Corporate Leniency and Incentives for Collusion

Tyra Merker

Master thesis in Economics
Department of Economics
University of Oslo

May 2015

Abstract

Leniency programs entail that the competition agency waives some fraction of the fines for a former cartel member, provided that it comes forward with information. I use a game theoretical collusion model from Harrington (2008), where the competition authority has an endogenous policy choice. The model is utilized to analyze how the degree of leniency impacts the firms' collusion decision. The result implies that it is in fact optimal to offer full amnesty to the first firm to come forward. Furthermore, when firms have private information about the probability of prosecution, the probability of whistle-blowing and conviction increases. This is provided that leniency is sufficiently generous.

Keywords: Collusion, Corporate Leniency Programs, Whistle-blowing, Industrial Organization, Game theory, Private information

2015

Corporate Leniency and Incentives for Collusion

Tyra Merker

http://www.duo.uio.no/

Print: Reprosentralen, Universitetet i Oslo

Preface

First and foremost, I would like to thank my supervisor Tore Nilssen for his support, valuable feedback and encouragement. I am truly grateful for his thorough efforts, which have raised the quality of my work.

I would also like to thank ESOP and the Department of Economics for providing me with a scholarship and office space for this past year. It has made the work with my thesis much more comfortable.

Finally, I would like to give a warm thank you to my friends and family. You have calmed me down during stressful exam periods - or at least attempted to do so - and always offered words of encouragement. Tom, thank you for our study groups and for giving feedback on a full draft of my thesis. Fernanda, thank you for putting things into perspective when I need it the most. And last, but certainly not least; to Yoseph, thank you for standing by me always, especially when I am at my worst.

Summary

This thesis analyzes the effect of corporate leniency programs on incentives for collusion. Participating in illegal collusion can result in large fines, possibly even imprisonment. Leniency programs entail that the competition agency waives some fraction of these fines for a cartel member, provided that it comes forward with information. Full amnesty from fines creates an incentive for cartel members to cooperate with anti-trust agencies, thus causing cartels to break down. On the other hand, it lowers the expected penalties the colluding firms must pay, which in turn can lead to more cartels being formed. The overall effect of leniency programs on collusion is therefore not immediately clear.

By utilizing a game theoretical model from Harrington (2008), I am able to assess how the incentives for collusion are affected by a leniency program. More specifically, I analyze how the degree of leniency, measured by the fraction of fines waived, impacts the firms' collusion decision. The model implies that it is in fact optimal to offer full amnesty, however leniency should only be awarded to the first firm to come forward. Allowing more firms to receive some degree of leniency can be justified if the first firm is unable to provide enough evidence to ensure a conviction.

I then turn to a model (Harrington and Chang, 2014) which incorporates more of the competition authorities actions, such as the size of penalties, prosecution and investigation efforts. The model shows how the cases created by the leniency program can crowd out resources spent on investigating active cartels. Thereby, the existence of a leniency program can actually *raise* the overall cartel rate. In order to prevent this, sufficiently large fines are needed.

Finally, I analyze the choice of whistle-blowing and the impact of information. The motivation is that leniency models generally feature symmetric firms, leading to the unrealistic result that either all or none of the firms apply for leniency. Using a model where firms have private information about the strength of the case (of the competition authority) against them, can lead to more realistic equilibrium outcomes. The model allows me to assess the firms' decisions under uncertainty, and demonstrates how the lack of information creates an incentive to pre-empt an opponent's application. As a result, the probability of whistle-blowing and conviction increases when firms have private information. This result holds provided that leniency is sufficiently large.

Contents

1 Introduction

There is a consensus that collusion is costly. When firms secretly agree to set high prices in order to earn higher profits, consumers must literally pay the price. It restricts supply and, in the long run, competitiveness and employment opportunities (European Competition Network, 2012). Firms may also collude in public tenders (e.g. through bid-rigging), thus making a large profit. In Norway, the so-called asphalt-cartel was able to charge the government several million NOK (Norwegian Kroner) in excess on single tenders, by dividing the market between each other and fixing prices (NRK, 2011). The cartel was unveiled when a cartel member alerted the competition authorities (CA), and the other cartel member was subsequently fined 140 million NOK (which later was reduced to 40 million NOK in a court of law) (Konkurransetilsynet, 2014).

As collusion is illegal in most countries, this illusive activity is hard to detect. Meetings between cartel members are held in secret, written agreements are rarely made and some cartels go to extraordinary lengths to keep their activities hidden. A creative example is the moon-phase cartel, which endured for several decades. According to their secret agreement, all firms entered high bids when selling electrical equipment to the US government in the 1960s. One predetermined winner would enter a slightly lower bid, thereby still making a large profit. The winner was determined by the phase of the moon at the day the auction was announced. Through this arrangement, the cartel was able to operate undetected for decades, as no meetings were necessary between its members. It is estimated that the cost to consumers was approximately 175 million US Dollars *each year* of the cartel's existence (Foundation for Economic Education, 1997).

Cartels are becoming increasingly sophisticated, and small countries like Norway face additional challenges when fighting them. Leading agents in an industry meet regularly through industry associations and other fora, as well as having studied at the same universities and colleges. Not only does it make collusion easier to commence, but personal relationships and loyalty can prevent collusion from breaking down (Meyer, 2011). Director of Konkurransetilsynet (the Norwegian CA), Christine Meyer (2011) points out that stable leader-groups in stable networks and associations have been the backbone of the largest and most damaging cartels in Norway.

1.1 Leniency Programs

A cartel conviction implies large fines for the firms and individuals involved, possibly even imprisonment. In order to facilitate the breakdown of cartels, the US Department of Justice offered their first leniency program in 1978. The Corporate Leniency Program was substantially revised in 1993, and the Leniency Policy for Individuals was in place in 1994. The first firm (and the first firm *only*) to come forward now automatically receives full amnesty from fines, and its individuals avoid personal fines, criminal convictions and prison terms. However, there is no amnesty from private lawsuits, for example put forth by victims seeking restitution (US Department of Justice, 2008). After the 1993 revision, a large number of leniency applications were submitted and the program was deemed a large success (Harrington and Chang, 2014). The Department of Justice has collected approximately 5 billion US Dollars in fines from cartel cases, where over 90 % of this figure is a result of leniency cases (Stolt-Nielsen, 2011).

