiency is balanced against possible increase in the emerging firm's market power.

share and a large fraction of the total collusive profit. In particular, I show that the net gain from collusion is higher for the low-cost firm than for its high-cost partner.[7] This induces the efficient firm to collude. At the same time, the small market share allocated to the inefficient firm limits the benefit for the efficient one to undercut. Under trigger strategies, a firm does not deviate if the forgone profit from future collusion exceeds the immediate gain from deviation. Since the low-cost firm faces both limited benefit from undercutting and large gains from collusion, it will have a strong incentive to collude, a result that may help explain why cartels often involve large and efficient firms.[8]

Having analyzed firms' negotiation under cost asymmetry, I proceed to examine how leniency programs affect firms' collusive agreements and their incentives to collude by changing firms' outside options. If a cartel survives leniency programs, the high-cost firm can threaten its low-cost partners by self-reporting.[9] In order to stabilize the cartel, the efficient firm will have to bribe its inefficient conspirator with a higher collusive profit and a larger market share. This would lead to collusive schemes where a larger quantity is produced by the high-cost firm. As a result, productive efficiency deteriorates. When market price increases with production cost, consumers will suffer so that allocative efficiency deteriorates as well. Such perverse effects, however, may not emerge in equilibrium, because the efficient firm will anticipate the hold-up problem and become more reluctant to form a cartel in the first place. In other words, leniency programs interfere the negotiation among firms and create ex-ante deterrence for cartels. The use of leniency programs therefore introduces a basic trade-off between ex-ante deterrence and ex-post efficiency. On the one hand, it creates a hold-up problem and makes it harder to form a cartel. On the other hand, if a cartel remains sustainable under leniency programs, it will be less efficient both in production and allocation.

To the best of my knowledge, only two papers study leniency programs for asymmetric firms. Motchenkova and Van der Laan (2005) consider firms that collude in one market but differ in their presence in other non-cartelized markets. As antitrust conviction leads to a loss of

[7] The net gain is defined as a firms' collusive profit minus its competitive profit.

[8] Another explanation for the observed pattern can be that a cartel that exclude large firms can easily fail because the competitive fringe can effectively challenge the collusive price. It is also noteworthy that existing results are diverse on how incentives to collude relate to firm characteristics. While some conclude large firms are more inclined to cartelize the market, e.g., Vasconcelos (2005), Kuhn and Motta (1999) and Bos and Harrington (2010), others argue large firms are harder to be disciplined and tend to deviate, e.g., Compte, Jenny, and Rey (2002). The diverse results partly reflects diverse assumptions adopted in modeling. For example, Rothschild (1999) assumes colluding firms minimize the total production cost. Vasconcelos (2005) assumes production quotas are proportional to firms' capacities.

[9] Even though the efficient firm can also threaten to report, the threat is less severe when the antitrust fines are positively related to firms' illegal profits.

reputation and reduces sales in all markets, a better diversified firm faces a greater reputational cost and is less willing to self-report. Motchenkova and Leliefeld (2010) study firms that differ in accumulated profits, showing that firms with high accumulated profits can prevent their small conspirators from self-reporting by predation. Unlike the current study, both of the papers assume that firms produce at the same cost, and do not allow the asymmetry to affect firms' collusive agreements.

The ex-ante deterrence predicted by the current paper seems consistent with empirical and experimental findings that leniency programs destabilize cartels.[10] Using U.S. data, Miller (2009) shows that the number of convicted cartels increased instantaneously at the introduction of the revised leniency program, and remained low afterwards, indicating that leniency increases detection power and deters cartels persistently. Taking an experimental approach, Hinloopen and Soetevent (2008) and Bigoni, Fridolfsson, Le Coq, and Spagnolo (2008) show that fewer cartels are formed at the presence of leniency programs. Harrington and Chang (2009) derive a structural model, in which the life length of detected cartels is an informative indicator of the number of active cartels. And their calibration shows that leniency programs reduce the duration of collusion and therefore deter cartel formations.

Related to the current paper, Aubert, Rey, and Kovacic (2006) highlight another potential inefficiency due to leniency programs. Assuming bounties to individual informants, the authors argue that if current managers are involved in cartel negotiation and need to be dismissed for efficient corporate restructuring, they will have to be bribed not to take antitrust bounties or reveal the cartel to competition authorities. Firms therefore will forgo efficient restructuring should the bribery be sufficiently high.[11]

I also provide a comparison between leniency programs and traditional antitrust enforcement. When antitrust investigations convict cartels with certain probabilities and charge fines proportional to firms' illegal profits, traditional enforcement can both deter cartels and increase productive and allocative efficiencies. This poses a clear contrast to leniency programs, which achieve deterrence only at the potential cost of inefficiencies. Leniency programs therefore

[10]The only exception is Brenner (2009). The author shows leniency programs have no significant effects in Europe. But the the result should be best interpreted with caution, because the data set covers year 1990-2003, largely before the revision of EU leniency program in 2002. For the period covered, the discretion in granting leniency may have severely undermined the effectiveness of EU leniency programs.

[11]A natural extension not in the paper is to consider shareholders' ex-ante incentives to cartelize markets. Arguably, shareholders will anticipate the cost of bribing employees and the forgone opportunity of efficient restructuring, which makes collusion less attractive and deters cartel formation in the first place. Therefore, the principal-agent model can potentially generates a similar trade-off between efficiency and deterrence, a result closely related to the current paper but generated with different reasonings.

provide no free lunch and should only be seen as a second-best solution for budget-constrained competition authorities. This result is in line with Motta and Polo (2003), who show while installing a leniency program saves investigation costs, it may also encourages collusion because fine reductions make it cheaper to participate in cartels.

It should also be recognized that leniency programs cannot work independently of traditional investigations. As well acknowledged in anti-cartel practices (e.g. Hammond (2004)), intensive investigations and severe sanctions are essential to induce self-reporting.[12] Spagnolo (2004) shows that for leniency programs to make collusion a risky equilibrium, a critical level of traditional enforcement must be guaranteed. Chang and Harrington (2008) warn of the danger that if investigations slack because leniency applications bring up additional caseload, leniency programs can even backfire. The current paper resonates with the extant ones and brings the argument one step further. I show that traditional antitrust enforcement reduces inefficiencies that leniency programs can cause. Even if leniency programs result in a trade-off between deterrence and efficiency, the side-effect can be mitigated by tranditional investigations.[13] In short, leniency programs are no substitute to traditional investigations, but rather rely on them.

The paper proceeds as follows. Section 2 sets up the model. Section 3 characterizes firms' collusive contracts in the absence of any antitrust enforcement. In section 4, traditional antitrust investigations are analyzed for its impacts on allocation and deterrence. Section 5 turns to leniency programs. I first examine productive and allocative inefficiencies when a cartel survives leniency programs, and further analyze firms' ex-ante incentives to cartelize the market. Section 6 studies the case where traditional enforcement and leniency programs are both at play, showing that traditional investigations can reduce the ex-post inefficiencies caused leniency programs. Section 7 concludes.

2 The model

The model setup unfolds in two steps. It starts with one-shot games where firms negotiate to divide a market. In these static settings, Nash bargaining outcomes are formulated under

[12] The point is also clear in academic papers such as Chang and Harrington (2008): "A cartel member will apply for leniency only if it believes that doing so is better than running the risk of being caught and paying full penalties. Thus, the probability of being caught and convicted is integral to inducing firms to apply for leniency"

[13] In fact, the current paper suggests that the welfare consequnces of leniency programs depend crucially on where the non-leniency enforcement comes from. When it is due to a higher fine scheme, inefficiency aggravates when a cartel sustains. On contrast, if the enforcement is due to a higher probability of detection, both deterrence and efficiency can improve.

different policy regimes. Afterwards, the model is extended to incorporate dynamic elements such as repeated games, which implement the static bargaining outcomes.

2.1 The static game

I study a two-firm industry that faces a linear demand. The inverse demand function is normalized to $P = 1 - Q$, where P is the market price and Q is the industry output. Two firms are assumed to produce at constant marginal costs, with cost function $C_i(q) = (1-\theta_i)q$, $i = 1, 2$. The two firms, however, differ in their productive efficiency: With $\theta_2 < \theta_1 < 2\theta_2$, firm 1 is assumed to be more efficient and produces at a lower cost. The inequality $\theta_1 < 2\theta_2$ precludes the uninteresting case where the monopolistic price of firm 1 is lower than firm 2's marginal cost and the inefficient firm will be entirely driven out of the market. Denote firm 1's relative cost advantage by $x \equiv \theta_1/\theta_2$. The assumption can be written compactly as $x \in (1, 2)$.

When the two firms collude and form a cartel, they are assumed to "divide markets" as defined in Schmalensee (1987). That is, they agree not only on market shares but also on the allocation of potential consumers.[14] In other words, firms agree not to serve consumers allocated to their conspirators; each of them is able to charge a uniform monopolistic price in its own consumer pool.[15] Since firms prefer different collusive prices under cost asymmetry, the possibility of dividing markets simplifies the analysis by letting a collusive agreement specify only market shares and delegate pricing decisions to individual firms.[16] In particular, I assume that colluding firms divide the market according to a fraction $g \in [0, 1]$. Firm 1 takes a market share of g, and firm 2 takes $1 - g$.

[14]The assumption is more than a pure theoretical simplification. In the survey of Levenstein and Suslow (2006), it is observed "market allocation" contributes to stabilizing cartels in about 30% of the studied cases. The concept "market allocation" is broad, including the use of production quotas, division of markets, division of territories, and allocation of customers. In essence, each of the terms divides the market and delegates the pricing decision to individual firms. Market division also naturally holds in some other situations like repeated auctions. For example, firms can collude on which firm is to submit the winning bid for an order, with the winning firm rotating every few weeks. This arrangement works well for products whose orders are relatively small and frequent.

[15]Since firms prefer different prices under cost asymmetry, a no-arbitrage condition is assumed throughout the paper. The assumption may be naturally satisified for intermediate goods or in situations such as auctions.

[16]A similar idea can be found in Ivaldi, Jullien, Rey, Seabright, and Tirole (2003). In analyzing cost asymmetry, the authors assume the demand is inelastic: firms can sell an arbitrary amount D of the product as long as the price does not exceed consumers' reservation level. So firms will readily agree on a collusive price equal to consumers' willingness to pay, whatever their costs are. But if consumers buy from the two firms with equal probabilities, one has to assume side payment if the collusive profit is to divided unevenly between cartel members. The presence of side payment also makes it difficult to draw any conclusion as to production efficiency, as in that case, the inefficient firm will be optimally shut down.

Being a monopoly in its own market, firm 1 solves the following optimization programme.

$$\max_{q_1} \left\{ \left[(1 - \frac{q_1}{g}) - (1 - \theta_1) \right] q_1 \right\} \tag{1}$$

It produces $q_1^C = g\theta_1/2$, and makes a profit

$$\Pi_1^C = g\theta_1^2/4, \tag{2}$$

where the superscript C denotes collusion. Similarly, we can solve for the output and profit of firm 2. That is, $q_2^C = (1-g)\theta_2/2$, and

$$\Pi_2^C = (1-g)\theta_2^2/4. \tag{3}$$

Firms Nash bargain to determine the market share g, which in turn determines the allocation of their collusive profits. Side payments are not allowed.[17] Thus a firm's profit must follow its allocated market share and the corresponding output. Conversely, from firms' profit allocation, one can solve for their market division. Formally, the profit allocation and the market share g is determined by the following Nash bargaining solution.

$$N(W,d) = \underset{(d_1,d_2)<(w_1,w_2)\in W}{argmax} (w_1 - d_1)(w_2 - d_2) = (\Pi_1^C(g), \Pi_2^C(g)) \tag{4}$$

I assume firms are risk neutral profit maximizers, so that W becomes the set of profit allocations achievable under market-dividing collusion. When collusion cannot be achieved, the two firms play a non-cooperative game, Γ, in which $d = (d_1, d_2)$ is the payoff associated with its Nash equilibrium.[18] In other words, the Nash equilibrium of game Γ is taken to be the disagreement event, and the disagreement point corresponds to the associated equilibrium payoffs.

The paper studies collusion in four scenarios: 1) a free market without any antitrust intervention, 2) a market where only traditional antitrust enforcement is present without leniency programs, 3) a market where only leniency programs are present without traditional investigations, and 4) a general case where both leniency programs and traditional investigation are at work. Notation-wise, I denote the four scenarios by superscripts F, T, L, and TL respectively.

[17] One justification for this assumption can be that monetary transfers leave trails of evidence that expose cartels to prosecution. For classic criticisms of side payments, see Bain (1948).

[18] I show in this context, only one equilibrium is strict. So the disagreement point d is unique.

Now I discuss in turn how the policy interventions are modeled and how the elements of firms' bargaining game (W, d, and Γ) vary with different policy regimes.

2.1.1 Collusion in a laissez-faire market

In the absence of antitrust interventions, firms are assumed to Bertrand compete with each other when they cannot agree to collude.[19] A firm's strategy is characterized by its posted price $P_i \in [1-\theta_i, 1)$. As firms undercut each other down to their marginal costs, in Nash equilibrium, firm 1 will charge a price of $P_1^B = (1-\theta_2) - \epsilon$ and serve all consumers. This leads to

$$d_1^F = \Pi_1^B = \theta_2(\theta_1 - \theta_2) \quad \text{and} \quad d_2^F = \Pi_2^B = 0.$$

The assumption of market division makes the set of feasible profit allocations convex and compact. To see the point, note from expression (2) and (3) that both firms' profits are linear in g. Therefore, the Pareto frontier of feasible allocation set will also be linear in profit space $\Pi_1 \times \Pi_2$. Specifically, it takes the form

$$\theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \theta_1^2\theta_2^2/4, \tag{5}$$

and is depicted by the solid line in Figure 1. Each point on the frontier corresponds to a unique market division g that firms can agree on. The frontier, together with non-negative conditions, $\Pi_1 \geq 0$ and $\Pi_2 \geq 0$, characterizes the set of feasible allocations, W^F. The result is summarized in the following lemma.

Lemma 1. *Under market division, the Pareto frontier of feasible profit allocation set is linear, and represented by* $\theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \theta_1^2\theta_2^2/4$.

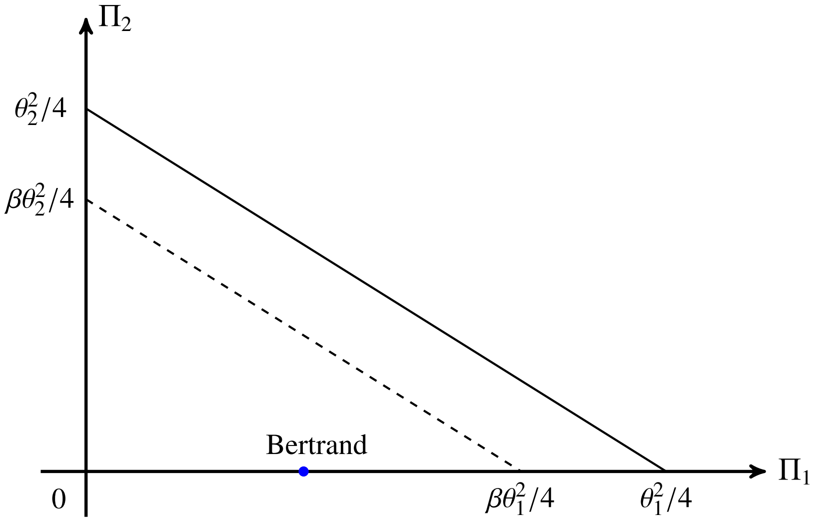
2.1.2 Collusion under traditional enforcement

Now we turn to the scenario where leniency programs are assumed to be absent and antitrust authorities solely rely on investigations to detect cartels, a case I call "traditional enforcement". I assume that antitrust investigations detect and convict cartels with probability s.[20] Once a

[19] To the extent that the efficient firms still earns a higher competitive profit and enjoys a favorable bargaining position, assuming Cournot competition will not qualitatively change the results, but will considerably complicate the computation.

[20] While colluding firms may not be detected, I assume that antitrust authorities make no type II error by convicting innnocent firms.