In a shipping-cartel, a Norwegian company (Stolt-Nielsen) was granted leniency and its competitor and co-conspirator (Odfjell) was fined 42.5 million US Dollars by the US Department of Justice. Due to the US authorities' insistence that price-fixers should go to prison, the Chief Executive Officer and Vice President of Odfjell voluntarily travelled to the US in order to serve four month and three month prison sentences, respectively. The Chief Executive, Sjaastad, was also sentenced to pay a 250,000 US Dollar fine. Additionally, a large number of private antitrust lawsuits were filed against the cartel (Bloomberg, 2003).

Due to the positive experiences with the US Corporate Leniency Program, the EU implemented its Model Leniency Program in 1996 (which was revised in 2012). After 10 years of the EU leniency program, 24 of the 27 EU members and over 50 countries worldwide had implemented their own national programs. In Spain, cartel members literally lined up at the day of implementation, in order to be the first to apply for leniency. In Germany, 122 leniency applications were filed in a five year period (Harrington and Chang, 2014). Approximately 70 % of all EU cartel cases are results of leniency applications (Stolt-Nielsen, 2011). The EU program is similar to that of the US in the sense that only the first firm to apply for leniency can be granted *full leniency* (i.e. that all penalties are waived). However, in the EU several other firms may also receive reductions in fines by cooperating with the authorities and revealing all evidence. The European Commission can issue fines of up to 10 % of a firm's turnover

(European Competition Network, 2012).

In 2010, 11 cargo-carriers were convicted for price-fixing by the European Commission. The airlines were fined a total of almost 800 million Euros, after reductions in penalties. Lufthansa received full leniency, being the first of the airlines to contact the European Commission with information. However, *all* of the other firms received reductions, in a range between 10-50 % of the original fines. The Norwegian company SAS was sentenced to pay penalties of over 70 million Euros after a 15 % reduction. Additionally, private actors suffering damages from the cartel activity were able to raise their own lawsuits against all 11 airlines (European Commission, 2010). As recent as 2014, the German logistics company Schenker filed a lawsuit against the 11 airlines, demanding restitution (Dagens Næringsliv, 2014).

The Norwegian leniency program was implemented much later, namely in 2004, as an effort to coordinate antitrust policies in the EU/EEA.[1] The results, however, were left wanting. Within the first four years of the program, only two leniency applications were submitted. Neither of these resulted in further processing. As opposed to the US Corporate Leniency Program, the Norwegian leniency program only applies to corporate fines and not to criminal prosecution. Therefore, leniency applicants still risk large criminal fines and up to three years imprisonment if found guilty. In 2008 the Norwegian CA (Konkurransetilsynet) and The National Authority for Investigation and Prosecution of Economic and Environmental Crime (Økokrim) released a policy memo. It stated that Økokrim would not launch investigations on violations of the competition law (Konkurranseloven) unless Konkurransetilsynet first had pressed charges. In following years after the joint statement, the number of leniency applications has been rising. Despite the slow start, trust in the leniency program seems to be rising and more applications are consequently being submitted (Stolt-Nielsen, 2011). The discovery of the previously mentioned asphalt cartel was also a result of this leniency program.

[1] The EEA (European Economic Area) includes all EU countries, as well as Iceland, Norway and Liechtenstein. It allows these countries to have access to the EU's single market. The countries have coordinated their legislation in a number of areas, such as labor migration, trade and competition (GOV.UK, 2014).

1.2 Theoretical Background

How firms interact in an industry and make strategic choices is the focus of Industrial Organization. In the so-called ""bible" of Industrial Organization" (Ludwig Maximilian University of Munich, 2014), the author and recent Nobel laureate in Economics, Jean Tirole, devotes one chapter to the topic of collusion. The book was published in 1988, a decade after the implementation of the US' first leniency program.

Collusion models are generally game theoretical models of repeated games. Each period, firms in an industry meet and simultaneously set prices.[2] In each period, the competitive equilibrium leads to low pay-offs, as each firm has an incentive to undercut its competitors' prices. However, as the firms are meeting repeatedly in the same market, it is possible for them to coordinate their actions and charge higher prices. If a firm now undercuts, it will receive a one-time large pay-off, but will be punished as cooperation breaks down (and the industry reverts to the competitive equilibrium with low pay-offs). Whether firms are able to maintain collusion then depends on the sizes of profits (in all contingencies: collusion, competition and deviation), the number of firms in the industry and the firms' patience (Tirole, 1988).[3]

There are a number of extensions to the simple collusion set-up. There are several different punishment strategies which can be implemented, other than the so-called grim-trigger strategy described in the previous section. It is for example possible for collusion to be reinstated after a predetermined number of punishment-periods (in the grim-trigger strategy, these punishment-periods last forever). For collusion to be feasible, it is still needed that firms are not tempted to deviate. Additionally, they must now be *willing* to punish other deviators.[4] Furthermore, profits are not constant in reality, and may be affected by a number of factors (e.g. market fluctuations and business cycles). It is therefore possible to include stochasticly determined profits in collusion models. As a result, we can find that collusion is hardest to maintain at the beginning of a downturn (Tirole, 1988).

Due to the covert nature of collusion, empirical research on the topic is scarce. Also common for these models is that government policies are considered exogenous. In

[2] Alternatively, competition can be modelled as quantity competition, where firms choose how much of a good to produce. The price is then determined by the consumer demand.

[3] As it is a repeated game, it is assumed that firms discount future time periods. The further into the future a period is, the less they value the pay-off in that period.

[4] To punish a deviator, all firms will receive reduced profits during the punishment phase. The other firms must therefore have incentives to execute a punishment, in order for it to be a credible threat.

reality, however, there is clearly interplay between government agencies (such as a CA) and firms. More sophisticated models are needed in order to determine optimal public policies. In the case of collusion, we need to model the CA's policy choices as game theoretical strategies. To choose the optimal policies, we must also be clear on what should be optimized. For example, should a government aim at destabilizing cartels (in order to cause them to *desist*) or should they instead prevent cartels from *forming*? It is not immediately clear that both goals warrant the same optimal policy.