Figure 1: Pareto frontiers of feasible profit allocation set



cartel is convicted, colluding firms have to pay fines proportional to their collusive profits.[21]

$$F_i^T = \alpha\Pi_i^C \quad \text{with} \quad \alpha > 0 \tag{6}$$

The expected profit for colluding firm i becomes

$$E(\Pi_i^C) = s(1-\alpha)\Pi_i^C + (1-s)\Pi_i^C = (1-\alpha s)\Pi_i^C.$$

Denote $\beta \equiv 1 - s\alpha$. Traditional antitrust investigation can be seen discounting firms' collusive profits by $\beta \in (0, 1)$. Specifically, denoting firm i's collusive profit under traditional antitrust enforcement by $\Pi_{i,T}^C$, we have

$$\Pi_{1,T}^C = \beta\Pi_1^C = \beta g\theta_1^2/4 \quad \text{and} \quad \Pi_{2,T}^C = \beta\Pi_2^C = \beta(1-g)\theta_2^2/4.$$

Since both $\Pi_{1,T}^C$ and $\Pi_{2,T}^C$ are linear in g, following the same procedure as in section 2.1.1, we can derive a new Pareto frontier.

$$\theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \beta\theta_1^2\theta_2^2/4$$

[21]This proportional fine scheme can be rationalized by Motchenkova and Kort (2006), who show fines proportional to the gains from collusion deters cartels more effectively than a fixed fine scheme would do.

The antitrust investigation lowers the intercept of the frontier, but does not change its slope. The new frontier is depicted by the dashed line in in Figure 1, which, together with the non-negative conditions $\Pi_1 \geq 0$ and $\Pi_2 \geq 0$, defines a new feasible profit allocation set W^T.

Parameter s reflects investigation intensity and the chance of successful conviction. A greater s means more effective traditional antitrust enforcement. For a given fine scheme α, which is often predetermined by a legal system, more intensive investigations contribute to stronger antitrust enforcement, captured by a lower β.

For leniency programs to be relevant, I assume that antitrust authorities cannot perfectly deter cartels solely by investigation. Therefore, forming a cartel should remain mutually profitable for the two firms. This implies that the Pareto frontier of the feasible allocation set stays to the right of Bertrand payoffs (also illustrated in Figure 1). Analytically, it requires $\beta\theta_1^2/4 > \theta_2(\theta_1 - \theta_2)$, or

$$s < \bar{s} \equiv \frac{1}{\alpha}\left[1 - 4\left(\frac{1}{x} - \frac{1}{x^2}\right)\right] \equiv \frac{1-2y}{\alpha}. \tag{7}$$

In the traditional enforcement, antitrust authorities deter and terminate cartels by increasing investigation intensity s. Inequality (7) is meant to reflect the limited resources and budget constraints that antitrust authorities have: Investigation intensity s cannot be set high enough to perfectly deter cartels, which is exactly why leniency programs need to be introduced. The notation $y \equiv 2/x - 2/x^2$ is introduced for the ease of future presentation, as the expression will be repeatedly used in the rest of the paper.

The presence of the antitrust authority does not expand the set of actions available to firms, whose possible action remains either to collude or to undercut. The same Bertrand game Γ will be played when firms fail to collude. So the disagreement point stays unaltered, $(d_1^T, d_2^T) = (d_1^F, d_2^F) = (\Pi_1^B, \Pi_2^B)$.

2.1.3 Collusion under corporate leniency

I assume that a leniency program waives antitrust fines completely for the first firm that self-reports.[22] Furthermore, colluding firms are assumed to be always able to provide hard information to convict cartels.[23] Therefore, being the first to report automatically guarantees

[22] Spagnolo (2004) and Harrington (2008) show the optimal leniency program entails amnesty only to the first reporter. In fact, the authors advocate bounties for self-reporting. The current paper focuses on existing leniency programs, and therefore excludes bounties.

[23] Aubert, Rey, and Kovacic (2006) provides an economic rationale why colluding evidence is kept.

Figure 2: Payoff matrix under the leniency program

	Firm 2	
Firm 1	Report	No-Report
Report	$\Pi_1^B - F_1^L/2, 0 - F_2^L/2$	$\Pi_1^B, -F_2^L$
No-Report	$\Pi_1^B - F_1^L, 0$	$\Pi_1^B, 0$

amnesty. I also assume a "winner-takes-all" approach in granting leniency: A firm late in self-reporting is not entitled to any fine reduction.[24]

Firms can report to the antitrust authority if they do not reach a collusive agreement. In the absence of traditional investigations, the set of feasible allocations remains the same as in the laissez-faire market ($W^L = W^F$), but the leniency program changes the non-cooperative game Γ to be played when firms disagree. Instead of being passive in investigations, firms can now self-report to apply for leniency. Accordingly, their action sets are expanded by the possibility of reporting, and can be characterized by a price and reporting decision pair (P_i, L_i). $L_i \in \{\text{Report}, \text{No-Report}\}$ adds to the action set which includes only pricing in the absence of leniency. As the reasoning of Bertrand competition applies, any outcome where $P_1 \neq 1 - \theta_1$ or $P_2 \neq 1 - \theta_2$ cannot qualify as an equilibrium. Therefore one can focus on the reduced game where firms charge Bertrand competitive prices and decide whether to self-report.[25]

Furthermore, I assume that firms are equally efficient at self-reporting. When they cannot reach an agreement and decide to report at the same time, each of them wins the race to the antitrust authority with a probability 1/2. With antitrust fines denoted by F_1^L and F_2^L for firm 1 and 2 respectively, table 1 formulates the bi-matrix game to be played by the two firms if they cannot reach a collusive agreement. Also note that F_1^L and F_2^L in principle can be different and depend on firm characteristics and the history of collusion. The exact specification will be fully specified in the dynamic setting that follows.

[24] In this respect, the setup mimics the corporate leniency program in the U.S.

[25] Alternatively one can think of a sequential game and look for subgame perfect equilibria: In the first stage firms choose whether to report or not. Afterwards, given what has happened in stage 1, firms set prices. Still only the equilibrium involves static Bertrand prices is subgame perfect.

With F_1^L and F_2^L both positive, the game has two pure strategy equilibria, where both firms report, or neither of the firms reports. While the first equilibrium is strict, the second is only weak. Arguably, an equilibrium that will be ultimately played should be stable: An infinitesimal probability that one firm plays the other equilibrium should not change the best response of the other firm. This is a requirement that the weak equilibrium does not satisfy. Therefore, the equilibrium where both firms report is taken as the disagreement event, and the associated expected payoffs constitute the disagreement point.

$$d_1^F = E(\Pi_1^B - F_1^L) = \Pi_1^B - F_1^L/2 \tag{8}$$

$$d_2^F = E(\Pi_2^B - F_2^L) = 0 - F_2^L/2 \tag{9}$$

Because firms win the race to the antitrust authority with equal probabilities, when both of them report, $F_i^L/2$ captures, from an ex ante perspective, firm i's expected fine if collusion fails.

2.1.4 Traditional investigation and leniency combined

The scenario where the traditional investigation and the leniency are both at work just combines the results from the last two subsections. While traditional investigations shift Pareto frontier to $\theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \beta\theta_1^2\theta_2^2/4$, the leniency program changes firms' disagreement point to (d_1^F, d_2^F). Therefore, we have $W^{TL} = W^T$ and $d^{TL} = d^L$.

2.2 The dynamic game

I now discuss the timing of the dynamic game, its stationary structure, and how it implements the static bargaining outcomes. The analysis again entails the discussion of different policy scenarios.

2.2.1 Timing and stationarity

For the benchmark case where there is no antitrust intervention, it is assumed that at the beginning of each period, firms decide on whether to form a cartel or to Bertrand compete. If a cartel is formed, firms Nash bargain over the market division g. Once an agreement is reached, production is carried out and collusive profits are earned. Under traditional anti-cartel enforcement, the addition is that at the end of each period, firms face antitrust investigation, which convicts their cartel with a probability s. The process of collusion, bargaining, and investigation (in case of traditional investigation), repeats infinitely. It should be clear that for both cases the repetition has a stationary feature: Firms' outside options and their bargaining outcomes do not vary with time.

Under the leniency program, firms choose at the beginning of a period to stay in the cartel or to deviate. If they deviate and there has been a cartel in the last period, both firm will undercut and self-report, as indicated by the game in Figure 2. It is assumed that evidence on collusion lasts for one period, so that upon the self-reporting in period t, the antitrust authority can only convict the cartel in period $t-1$. When a cartel is convicted because of self-reporting, the first self-reporter will receive full leniency. The other firm will be charged with a fine proportional to its last period profit, since the bargaining and reporting are assumed to happen at the beginning of the period and firms have not earned collusive profit in the current period. Therefore, the antitrust fine takes the following form.

$$F^L_{i,t} = \alpha \Pi^C_{i,t-1}. \tag{10}$$

The case becomes slightly different.