Despite the fact that the first leniency program was introduced in 1978 in the US, theoretical work on such programs were not produced before the beginning of the new millennium. Motta and Polo (2003) wrote a pioneering paper on leniency, providing a full game theoretical model on such programs (Harrington and Chang, 2014). As the more successful revised version of the US leniency program was not launched until 1993, this paper still came *10 years* after its introduction. It therefore seems that the theoretical assessment of leniency programs is falling, decades even, behind actual policies. With growing interest and proliferation of leniency programs, an increasingly larger body of research has been produced on the topic. The *fundamental question* to be answered is whether or not leniency programs reduce collusion. The main conclusion seems to be that they do (Motta and Polo, 2003; Spagnolo, 2004; Aubert et al., 2006; Harrington, 2008; Chen and Rey, 2013; Harrington and Chang, 2014).

1.3 Research questions and structure of the thesis

Economic research on the effects of leniency programs began a full decade after the programs were implemented. Considering the high cost collusion has for consumers and the public (European Competition Network, 2012), it is crucial that these effects are determined. As it is impossible to measure the actual cartel rate in a country empirically, it increases the need for comprehensive theoretical models. They would seem to be the best tool in order to determine the effect these policies actually have. On the one hand, firms now have an incentive to report the cartel to the CA, as they can avoid all penalties. This can cause cartels to break down. On the other hand, the expected punishment from part-taking in a cartel is reduced. This can make cartels more profitable, thereby increasing cartelization. It is not obvious which effect is stronger, and hence whether or not a leniency program reduces collusion (Harrington, 2008).

We can assess the opposing effects of the leniency programs by constructing a game theoretical model. As mentioned in the previous section, the general result of these models is that leniency programs reduce collusion. In order to identify the effects at play, I have chosen to base my analysis on a model presented by Harrington (2008). It isolates the CA's policy choices merely to the degree of leniency. This allows us to clearly identify the effects at play. Furthermore, we can define *optimal leniency*. Namely, we can answer the questions: Is it in fact optimal to offer a leniency program? If yes, how large should the fraction of fines waived ideally be?

Section 2 explains the mechanics of Harrington's model. I present which agents are at play, what choices they can make, the timing of their choices and what their preferences are. The main focus of the model is how a leniency program affects cartel *desistence*. As Harrington (2008) argues, if collusion breaks down more frequently, it reduces the incentive to collude *ex ante*. Thereby, if desistence occurs in enough contingencies, it will lead to deterrence.

In section 3, I start by presenting the main findings of the model. The main conclusion is that it is optimal to offer *full* leniency. As previously discussed, the US and European leniency programs differ in one important aspect. In the US, only the first firm to come forward receives leniency. However, in the EU and Norway, subsequent firms to cooperate with the CA will also receive some degree of leniency. I therefore enter into a discussion of which program structure is optimal in section 3.2. All leniency programs mentioned here state conditions for granting leniency (US Department of Justice, 2008; European Competition Network, 2012; Konkurranseloven, 2004). In many cases, the CA has already gathered some evidence against the cartel, but it might not be sufficient for a conviction. I therefore turn to the important question of when a CA should accept a leniency application and when it should reject it.

When discussing the results of the model, we must keep in mind the underlying assumptions and simplifications. Do these assumptions drive the results? Could optimal leniency change if the model includes more of the CA's policy choices (such as the magnitude of fines, investigation effort, prosecution efforts, etc.)? What would happen if the CA is resource constrained? Would the results be different if broken-down cartels can re-cartelize in the future? I address these questions in section 3.2.2.

As we are using theoretical models in order to assess real-life policies, it is important that they encompass important features of reality. Models of leniency programs generally consist of symmetric firms. As a result, either all or none of the firms apply for leniency. In reality, however, it is rarely the case that several firms simultaneously rush to the CA with leniency applications. It is reasonable to assume that firms have private information about (put differently; a perception of) the amount of evidence that the CA has against them. It is this perception, or signal, of the probability that they can be prosecuted that urges them to apply for leniency. Additionally, firms worry that their co-conspirators may apply for leniency first. This causes an incentive for pre-emption. Therefore, I ask: What drives *the choice of whistle-blowing*? How can the CA use this to its advantage? In section 4, I hence present a model for the choice of whistle-blowing, based on another model by Harrington (2013).

Finally, I summarize the main findings of my thesis in section 5 and revisit the questions raised in this section. I also offer some final remarks.

2 A Leniency Model

I will base my analysis on a model originally presented by Harrington (2008). The model aims to find optimal corporate leniency programs, in terms of the conditions and the extent of leniency granted.

The set-up of the model is based on a repeated Prisoners' Dilemma - collusion game with grim-trigger strategies as described in Tirole (1988). Competition can be thought of as a classical Bertrand price game. I will focus mainly on the collusive equilibrium of the model, as I aim to investigate how a leniency program may cause collusion to desist.

2.1 Players and their strategies

The agents in the model are the competition authority (CA) and the n firms in the industry.

We are interested in optimal leniency, in the sense of the *degree* of leniency granted by the CA. Therefore, the CA makes a policy choice of $\theta \ \epsilon[0,1]$, which is the fraction of fines F a firm pays when granted leniency. Thus, $1-\theta$ is the fraction of fines waived by the CA and also the degree of leniency. Let $\omega \ \epsilon(0,1]$ denote the probability that the CA opens an investigation in the industry. In the event that an investigation is launched (and no firm applies for leniency), $\rho \ \epsilon[0,1]$ is the probability that the CA successfully detects and prosecutes colluding firms. Note that $\rho = 0$ is equivalent to no investigation being launched. Therefore, there will be no loss in generality by only examining the equilibrium in the case of an investigation. We will assume that ω and ρ are exogenously determined,[5] and that ρ is randomly distributed according to a twice differentiable cdf $G : [0,1] \rightarrow [0,1]$.[6] We also assume that detection and conviction are certain when a firm applies for (and is granted) leniency and that only the first firm to apply will be granted leniency.[7]

The firms independently choose whether to collude (c) or not (nc). Subsequently,

[5]Realistically, the CA also chooses the size of the fine F, as well as being able to influence ω and ρ by resources invested in investigation efforts. In this analysis, I will however focus on the degree of leniency only, as the aim is to assess the leniency program itself. The size of the fine and efforts exerted in investigations will therefore not be a part of the evaluation here. I will however return to a short discussion of the CA's choices in section 3.