Since the fine enters firms' disagreement payoffs and affects bargaining outcomes, having the antitrust fine being a function of the last period profit implies an iterative relationship between collusive profits in period t and $t-1$. To simplify the analysis, I focus on the case where firms have reached stationarity: after a large number of colluding periods, firms' collusive profits no longer change over time. The contemporary profits result in possible fines that affect the disagreement points in the next period bargaining. Yet based on the fine, the bargaining restores the exact level of profits, making firms' outside options and bargaining outcomes both time-invariant. Formally, as the disagreement point d^L_t evolves according to $d_{i,t} = H(d_{i,t-1})$, as indicated by equation (8) and (9), the following equality should hold.

$$N(W^L, d^L_t) = N(W^L, H(d^L_{t-1})) = N(W^L, d^L_{t-1}) \tag{11}$$

I show in Appendix A that for an arbitrary initial collusive agreement, the bargaining outcome will converge to such a steady state. By focusing on the stationary case, I drop the subscript t in the rest of the paper.

2.2.2 Cartel sustainability

The illegal nature of price-fixing makes it necessary for any collusive scheme to be self-enforcing. I focus on the implementation by infinitely repeated games and trigger strategies: It is assumed that once deviation is observed, whether it is undercutting or self-reporting, Bertrand competition is played forever. On contrast, when a cartel is detected by traditional antitrust investigation, the colluding firms will continue to cartelize the market in periods afterwards.

For a cartel to be successful, it must provide firms sufficient economic incentives to collude. Specifically, the discounted profit must be higher from collusion than from deviation,

$$\Pi_i^C + \Pi_i^D + \delta_i \Pi_i^B + \delta_i^2 \Pi_i^B + \ldots < \Pi_i^C + \delta_i \Pi_i^C + \delta_i^2 \Pi_i^C + \ldots,$$

where Π_i^D is the *extra* profit from the most profitable deviation. The aforementioned stationarity guarantees the time ivariance of Π_i^C and Π_i^D. The incentive compatibility constraint implies a critical discount factor $\hat{\delta}_i$ for firm i to be willing to participate in the cartel. A cartel is sustainable if and only if $\delta_i > \hat{\delta}_i$ for all firms.

$$\hat{\delta}_i \equiv \frac{\Pi_i^D}{\Pi_i^D + (\Pi_i^C - \Pi_i^B)} \tag{12}$$

Following the literature on collusion in infinitely repeated games, the critical discount factor $\hat{\delta}_i$ measures how firm i is willing to collude: a lower critical discount factor indicates a greater appetite for collusion.

In the rest of the paper, I analyze in turn the different policy regimes: the laissez-faire market, the traditional anti-cartel investigation, and the corporate lenciency program. For each of the regimes, I discuss first the collusive allocation in a static setting, and then study cartel sustainability as well as firms' incentives to collude in a dynamic context.

3 Collusion in a free market

3.1 Collusive outcome

I start with the benchmark case where no antitrust intervention is present. Recall that in case of disagreement, the two firms earn $\Pi_1^B = \theta_2(\theta_1 - \theta_2)$ and $\Pi_2^B = 0$ respectively from Bertrand competition. Together with the Pareto frontier $\theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \theta_1^2\theta_2^2/4$, one can formulate a Nash bargaining program as follows.

$$
\begin{aligned}
&\max_{\{\Pi_1,\Pi_2\}} \{(\Pi_1 - \Pi_1^B)(\Pi_2 - \Pi_2^B)\} \\
&s.t. \qquad \theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \theta_1^2\theta_2^2/4
\end{aligned} \tag{13}
$$

The solution suggests when the two firms collude, they will obtain the following profit allocation.

$$
\Pi_{1,F}^C = \Pi_1^C = \frac{\theta_1^2}{8} + \frac{\theta_2}{2}(\theta_1 - \theta_2) \tag{14}
$$

$$
\Pi_{2,F}^C = \Pi_2^C = \frac{\theta_2^2}{8} - \frac{\theta_2}{2}\frac{\theta_2^2}{\theta_1^2}(\theta_1 - \theta_2) \tag{15}
$$

Correspondingly, the *extra* profits from cartelizing the market can be written as follows.

$$
\begin{aligned}
\Pi_{1,F}^C - \Pi_1^B &= \frac{\theta_1^2}{8} - \frac{\theta_2}{2}(\theta_1 - \theta_2) \\
\Pi_{2,F}^C - \Pi_2^B &= \frac{\theta_2^2}{8} - \frac{\theta_2}{2}\frac{\theta_2^2}{\theta_1^2}(\theta_1 - \theta_2)
\end{aligned}
$$

With $\theta_1^2/4 > \theta_2(\theta_1 - \theta_2)$, the sum of monopolistic profits is always higher than the sum of Bertrand competitive profits. While both firms gain from cartelization, the efficient firm gains more.

$$
\Pi_{1,F}^C - \Pi_1^B > \frac{\theta_2^2}{\theta_1^2}\left(\Pi_{1,F}^C - \Pi_1^B\right) = \Pi_{2,F}^C - \Pi_2^B \tag{16}
$$

Proposition 1. *Under Nash bargaining, firm 1 and 2 earn collusive profits* $\Pi_{1,F}^C = \frac{\theta_1^2}{8} + \frac{\theta_2}{2}(\theta_1 - \theta_2)$ *and* $\Pi_{2,F}^C = \frac{\theta_2^2}{8} - \frac{\theta_2}{2}\frac{\theta_2^2}{\theta_1^2}(\theta_1 - \theta_2)$ *respectively. While both firms gain from collusion, the net gain is greater for the efficient firm,* $\Pi_{1,F}^C - \Pi_1^B > \Pi_{2,F}^C - \Pi_2^B$.

Under the assumption of market division, pricing decisions are delegated to individual firms, and the agreed market share g fully characterizes firms' collusive agreement. Furthermore, g can be solved analytically. To see the point, note that we have calculated firms' collusive profits in two ways: The profits can either be solved from the Nash bargaining program, i.e., equation (14) and (15), or be solved from firms' monopolistic pricing decisions, i.e., equation (2) and (3). One can solve for g by equalizing corresponding expressions.

$$g_F = \frac{1}{2} + \frac{2\theta_2}{\theta_1^2}\left(\theta_1 - \theta_2\right) \tag{17}$$

The result suggests that the low cost firm holds a collusive market share more than one half. This is because its cost advantage gives it a better outside option and a better bargaining position. In fact, the efficient firm's collusive market share monotonically increases in its cost advantage. To see so, recall that we defined firm 1's relative cost advantage by $x \equiv \theta_1/\theta_2 \in (1, 2)$. With this defintion, expression (17) can be rewritten as

$$g_F = \frac{1}{2} + \frac{2}{x} - \frac{2}{x^2}. \tag{18}$$

Taking first order derivative with respect to x, one can show $\partial g_F/\partial x = -2x^{-2} + 4x^{-3} > 0$, for any $x \in (1, 2)$. Given the monotonicity, it is also straightforward to verify that firm 1's market share lies between $1/2$ and 1. The results are summarized in the following corollary.

Corollary 1. *In collusion, the efficient firm will be assigned a market share* $g_F = 1/2 + 2/x - 2/x^2 \in (1/2, 1)$. *And the firm's market share increases with its cost advantage.*

The market division g has direct implications for firms' profits and social welfare. On the one hand, a higher market share translates into a higher profit, as evident from equation (2) and (3). On the other hand, a higher g implies improved productive and allocative efficiency. To see this, note that for $\theta_1 > \theta_2$, the total output $g\theta_1/2 + (1-g)\theta_2/2$ increases with g. For a given total output, a higher fraction of the products are produced by the low cost firm, leading to an increase in productive efficiency and the aggregate profit. Furthermore, the prices that firms charge do not vary with their market share g. For example, rewrite firm 1's optimization programme (1) in its P_1.

$$\max_{P_1} \{(P_1 - 1 + \theta_1)g(1 - P_1)\}$$

It is clear that g acts only as a multiplier in the programme, which does not affect the profit maximizing price. Therefore, a higher g means more consumers are served in the cartelized market; and for those who get the products, a larger portion of them pay less and buy from the efficient firm. Both lead to an increase in consumer surplus. At the meanwhile, for a given total output, a higher fraction of it is produced by the low-cost firm, leading to an increase in productive effiency and firms' aggregate profit.