[6]Note that ρ is a probability, namely the probability of being detected. At the same time, ρ itself is also randomly distributed over the cumulative distribution function G, and may therefore take on a different value in each period.

[7]I will return to the possibility of multiple firms receiving leniency later in the thesis.

they choose whether to apply for leniency (l) or not to apply (nl). This choice occurs only if they in fact colluded in the *previous* time period. I will return to the timing of the game in the next subsection. If all firms collude, then each firm receives a pay-off of π^c. If a firm chooses to deviate from collusion while the other firms stick to the collusive agreement, then the deviating firm receives a profit of π^d. If none of the firms collude, then they all receive π^{nc}. We will assume that $\pi^d > \pi^c > \pi^{nc}$, which is in line with a price competition collusion game.

As the firms are playing a repeated Prisoners' Dilemma game with grim trigger strategies, detection or deviation from collusion will result in the Nash equilibrium forever, i.e. that each firm chooses nc (with pay-off π^{nc}) forever.

2.2 Timing

In period $t = 0$, the CA starts by making its policy choice of θ. With probability ω, it opens an investigation in the industry. If an investigation is opened, ρ is realized. After observing ρ (which is common knowledge amongst the firms) in the beginning of period $t = 1$, each firm chooses whether or not to collude. Period $t = 2$ again starts by a new value of ρ being realized.[8] After observing ρ, each firm again chooses whether or not to collude. Subsequently, they also choose whether or not to apply for leniency (for their actions in the *previous* period).[9]

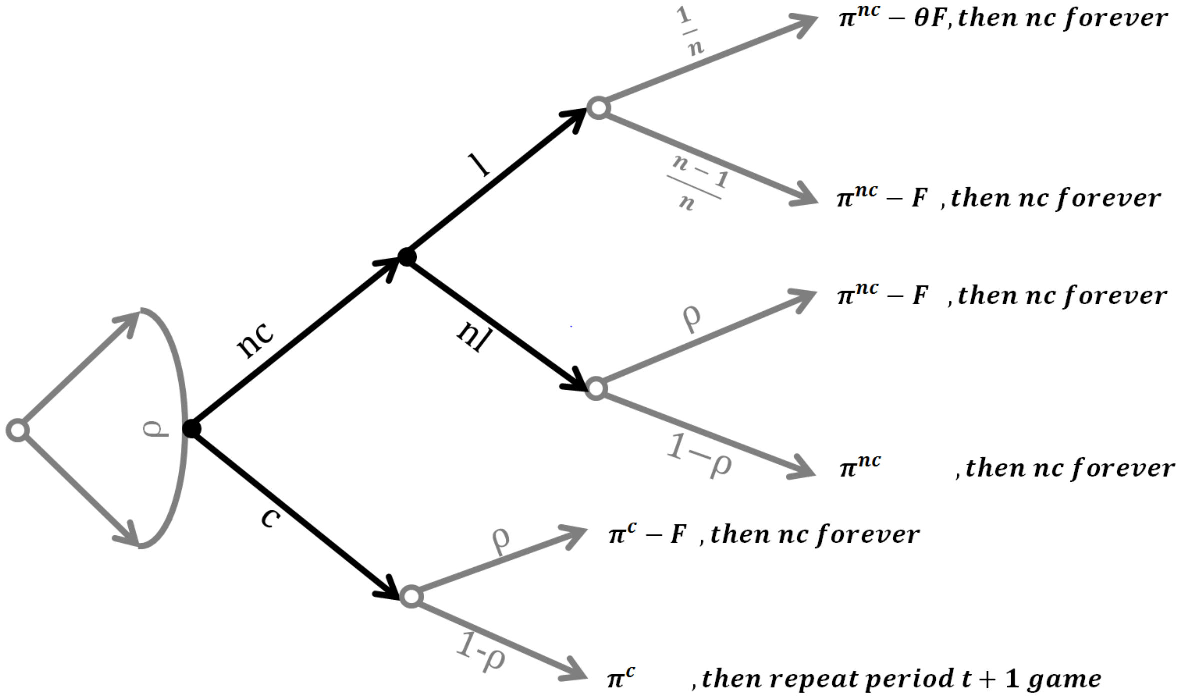
If firms choose the stage Nash (no collusion), they will remain there forever. If they choose to collude and are discovered and dissolved by the CA in the subsequent period, they also revert to the stage Nash equilibrium. However, if they choose to collude and are *not* discovered by the CA, the period repeats itself. Let us call this repeated period $t = t'$. It always begins with a new realization of ρ, followed by the collusion choice and the choice of a leniency application.

If ρ is low enough, firms will choose to collude. As $\pi^d > \pi^c$, it is clear that firms would choose to deviate (not collude) if they are planning to apply for leniency. This is because collusion would break down, and the (previously infinitely) repeated game now becomes finite. This then becomes the last period of the repeated game, where deviating from collusion is the dominant strategy. Therefore, we can assume that they

[8] As before, ρ is realized if an investigation is opened.

[9] This is a simplifying assumption of the model. However, Harrington (2008) argues that it is a plausible assumption nevertheless. Even though evidence depreciates over time, it is quite realistic that discovery and procecution is possible in the subsequent period.

Figure 1: Feasible strategies, probabilities and outcomes of period $t = t'$



will never apply for leniency if they have chosen to collude. In figure 1, I have depicted period $t = t'$ of the game.[10] The realization of ρ (in the beginning of the period) and the final pay-offs and outcomes are determined by probabilities (of getting an application for leniency approved, or of being discovered and prosecuted). I have depicted these probabilities in grey, as they are not to be confused with the firms' strategies. Finally, the firms *feasible* strategies are depicted in black.