Corollary 2. *A firm's profit monotonically increases with its allocated market share. When a market is cartelized, both allocative and productive efficiencies increase when a greater market share is allocated to the low-cost firm. Total welfare will improves with a higher g.*

3.2 Incentive to collude

I now discuss the conditions under which a collusive agreement above can be implemented. The critical discount factor for the implementation, $\hat{\delta}_{i,F} = \Pi^D_{i,F}/[\Pi^D_{i,F} + (\Pi^C_{i,F} - \Pi^B_i)]$, also reflects firms' incentives to form a cartel. Naturally δ_i monotonically increases in $\Pi^D_{i,F}$ and decreases in $\Pi^C_{i,F} - \Pi^B_i$: a firm has less incentive to be in a cartel if it profits more from deviation, and is more willing to collude if it gains more from it.

In the current model, the deviation profit is always higher for the inefficient firm. To see so, note that the most profitable deviation for the inefficient firm is to undercut in the efficient firm's market but to charge the monopolistic price in its own market. Recall that the monopolistic prices in each market are $P^C_1 = 1 - \theta_1/2$ and $P^C_2 = 1 - \theta_2/2$. When Firm 2 marginally undercuts in Firm 1's market, the *extra* profit from deviation is given by

$$\Pi^D_{2,F} = g_F \left(\theta_2 - \frac{\theta_1}{2}\right) \frac{\theta_1}{2}.$$

On the other hand, the most profitable deviation for the efficient firm is to charge its monopolistic price in both markets. The *extra* profit from the deviation can be written as

$$\Pi^D_{1,F} = (1 - g_F) \frac{\theta_1^2}{4}.$$

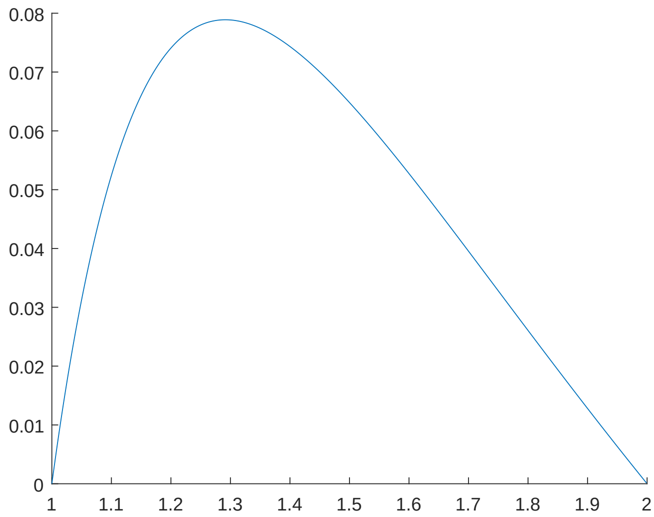
Substituting into the expression for g_F from equation (17), one can show that the extra profit from deviation is lower for Firm 1 than for Firm 2. Specifically, we have

$$\Delta\Pi^D_F(x) \equiv \frac{1}{\theta_2^2}\left(\Pi^D_{2,F} - \Pi^D_{1,F}\right) = -\frac{1}{4} + \frac{1}{4x} + \frac{1}{x^2} - \frac{1}{x^3} \geq 0 \qquad \forall x \in (1, 2).$$

Since it has already been shown the efficient firm gains more from the cartel, this completes the proof that $\hat{\delta}_{1,F} < \hat{\delta}_{2,F}$: the low cost firm has more incentive to form a cartel.

Proposition 2. *For market-dividing cartels, the low cost firm gains less from deviation, giving it more incentive to collude.*

As a matter of fact, the function $\Delta\Pi_F^D(x)$ is an inverted U curve on the interval $x \in (1, 2)$, with $\Delta\Pi_F^D(x = 1) = \Delta\Pi_F^D(x = 2) = 0$, as shown by the figure below.



The intuition is that in the collusive contract, a larger market share is always assigned to the efficient firm, making it more profitable for the inefficient firm to undercut. It is noteworthy that, as the inefficient firm, both its higher production cost and the lower price in the other firm's market reduce the unit profit from cheating. Yet mostly driven by functional forms, the temptation of large market share always dominates in the current model.

4 Traditional antitrust enforcement

4.1 Allocation under investigation and prosecution

Since traditional antitrust enforcement only discounts collusive payoffs by β, the main results from the previous section hold qualitatively. Indeed, collusion in a free market can be considered as a special case where $s = 0$ and $\beta = 1$. Now I generalize the analysis by introducing an active antitrust authority that sets $s > 0$. Recall that the new Pareto frontier will be $\theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \beta\theta_1^2\theta_2^2/4$. Traditional antitrust enforcement lowers its intercept, but does change its slope. Since Nash bargaining solution suggests firms agreeing on the middle point of the frontier, firm 1 will continue to gain more from collusion. Formally, the new Nash bargaining programme can be written as follows.

$$\max_{\{\Pi_1,\Pi_2\}} \{(\Pi_1 - \Pi_1^B)(\Pi_2 - \Pi_2^B)\}$$
$$s.t. \qquad \theta_2^2\Pi_1 + \theta_1^2\Pi_2 = \beta\theta_1^2\theta_2^2/4.$$

It is straightforward to verify that

$$\Pi_{1,T}^C = \beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2}(\theta_1 - \theta_2) \tag{19}$$
$$\Pi_{2,T}^C = \beta\frac{\theta_2^2}{8} - \frac{\theta_2}{2}\frac{\theta_2^2}{\theta_1^2}(\theta_1 - \theta_2), \tag{20}$$

and $\Pi_{1,T}^C - \Pi_1^B > \Pi_{2,T}^C - \Pi_2^B$. Firm 1 continues to gain more from collusion. Correspondingly, the collusive market share becomes

$$g_T = \frac{1}{2} + \frac{1}{\beta}(\frac{2}{x} - \frac{2}{x^2}). \tag{21}$$

Moreover, it remains the case that the low-cost firm has a stronger incentive to collude. Since the price margin remains the same in undercutting, and more market share is allocated to the efficient firm, deviation becomes even more profitable for the inefficient firm, i.e. $\Pi_{2,T}^D > \Pi_{1,T}^D$. Together with $\Pi_{1,T}^C - \Pi_1^B > \Pi_{2,T}^C - \Pi_2^B$, it is easy to verify that $\delta_{1,T} < \delta_{2,T}$.

Ex post efficiency improves with traditional antitrust enforcement. To see this, note that the market share of low-cost firm increases with investigation intensity s. According to corollary 2, this promotes both consumer surplus and production efficiency. The more intensive the investigation, the less the distortion. When investigation intensity approaches the perfect deterrence $\bar{s}$, all market share goes to the efficient firm.

$$\lim_{s\to\bar{s}} g_T = 1$$

The welfare though is still lower than a competitive level: Even if the low-cost firm produces all the output, it charges a monopolistic price so that allocative distortion remains. Therefore, as the antitrust investigation intensifies, productive inefficiency vanishes but allocative inefficiency persists.

Proposition 3. *Facing antitrust investigation, a market-dividing cartel allocates a higher market share to the efficient firm, $\partial g_T/\partial s > 0$. Even if the cartel sustains under investigation, its productive and allocative efficiency improves as investigation intensity increases.*

4.2 Ex ante deterrence

Effective antitrust enforcement also adds to ex ante deterrence. First of all, facing lower collusive profits in expectation, both firms' incentive to collude will be decreased. If no antitrust fine applies in case of deviation, defined as at least one of the firms undercutting, the deviation profit becomes even higher for both firms, because neither of them needs to bear the expected fine from investigation. But the effects are asymmetric. As a larger market share is allocated to firm 1, the inefficient firm gains more from deviation and undercutting. Its incentive to stay in the cartel is therefore unambiguously decreased.

The situation is more subtle for the efficient firm. On the one hand, antitrust enforcement discounts its collusive profit, making it less profitable to form the cartel. On the other hand, its collusive market share increases with the antitrust enforcement, making collusion more attractive. The deterrence effect, nevertheless, dominates in this model.[26]

Proposition 4. *Traditional antitrust enforcement reduces both the efficient and inefficient firms' incentive to collude. That is, $\partial\hat{\delta}_{1,T}/\partial\beta < 0$ and $\partial\hat{\delta}_{2,T}/\partial\beta < 0$.*

Proof. I focus on firm 1, as the case for firm 2 is already evident. Formally, the efficient firm finds it less attractive to join the cartel if and only if

$$\Pi^D_{1,T}/[\Pi^D_{1,T} + (\Pi^C_{1,T} - \Pi^B_1)] \geq \Pi^D_{1,F}/[\Pi^D_{1,F} + (\Pi^C_{1,F} - \Pi^B_1)].$$

To check whether this is true, take the inverse of the fractions to have

$$\frac{\Pi^D_{1,T} + \Pi^C_{1,T} - \Pi^B_1}{\Pi^D_{1,T}} \leq \frac{\Pi^D_{1,F} + \Pi^C_{1,F} - \Pi^B_1}{\Pi^D_{1,F}}$$

or

$$1 + \frac{\Pi^C_{1,T} - \Pi^B_1}{\Pi^D_{1,T}} \leq 1 + \frac{\Pi^C_{1,F} - \Pi^B_1}{\Pi^D_{1,F}}.$$

[26] The proof below assumes implicitly that the deviating firm is still subject to an antitrust fine in its own market, but is not exposed to antitrust sanction in the market where it undercuts. Under the alternative assumption that no antitrust fine applies in neither market when a firm undercuts, the result still holds.

For

$$\Pi^D_{1,T} = (1-g_T)\frac{\theta_2}{2}(\theta_1 - \frac{\theta_2}{2}) \quad \text{and} \quad \Pi^D_{1,F} = (1-g_F)\frac{\theta_2}{2}(\theta_1 - \frac{\theta_2}{2}),$$

the expression boils down to

$$f(\beta) \equiv \frac{\beta \cdot g_T \cdot \theta_1^2/4 - \theta_2(\theta_1 - \theta_2)}{1-g_T} \leq \frac{g_F \cdot \theta_1^2/4 - \theta_2(\theta_1-\theta_2)}{1-g_F}.$$

It is straightforward to see that when β approaches 1 the equality holds. Also note that the right hand side of the inequality is not a function of β. Therefore it is sufficient to prove $f(\beta)$ is monotonically increasing. I now show this is indeed the case. Writing g_T explicitly, note that the left hand side further simplies to

$$f(\beta) = \frac{\beta \cdot \theta_1^2/8 - \theta_2(\theta_1 - \theta_2)/2}{1-g_T}.$$

Taking the first order derivative, one has

$$\begin{aligned}
\frac{\partial f(\beta)}{\partial \beta} &= \frac{1}{(1-g_T)^2}\Big[\frac{\theta_1^2}{8}(1-g_T) - (-\frac{\partial g_T}{\partial \beta})\Big(\beta\frac{\theta_1^2}{8} - \frac{1}{2}(\theta_1-\theta_2)\theta_2\Big)\Big] \\
&= \frac{1}{(1-g_T)^2}\Big[\frac{\theta_1^2}{4}\Big(\frac{1}{2}(\frac{1}{2} - \frac{y}{\beta}) + (-\frac{y}{\beta^2})(\frac{\beta}{2} - y)\Big)\Big] \\
&= \frac{1}{(1-g_T)^2}\frac{\theta_1^2}{4}\Big[\frac{1}{4} - \frac{y}{\beta} + \frac{y^2}{\beta^2}\Big] \\
&= \frac{1}{(1-g_T)^2}\frac{\theta_1^2}{4}(\frac{1}{2} - \frac{y}{\beta})^2 > 0.
\end{aligned}$$

This completes the proof: stronger traditional enforcement, as represented by a lower β, always reduces firms' incentives to form a cartel. □

Traditional antitrust investigation has desirable impacts not only on the efficiency of surviving cartels, but also on the deterrence of cartels. However antitrust authorities may face resource and budget constraints so that the investigation intensity cannot be set as high as desired. In light of such limitations, corporate leniency can serve a useful supplement.

5 Corporate leniency

5.1 Allocation under leniency programs

Unlike traditional antitrust enforcement, leniency programs do not change the set of feasible profits available to firms, but instead alter their payoffs in the event of disagreement. When firms cannot reach a collusive agreement, Nash equilibrium suggest that firms self-report and undercut at the same time. While undercutting leads to Bertrand profits, simultaneous reporting leaves firms equal probabilities to be the first self-reporter and to obtain leniency. When their cartel fails, firms will in expectation receive Bertrand profits net of expected antitrust fines. Therefore the disagreement point is $(\Pi^B_{1,t} - F^L_1/2, 0 - F^L_2/2)$, with $F^L_i = \alpha \Pi^C_{i,L}$, and the bargaining program can be formulated as below.

$$\begin{aligned}
&\max_{\{\Pi_1,\Pi_2\}} \{(\Pi_1 - \theta_2(\theta_1 - \theta_2) + \frac{\alpha}{2}\Pi_1)(\Pi_2 + \frac{\alpha}{2}\Pi_2)\} \\
&s.t. \qquad \theta_2^2 \Pi_1 + \theta_1^2 \Pi_2 = \theta_1^2 \theta_2^2 / 4.
\end{aligned}$$

Solving the program, one would obtain the following bargaining outcome.

$$\Pi^C_{1,L} = \frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1 - \theta_2) \tag{22}$$

$$\Pi^C_{2,L} = \frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)} \cdot \frac{\theta_2^2}{\theta_1^2}(\theta_1 - \theta_2) \tag{23}$$

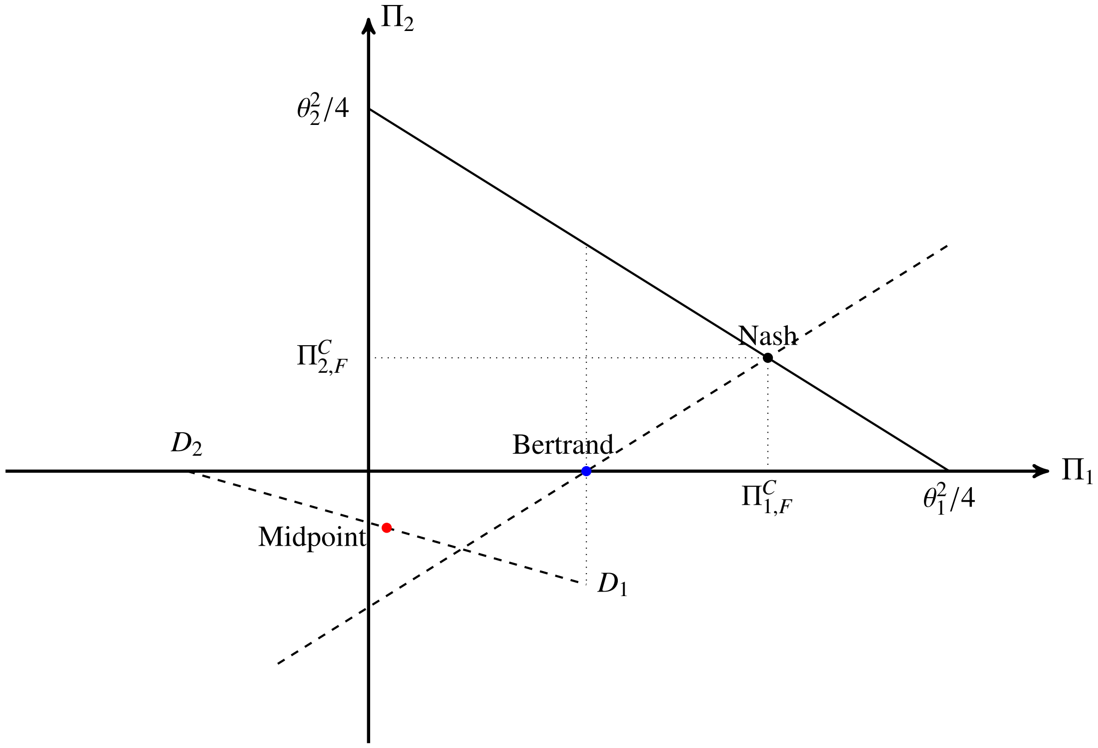
The collusive profits imply a market division

$$g_L = \frac{1}{2} + \frac{1}{1+\alpha/2}\Big(\frac{2}{x} - \frac{2}{x^2}\Big).$$

In the new collusive outcome, firm 1 earns a lower profit and takes a smaller market share as compared to the case where leniency is absent.