In equilibrium, the probability of getting a leniency application approved is $\frac{1}{n}$. This is because it is a symmetric game, and if it is rational for one firm to apply for leniency, it is also rational for the $n-1$ other firms to do so.[11] If a firm applies for leniency (for the collusive actions in the previous period), it will therefore receive a pay-off of $\pi^{nc} - \theta F$ with probability $\frac{1}{n}$. On the other hand, it could also be the case that another firm is granted leniency instead. This occurs with probability $\frac{n-1}{n}$, and the firm must pay the full fine. Its pay-off in that contingency is $\pi^{nc} - F$. Thus, the expected pay-off when applying for leniency is:

$$\pi^{nc} - E\left[\theta F\right] = \pi^{nc} - (\frac{1}{n})\theta F - (\frac{n-1}{n})F = \pi^{nc} - (\frac{\theta + n - 1}{n})F$$

[10]Note that this figure is not intended as an extensive representation of the game (a so-called game tree), but simply a visualization of the timing in period $t = t'$.

[11]If at least one other firm applies for leniency, a firm can lower its expected penalty by applying for leniency as well, rather than paying the full fine F for sure.

2.3 Preferences and Incentive Compatibility Constraints

From figure 1 and backward induction, we can deduce that a firm will prefer to apply for leniency if the *expected* penalty when applying is lower than the expected penalty when not applying:

$$\underbrace{E\left[\theta F \mid l\right]}_{\text{Expected penalty when applying for leniency}} \leq \underbrace{E\left[F \mid nl\right]}_{\text{Expected penalty when not applying for leniency}}$$

$$\frac{1}{n}\theta F + \frac{n-1}{n}F \leq \rho F + (1-\rho)\cdot 0$$

$$\rho \geq \frac{\theta}{n} + \frac{n-1}{n}$$

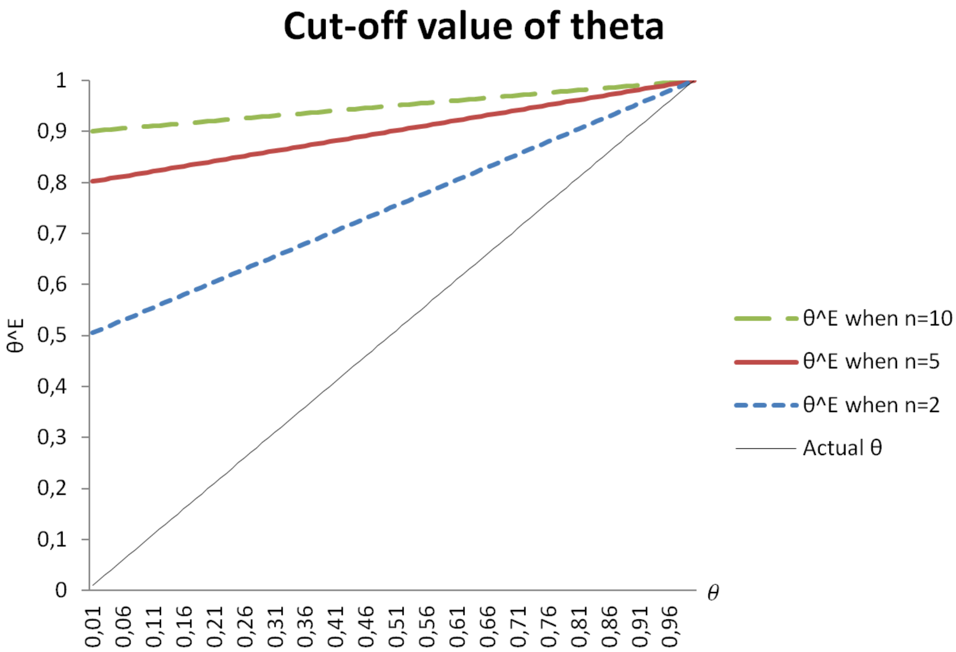
$$\rho \geq \frac{\theta + n - 1}{n} \equiv \theta^E,$$

where θ^E denotes the expected fraction of penalties paid in equilibrium. It is clear that $\theta^E\ \epsilon[\frac{1}{2}, 1]$.[12] We must therefore have a cut-off value of θ, which indicates whether or not firms prefer the equilibrium where they apply for leniency. The expression above and figure 2 show that θ^E increases with the number of firms. As the number of firms increases, the chance of being selected as the one firm to receive leniency diminishes. Therefore, it becomes less attractive for firms to apply for leniency. However, few firms applying for leniency does not necessarily imply that the existence of the leniency program is in vain, as it still may reduce the (unobservable) frequency of collusion in the industry.

Despite the fact that firms may prefer a situation where no one applies for leniency, it may not be an equilibrium. A commitment issue may arise, as deviating from a non-leniency agreement (by in fact applying for leniency when no other firm does) could yield a higher pay-off than continuing the agreement. In that case, not applying for leniency would no longer be sustainable as an equilibrium, despite yielding higher expected pay-offs. It therefore becomes clear that (given θ) we must have a cut-off value for ρ, determining whether or not firms prefer to collude. Let us denote this cut-off value by ρ^0. The realization of ρ subsequently determines whether or not firms apply for leniency. Harrington (2008) finds that there are three possible equilibrium strategies:

[12] $\theta^E = \frac{\theta+n-1}{n}$. Therefore, as we let $n \to 2$ and $n \to \infty$, we find that $\theta^E\ \epsilon\ [\frac{1+\theta}{2}, 1]$. Finally, as we know that $\theta \epsilon [0,1]$, we must have $\theta^E \epsilon [\frac{1}{2}, 1]$.

Figure 2: The cut-off value θ^E compared to the actual θ by cartel size



1. if $\rho \, \epsilon [0, \rho^0]$ then firms collude.[13]

2. if $\rho \, \epsilon \, (\max\{\rho^0, \theta\}, 1]$ then firms do not collude and apply for leniency.

3. if $\rho \, \epsilon (\rho^0, \theta]$ then firms do not collude and do not apply for leniency.