The result can be best visualized in Figure 3. Note that all profit pairs to the left of Bertrand-Nash line generate bargaining outcomes in favor of the high-cost firm. Denote by D_i the payoffs where firm i is the first self-reporter. The new disagreement point will be the middle point of segment $[D_1, D_2]$. As the middle point is to the left of Bertrand-Nash line, the efficient firm's bargaining position weakens and its market share shrinks. Intuitively, being able to hurt the other party in the disagreement event increases a firm's profit if a cooperative agreement is

Figure 3: The Impact of Leniency Program on Bargaining Outcome



actually reached. The presence of leniency improves the bargaining position of the inefficient firm by allowing it to threaten its efficient partner.[27]

The leniency program hurts efficiency if a cartel survives. To see so, just recall corollary 2. A lower market share for the low-cost firm implies deterioration in both consumer surplus and firms' aggregate profit; both allocative and productive efficiency will drop. Furthermore, given the monotonic relationship between welfare and market division g, the loss of efficiency can be measured by the following expression.

$$g_F - g_L = \left(1 - \frac{1}{1+\alpha/2}\right)\left(\frac{2}{x} - \frac{2}{x^2}\right)$$

Note that the higher the fine scheme α, the greater the distortion.[28]

Proposition 5. *Compared to the case where leniency is absent, both allocative and productive efficiency will drop if a cartel remains stable under the leniency program. The efficiency loss increases with antitrust fine scheme α.*

[27]Even though the efficient firm can also threaten to report, the threat is less severe when the antitrust fines are positively related to firms' illegal profits.

[28]When α approaches zero, equation (19) restores.

5.2 Ex ante deterrence

The perverse effects of leniency programs, however, only prevail when the cartel remains stable. As the low cost firm foresees the hold-up problem and the lower gains from collusion, it will become more reluctant to join the cartel in the first place. The incentive to collude is further reduced by its higher deviation profit. Since a larger market share is allocated to its inefficient partner, undercutting becomes more attractive for the efficient firm. The use of leniency programs, therefore, introduce a trade-off for antitrust authorities. To reduce the ex ante incentive to collude for the efficient firm, the antitrust authority takes the risk that ex post efficiency aggravates for the surviving cartels.

Proposition 6. *By allocating a higher market share and profit to the inefficient firm, the leniency program reduces the efficient firm's ex ante incentive to join the cartel: $\hat{\delta}_1$ increases under leniency program. And the opposite holds for the inefficient firm: $\hat{\delta}_2$ falls.*

Compared with traditional investigation, which also helps deter cartels, the deterrence of the leniency program comes with a cost. It distorts the market share and production in favor of the inefficient firm. Bearing this limitation in mind, the leniency program has its advantage in cost saving. It saves not only the cost of cartel detection but also facilitates the proof of the existence of a cartel, which in many cases is challenging. The leniency program can supplement investigation when an antitrust authority faces a resource constraint, but attention should be paid to its side effects on ex post efficiency.

6 Leniency and investigation: substitutes or complements?

Up to now, leniency programs and traditional antitrust investigation have been examined separately. A natural question is whether they are substitutes or complements. This section combines the two policy tools to explore their interaction. In particular, I show that intensive antitrust investigation helps to reduce the ex-post productive inefficiency resulted from leniency programs, suggesting a synergy between the two policy tools.

When leniency programs and traditional antitrust enforcement are both at play, we have a general case where both feasible allocation sets and the disagreement point are changed. Nash bargaining programme becomes

$$\max_{\{\Pi_1,\Pi_2\}} \{(\Pi_1 - \theta_2(\theta_1 - \theta_2) + \frac{\alpha}{2}\Pi_1)(\Pi_2 + \frac{\alpha}{2}\Pi_2)\}$$
$$s.t. \qquad \Pi_2 = \beta\frac{\theta_2^2}{4} - \frac{\theta_2^2}{\theta_1^2}\Pi_1.$$

It has bargaining outcome

$$\Pi^C_{1,TL} = \beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1 - \theta_2) \tag{24}$$
$$\Pi^C_{2,TL} = \beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)} \cdot \frac{\theta_2^2}{\theta_1^2}(\theta_1 - \theta_2). \tag{25}$$

The agreed market share takes the form

$$g_{TL} = \frac{1}{2} + \frac{1}{\beta}\frac{1}{1+\alpha/2}(\frac{2}{x} - \frac{2}{x^2}).$$

Measuring the productive inefficiency due to leniency programs by

$$g_F - g_{TL} = \left(1 - \frac{1}{\beta}\frac{1}{1+\alpha/2}\right) \cdot y$$

I show the distortion diminishes as traditional antitrust enforcement strengthens. First, it is straightforward to verify that

$$\frac{\partial(g_F - g_{TL})}{\partial s} = \frac{1}{\beta^2}\frac{\partial\beta}{\partial s}\frac{1}{1+\alpha/2} \cdot y < 0.$$

As the investigation intensity s rises, the size of production distortion decreases. Second, even though a higher fine scheme α unambiguously increases efficiency loss in the absence of the traditional enforcement, it can help to reduce the efficiency loss once the traditional enforcement comes into play. To see this, it is sufficient to examine the comparative statics with respect to α.

$$\begin{aligned} \frac{\partial(g_F - g_{TL})}{\partial\alpha} &= \frac{-1}{\beta^2(1+\alpha/2)^2}\frac{\partial(1-\alpha s)(1+\alpha/2)}{\partial\alpha} \cdot y \\ &= \frac{-1}{\beta^2(1+\alpha/2)^2}(-s + \frac{1}{2} - \alpha s) \cdot y \end{aligned}$$

The expression is negative if and only if $(1 + \alpha)s < 1/2$. The result is summarized in the corollary below.

Corollary 3. *The productive inefficiency that results from leniency programs can be reduced by stronger traditional enforcement. In particular, the distortion unambiguously decreases with more intensive investigation s, and decreases with fine scheme α if and only if* $(1 + \alpha)s < 1/2$.

In the author's opinion, the condition $(1+\alpha)s < 1/2$ may represent a reasonable case: as the small number of cartels that were sanctioned before leniency programs seems to indicate, the conviction probability s can be rather low in practice, perhaps because of the intrinsic difficulty in convicting cartels,[29] or because of the limited resources (budget, staff, etc) that antitrust authorities command.

7 Concluding remarks

In the paper, I have analyzed how cost asymmetry influences the effectiveness of leniency program. I show that leniency programs can be useful in assisting cartel detection, especially when budget constraints of antitrust authorities' allow for only limited investigation intensity. However, the detective power of leniency programs is not a free lunch. When a cartel survives leniency programs, the inefficient firm will produce more, aggravating both productive and allocative inefficiency. In using leniency programs, antitrust authorities have to trade off between ex-ante deterrence and ex-post efficiency. This contrasts traditional investigation, which can both deter cartels and reduces productive as well as allocative inefficiency. Furthermore, I show that leniency programs and traditional investigation are compliments rather than substitutes: intensive traditional investigation helps to reduce the potential inefficiency caused by leniency programs.

Compared with the efficiency and deterrence trade-off in Aubert, Rey, and Kovacic (2006), the current paper suggests more degrees of freedom for policy makers. Specifically, the antitrust authority can set the fine scheme higher, if it wants to maximize the deterrence effects of leniency. And conversely, if more weight is placed on ex post efficiency, the fine scheme can be scaled down. The result also relates to Motta and Polo (2003) in the sense that leniency is a second-best solution for antitrust authorities. The best way to fight cartels is still traditional

[29] Colluding firms are aware of the severe punishment of price-fixing and try their best to conceal incriminating evidence.

investigation, which deters cartels ex ante as well as improves ex post efficiency in case the cartel is not detected.

Introducing cost asymmetry also helps to explain self-reporting after leniency programs are installed. The existing literature has difficulty in explaining there occurrences. Since firms are rational, they will not form a cartel when anticipating that some of them have the incentive to self-report. Cartels either do not form in the first place or are sustainable under leniency. No self-reporting will be observed in equilibrium. But this phenomenon can be interpreted as a result of the change in firms' relative costs. It might be the case that firms' relative production cost changes after they are already in a cartel, so that they find the old collusive agreement no longer sustainable and race to the court to report.

Appendix A Convergence to steady state

I discuss the most general case where leniency and traditional enforcement are both at work. The other regimes are just special cases with $s = 0$ or $\alpha = 0$.