First, let us assess whether 2. and 3. are incentive compatible. We wish to determine if the three strategies above can be equilibrium strategies. We are therefore assessing whether there could be an incentive to *deviate* from them, and are not assessing the actual equilibrium pay-offs. Note that it is not optimal to act collusively if the other firms act competitively , therefore there is no incentive to deviate from nc.[14] If other firms apply for leniency, then prosecution is certain. Therefore, a firm will pay a lower expected penalty by also applying for leniency (i.e. $\theta^E F$) than paying F for sure. Thus, there is no incentive to deviate from 2. If the other firms do not apply for leniency and $\rho \, \epsilon (\rho^0, \theta]$, then the expected penalty by *not* applying for leniency (i.e. ρF) is lower than paying θF for sure.[15] A risk neutral firm again has no incentive to deviate from 3.

[13] As discussed previously, it is not rational to apply for leniency if a firm chooses to collude.

[14] The Nash equilibrium in this Prisoners' Dilemma game is for all players to play nc (no collusion) and the collusive equilibrium is for all firms to play c (collude).

[15] Note that we have $\rho \, \epsilon (\rho^0, \theta]$ and not $\rho \, \epsilon (\rho^0, \theta^E]$. If the condition in 3. were $\rho \, \epsilon (\rho^0, \theta^E]$ and $\theta \leq \rho \leq \theta^E$, then firms would have an incentive to deviate from not applying. Thus, 3. could no longer be an equilibrium.

It only remains to assess whether firms have an incentive to deviate from 1. Assume that firms discount future time periods by a rate of $\delta \epsilon [0,1]$.[16] Let $W \equiv \frac{\pi^{nc}}{1-\delta}$ denote the net present value of forever playing the stage Nash (with no collusion),[17] and let $E[V|\rho^0,\theta]$ denote the (expected) value of continued collusion. I will return to the value of continued collusion later in this section. For collusion to be maintained, i.e. for firms to choose strategy c in figure 1, we get the following incentive compatibility constraint (ICC):

$$\text{Life-time value of collusion} \geq \text{Life-time value of deviation}$$
$$\Downarrow$$
$$\pi^c + \underbrace{\delta(1-\rho)E[V|\rho^0,\theta]}_{\text{Value of future collusion if not detected}} + \underbrace{\delta\rho(W-F)}_{\text{Value if detected}} \geq \pi^d + \delta W - \delta \cdot \underbrace{min\{\rho,\theta\}F}_{\text{Expected penalty}} \quad (1)$$

On the left-hand side (LHS) of equation (1), we first find the pay-off from colluding in the present period, π^c. In the next period, there is a probability of $(1-\rho)$ that the cartel will be undetected, and thus continue to collude. With time-discounting, the pay-off from continued collusion hence becomes $\delta E[V|\rho^0,\theta]$, which is multiplied with the probability of this contingency being realized. However, the cartel may instead be detected and prosecuted in the following period, which occurs with probability ρ. The cartel then reverts to the Nash equilibrium with pay-off W and has to pay the fine F. With time-discounting and multiplying the probability of the event, this constitutes the last part of the LHS expression.

On the right-hand side (RHS) of equation (1), we first find the pay-off when deviating from the collusive agreement. As the firm considers breaking out of a collusive agreement, we assume that it is the only firm to do so, hence receiving the pay-off π^d. A deviation causes the cartel to break down for sure and the firms stay in the Nash equilibrium forever, starting from the next period. Therefore we discount the value W for one period. If the firm applies for leniency when deviating, it will be the only firm to do so (and a leniency application is granted for sure). A risk-neutral firm wishes to minimize the *expected* fine it has to pay, therefore weighing the risk of detection against the fine-reduction under leniency. If the risk of detection, ρ, is low, it

[16]In a different interpretation, δ can be considered a measure of the firms' patience. A higher δ then implies that firms are more patient, or equivalently that they discount the future less.

[17]This is because no collusion forever has a value of $W = \Sigma_{t=0}^{\infty}\delta^t\pi^{nc} = \frac{\pi^{nc}}{1-\delta}$.

will choose not to apply for leniency. On the other hand, if the probability of detection is high, it would rather pay the reduced fine θF for sure.[18] The last part of the RHS expression is therefore the min-function, where the deviating firm either plays strategy 2 or 3 above.

The *expected* value of future collusion ($E[V|\rho^0, \theta]$) depends on the cut-off value, ρ^0, for collusion to be maintained, and the CA's policy choice of θ. Using that $\rho = 0$ is equivalent to no investigation being launched, we have:[19]

$$E[V|\rho^0, \theta] = (1-\omega)V(0, \rho^0, \theta) + \omega \int_0^1 V(\rho, \rho^0, \theta) dG(\rho), \tag{2}$$

where

$$V(\rho, \rho^0, \theta) = \begin{cases} \pi^c + \delta(1-\rho)E[V|\rho^0, \theta] + \delta\rho(W - F), & 0 \leq \rho \leq \rho^0 \\ W - \left(\dfrac{n-1+\theta}{n}\right)\delta F, & max\{\rho^0, \theta\} < \rho \\ W - \rho\delta F, & \rho^0 < \rho \leq \theta \end{cases} \tag{3}$$

The expected value of future collusion in (2) is an expectation over ω (the probability of an investigation) and *all possible* values of ρ, as ρ is a stochastic variable. First, we have the value of collusion when no investigation occurs, which happens with probability $(1-\omega)$ and implies that $\rho = 0$. If an investigation is launched (with probability ω), ρ can take on any value in the interval $[0, 1]$. As ρ is continuously distributed by the cdf $G(\rho)$, the expectation over ρ becomes the integral in (2) by definition.

The three terms in (3) correspond to the three equilibrium strategies stated above, and represent the value of each strategy. The first line is equivalent to strategy 1., and to the LHS of equation (1), the incentive compatibility constraint. This is the lifetime value of colluding. The second line corresponds to strategy 2., where ρ is above the collusion threshold ρ^0, as well as the fraction of fines paid under the leniency program (θ). Therefore, firms stop colluding, receiving the value W. All firms apply

[18] As it is the only firm applying for leniency, the probability of having its application approved is now 1 (compared to $\frac{1}{n}$ before). The expected fine paid (if applying for leniency) is θF, as opposed to $\theta^E F$ when all firms apply simultaneously.

[19] The probability that no investigation is launched is $(1-\omega)$. In the situation where no investigation is launched, $\rho = 0$. This constitutes the first part of (2). ω is the probability that an investigation *is* launched, and the realization of ρ is stochastically determined. To determine the expectation of the (period-by-period) values of optimal strategies ($V(\rho, \rho^0, \theta)$), we integrate over the distribution of ρ (namely $G(\rho)$).

for leniency, and face either the reduced fine θF if it is granted, or the full fine F. As previously discussed, the expected fine paid then becomes $\theta^E F$.[20] As the fine is paid in the subsequent period, we again discount by δ. Finally, row 3 represents strategy 3., where the probability of detection is too high for collusion to be incentive compatible, but lower than the fine reduction if leniency is granted. As collusion breaks down, firms again receive W, but this time face an expected penalty of ρF in the next period.

By rewriting the ICC in equation (1), we can find the following expression:

$$\begin{aligned}\Phi(\rho, \rho^0, \theta) \equiv\ & \pi^c + \delta(1-\rho)E[V|\rho^0, \theta] + \delta\rho(W - F) - \pi^d - \delta W + \delta \cdot min\{\rho, \theta\}F \\ & \geq 0\end{aligned} \tag{4}$$

This means that collusion is maintained as long as $\Phi(\rho, \rho^0, \theta) \geq 0$. We can consider $\Phi(\rho, \rho^0, \theta)$ to be the margin between the pay-off from colluding and the pay-off from deviating from the collusive agreement. In this sense, a larger $\Phi(\rho, \rho^0, \theta)$ means that collusion is more stable, as the pay-off from deviating is much smaller compared to the collusive pay-off. On the other hand, if $\Phi(\rho, \rho^0, \theta) = 0$, a firm is indifferent between colluding and deviating. Finally, if $\Phi(\rho, \rho^0, \theta) < 0$ collusion is not sustainable, as the pay-off from deviation is larger than the pay-off from collusion, and collusion breaks down. As the realization of ρ is stochastic (and changes every period), it is interesting to examine how $\Phi(\rho, \rho^0, \theta)$ (the margin at which the ICC holds) changes with ρ. By differentiating $\Phi(\rho, \rho^0, \theta)$ with respect to ρ, we find that:[21]

$$\frac{\partial \Phi(\rho, \rho^0, \theta)}{\partial \rho} = \begin{cases} -\delta\{E[V|\rho^0, \theta] - W\} & , \rho \leq \theta \\ -\delta\{E[V|\rho^0, \theta] - W\} - \delta F & , \theta < \rho \end{cases}$$

$$\Downarrow$$

$$\frac{\partial \Phi(\rho, \rho^0, \theta)}{\partial \rho} < 0$$

As this derivative is strictly negative, it means that $\Phi(\rho, \rho^0, \theta)$ moves closer to zero as ρ increases. Intuitively, this means that if the probability of being detected and convicted increases, collusion becomes harder to maintain, as it raises expected penalties (thus decreasing the value of collusion).

[20] Where $\theta^E F = \left(\frac{n-1+\theta}{n}\right) F$.

[21] This assumes that $E[V|\rho^0, \theta] > W$, which must hold at the collusive cut-off (Harrington, 2008).

As discussed for the three equilibrium strategies, there must be a cut-off value for ρ, denoted by ρ^0, determining whether or not firms collude. Let us now assess what properties such a cut-off value must have. For a cut-off value, $\tilde{\rho}$, to be incentive compatible, it must satisfy the condition $\Phi(\tilde{\rho}, \tilde{\rho}, \theta) \geq 0$. As $\Phi(\rho, \rho^0, \theta)$ is decreasing in ρ, we therefore know that $\Phi(\rho, \tilde{\rho}, \theta) \geq 0 \; \forall \rho \leq \tilde{\rho}$. Let the optimal collusive cut-off ρ be:

$$\bar{\rho}(\theta) \equiv max\{\tilde{\rho} : \Phi(\tilde{\rho}, \tilde{\rho}, \theta) \geq 0\}$$
$$\Rightarrow \bar{\rho}(\theta) = max\{\tilde{\rho} : \Phi(\tilde{\rho}, \tilde{\rho}, \theta) = 0\}$$

The optimal cut-off ρ is the maximal ρ, that still satisfies equation (4) in the event that the *realized* ρ is equal to the cut-off itself (i.e. when $\rho = \rho^0$).[22] However, as we know that $\frac{\partial \Phi}{\partial \rho} < 0$, equation (4) will hold with equality in the maximum. As $\Phi(1, 1, \theta) = \left(\pi^c - \pi^d\right) - \delta \left(1 - \theta\right) F < 0$,[23] we also know that if a collusive equilibrium exists, then $\bar{\rho}(\theta) < 1$. Harrington (2008) shows (in Theorem 1) that if ω (the probability of an investigation) is sufficiently low and if δ (the firms' patience) is sufficiently high, then an optimal collusive cut-off equilibrium exists.[24] This implies that, if a collusive equilibrium exists, there is an optimal cut-off value for ρ, namely $\bar{\rho} \,\epsilon [0, 1]$. As $\bar{\rho} \,\epsilon [0, 1]$, it is possible for a period's realization of ρ to be either larger or smaller than the cut-off. Thus, even if a collusive equilibrium exists, collusion will still break down whenever the realized value of ρ is above this cut-off.

The very first realization of ρ determines whether or not a cartel *forms*. However, all subsequent realizations of ρ will determine whether or not an existing cartel *breaks down*. We can therefore, as does Harrington (2008), consider $\bar{\rho}(\theta)$ an index of cartel stability. It then remains for the CA to choose a degree of leniency, which destabilizes cartels as much as possible. This will be the topic of section 3.1.

[22] As $\pi^c > \pi^{nc}$, firms will prefer to set a cut-off value $\tilde{\rho}$ as high as possible, as this increases the number of contingencies (based on the realizations of ρ) in which collusion is sustainable.