Given the time varying Nash bargaining program

$$\max_{\{\Pi_{1,t},\Pi_{2,t}\}} \{(\Pi_{1,t} - \theta_2(\theta_1 - \theta_2) + \frac{\alpha}{2}\Pi_{1,t-1})(\Pi_{2,t} + \frac{\alpha}{2}\Pi_{2,t-1})\}$$
$$s.t. \qquad \Pi_{2,t} = \beta\frac{\theta_2^2}{4} - \frac{\theta_2^2}{\theta_1^2}\Pi_{1,t},$$

one can solve for an iterative relationship

$$\Pi^C_{1,t} = (1 + \frac{\alpha}{2})\beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2}(\theta_1 - \theta_2) - \frac{\alpha}{2}\Pi^C_{1,t-1}, \tag{A.26}$$

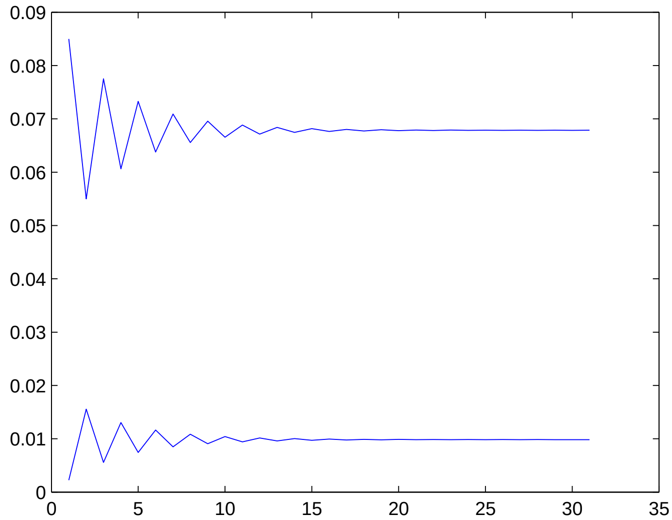
which gives the time path of bargaining outcomes for an arbitrary initial collusive agreement $(\Pi^C_{1,0}, \Pi^C_{2,0})$,

$$\Pi^C_{1,t} = \beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1 - \theta_2) + (\frac{\alpha}{2})^{t-1}[\Pi^C_{1,0} - (\beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1 - \theta_2))]. \tag{A.27}$$

Under a convergence condition $\alpha < 2$, solution (A.27) converges to equation (22) as t goes to infinity. Mathematically, the limit value is a fixed point to a contraction mapping derived from the first order condition of the Nash bargaining program (24).

$$\begin{aligned}\Pi^C_{1,\infty} &= \beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1-\theta_2) + (\frac{\alpha}{2})^{\infty}[\Pi^C_{1,0} - (\beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1-\theta_2))] \\ &= \beta\frac{\theta_1^2}{8} + \frac{\theta_2}{2(1+\alpha/2)}(\theta_1-\theta_2)\end{aligned}$$

For an illustration, the path is depicted in the figure below, with initial value $\Pi^C_{1,0} = 0.085$, parameters$\theta_1 = 0.6$, $\theta_2 = 0.4$, $\alpha = 1.5$ and $\beta = 1$. The upper part is the profits of firm 1, and the lower part is the profit of firm 2.



References

Aubert, C., P. Rey, and W. E. Kovacic (2006): "The impact of leniency and whistle-blowing programs on cartels," *International Journal of Industrial Organization*, 24(6), 1241–1266.

Bain, J. S. (1948): "Output Quotas in Imperfect Cartels," *The Quarterly Journal of Economics*, 62(4), 617–622.

Bigoni, M., S.-O. Fridolfsson, C. Le Coq, and G. Spagnolo (2008): "Fines, Leniency, Rewards and Organized Crime: Evidence from Antitrust Experiments," Discussion Paper N. 698, SSE/EFI Working Paper Series in Economics and Finance.

Bos, I., and J. Harrington (2010): "Endogenous cartel formation with heterogeneous firms," *The RAND Journal of Economics*, 41(1), 92–117.

Brenner, S. (2009): "An empirical study of the European corporate leniency program," *International Journal of Industrial Organization*, 27(6), 639–645.

Chang, M. H., and J. E. Harrington (2008): "The Impact of a Corporate Leniency Program on Antitrust Enforcement and Cartelization," Discussion paper, Economics Working Paper Archive, The Johns Hopkins University.

Compte, O., F. Jenny, and P. Rey (2002): "Capacity constraints, mergers and collusion," *European Economic Review*, 46(1), 1–29.

Ellis, C. J., and W. W. Wilson (2001): "What Doesn't Kill us Makes us Stronger: Analysis of Corporate Leniency Policy.," Discussion paper, Department of Economics, University of Oregon.

Fung, K. C. (1992): "Economic Integration as Competitive Discipline," *International Economic Review*, 33(4), 837–847.

Ganslandt, M., L. Persson, and H. Vasconcelos (2008): "Asymmetric Cartels - a Theory of Ring Leaders," Discussion paper, CEPR Discussion Paper no. 6829.

Hammond, S. D. (2001): "When Calculating the Costs and Benefits of Applying for Corporate Amnesty, How do You Put a Price Tag on an Individual's Freedom?," in *The Fifteenth Annual National Institute on White Collar Crime*.

——— (2004): "Cornerstones of an Effective Leniency Program," in *ICN Workshop on Leniency Programs, Sydney, Australia*.

——— (2005): "Cracking Cartels With Leniency Programs," in *OECD Competition Committee, Working Party No. 3 Prosecutors Program*.

Harrington, J. (2006): "How Do Cartels Operate?," *Foundations and Trends in Microeconomics*, 2(1), 1–105.

——— (2008): "Optimal Corporate Leniency Programs," *The Journal of Industrial Economics*, 56(2), 215–246.

Harrington, J. E. (1991): "The Determination of Price and Output Quotas in a Heterogeneous Cartel," *International Economic Review*, 32(4), 767–792.

Harrington, J. E., and M. Chang (2009): "Modeling the Birth and Death of Cartels with an Application to Evaluating Competition Policy," *Journal of the European Economic Association*, 7(6), 1400–1435.

Harrington, J. J. E., and J. Chen (2005): "The Impact of the Corporate Leniency Program on Cartel Formation and the Cartel Price Path," Discussion paper, The Johns Hopkins University,Department of Economics.

Hinloopen, J., and A. R. Soetevent (2008): "Laboratory evidence on the effectiveness of corporate leniency programs," *The RAND Journal of Economics*, 39(2), 607–616.

Ivaldi, M., B. Jullien, P. Rey, P. Seabright, and J. Tirole (2003): "The Economics of Tacit Collusion," Discussion paper, IDEI, Toulouse.

Kuhn, K. U., and M. Motta (1999): "The Economics of Joint Dominance," Discussion paper, European University Institute.

Levenstein, M. C., and V. Y. Suslow (2006): "What Determines Cartel Success?," *Journal of Economic Literature*, 44(1), 43–95.

Miller, N. H. (2009): "Strategic Leniency and Cartel Enforcement," *American Economic Review*, 99(3), 750–768.

Motchenkova, E., and P. M. Kort (2006): "Analysis of Current Penalty Schemes for Violations of Antitrust Laws," *Journal of Optimization Theory and Applications*, 128(2), 431–451.

Motchenkova, E., and D. Leliefeld (2010): "Adverse Effects Of Corporate Leniency Programs In View Of Industry Asymmetry," *Journal of Applied Economic Sciences*, 5(2), 114–128.

Motchenkova, E., and R. Van der Laan (2005): "Strictness of Leniency Programs and Cartels of Asymmetric Firms," Discussion paper, TILEC Discussion Paper.

Motta, M., and M. Polo (2003): "Leniency programs and cartel prosecution," *International Journal of Industrial Organization*, 21(3), 347–379.

OECD (2002): *Fighting Hard Core Cartels: Harm, Effective Sanctions and Leniency Programmes*. Publications de l'OCDE.

Rothschild, R. (1999): "Cartel stability when costs are heterogeneous," *International Journal of Industrial Organization*, 17(5), 717–734.

Schmalensee, R. (1987): "Competitive advantage and collusive optima," *International Journal of Industrial Organization*, 5(4), 351–367.

Spagnolo, G. (2004): "Divide Et Impera: Optimal Leniency Programs," Discussion Paper No. 4840, CEPR Discussion Paper.

——— (2006): "Leniency and Whistleblowers in Antitrust," Discussion paper, CEPR Discussion Paper No. 5794.

Vasconcelos, H. (2005): "Tacit Collusion, Cost Asymmetries, and Mergers," *The RAND Journal of Economics*, 36(1), 39–62.

Veugelers, R., and H. Vandenbussche (1999): "European anti-dumping policy and the profitability of national and international collusion," *European Economic Review*, 43(1), 1–28.