[23] Recall that $\pi^d > \pi^c$ and that $\theta \,\epsilon [0, 1]$.

[24] For the full proof of existence, I refer to Harrington (2008).

3 Implications of the Model

As the CA has the ability to change the degree of leniency, θ, they may be able to reduce cartel stability. As previously mentioned, the CA may also care about the revenue they collect through fines (F), where the leniency program could reduce this income.[25] As the agency also is likely to have limited resources, weighing the number of investigations (affecting ω) and resources spent on prosecution (thus affecting the distribution of ρ, namely $G(\cdot)$) against the loss in revenue caused by the leniency program becomes an important issue. Perhaps even more important is how the loss in resources caused by the leniency program (and efforts exerted in pursuing leniency-cases) may limit investigations and prosecution in non-leniency cases. It is indeed possible that the existence of a leniency program may cause the overall cartel-rate to rise, by crowding out resources otherwise used for discovering and prosecuting (non-leniency) active cartels (Harrington and Chang, 2014). I will return to the discussion of limited CA-resources in section 3.2, also providing conditions under which a leniency program can reduce the instances of cartels.

Given that a CA has in fact opted for a leniency program, the question remains to what extent leniency should be granted. Assessing the optimal degree of leniency was the aim of the model presented in section 2. In section 3.1 we therefore assess the CA's optimal choice of θ, treating F, ω and $G(\cdot)$ as exogenous. Let us for this purpose assume that patience is sufficiently high and that the penalty fine F and the probability of an investigation ω are sufficiently low, such that a collusive equilibrium always exists. In other words, we now assume that there always is a possibility that collusion is optimal (in a contingency where ρ is low enough). The question now becomes how to destabilize it, causing it to break down.

3.1 Optimal Choice of Leniency Policy

Collusion breaks down when $\rho > \bar{\rho}(\theta)$. Therefore, the CA should aim at minimizing this collusion threshold through theta. Hence, we turn to finding the optimal policy θ^*:

[25] On the one hand, firms applying for leniency will get reduced fines, thus lowering the revenue collected by the CA. On the other hand, more collusion-cases may be discovered and prosecuted, as leniency leads to self-reporting and a sure conviction. The overall effect on revenue collected through fines is therefore ambiguous.

$$\theta^* \in argmin \ \{\bar{\rho}(\theta)\}$$

Harrington (2008) points out that, despite $\Phi(\tilde{\rho}, \tilde{\rho}, \theta)$ being continuous in $\tilde{\rho}$, it is not necessarily the case that solutions to $\Phi(\tilde{\rho}, \tilde{\rho}, \theta) = 0$ are continuous. As cartel stability is $\bar{\rho}(\theta) = max\{\tilde{\rho} : \Phi(\tilde{\rho}, \tilde{\rho}, \theta) = 0\}$, it may also be the case that $\bar{\rho}(\theta)$ is discontinuous. However, we can find the optimal policy by assessing how policy changes affect the ICC in equation (1) on the margin. Note that $\Phi(\bar{\rho}(\theta), \bar{\rho}(\theta), \theta)$ is the ICC with $\bar{\rho}(\theta)$ as the cut-off value for collusion used by the firms. At the same time, the *realized* value of ρ in that specific period is $\bar{\rho}(\theta)$. As the actual ρ in the period is equal to the cut-off, firms are indifferent between continuing to collude and to deviate (which would cause collusion to break down). This then becomes the ICC on the margin, and we want to assess if a change in θ will sway firms towards continued collusion or ceasing to do so.

$\Phi(\bar{\rho}(\theta), \bar{\rho}(\theta), \theta) = 0$ is a necessary condition for $\bar{\rho}(\theta)$ to be the optimal cut-off for the firms. By assessing how this condition (which defines $\bar{\rho}(\theta)$) changes with θ, we may be able to assess how $\bar{\rho}(\theta)$ is impacted by θ. Substituting for $\bar{\rho}(\theta)$ into the expression for $\Phi(\rho, \rho^0, \theta)$ in (4) we find:

$$\begin{aligned} \Phi(\bar{\rho}(\theta), \bar{\rho}(\theta), \theta) = 0 = &\pi^c + \delta(1 - \bar{\rho}(\theta))E[V|\bar{\rho}(\theta), \theta] + \delta\bar{\rho}(\theta)(W - F) \\ &- \pi^d - \delta W + \delta \cdot min\{\bar{\rho}(\theta), \theta\}F \end{aligned} \tag{5}$$

3.1.1 When leniency is large initially: $\theta < \bar{\rho}(\theta)$

Recall that the fractions of fines waived is $(1 - \theta)$, which can be considered the degree of leniency. When θ is small, the fraction of fines *paid* is small and the fraction of fines *waived* (and the degree of leniency) is large. Let us start by assessing policies such that $\theta < \bar{\rho}(\theta)$. As long as $\rho < \bar{\rho}(\theta)$, collusion will be sustained. However when $\rho = \bar{\rho}(\theta)$ a firm may consider deviating from the collusive arrangement. As $\theta < \bar{\rho}(\theta)$ implies that $\theta < \rho$,[26] a deviator will apply for leniency. The increase in the pay-off from deviation (caused by an increase in leniency/decrease in θ) is what Harrington (2008) calls "The Deviator Amnesty Effect" of the leniency program. It makes collusion harder to maintain, as more leniency raises the incentive to deviate from collusion. A change in θ will also affect the pay-off of continued collusion. If the firm continues to

[26] In this case the min-expression in (5) yields: $min\{\bar{\rho}(\theta), \theta\} = \theta$. The expected penalty when applying for leniency (θF) is smaller than the expected penalty from not applying ($\bar{\rho}(\theta)F = \rho F$). Thus a firm will apply for leniency if it deviates.

collude, θ does not affect the current-period pay-off.[27] However, we determined that $\bar{\rho}(\theta) < 1$, which means that there is a strictly positive probability that $\rho > \bar{\rho}(\theta)$ in some future time period.[28] This implies that there is also a positive probability that collusion breaks down in the future. As we know that $\theta < \bar{\rho}(\theta)$, therefore $\theta < \rho$ when collusion breaks down (as we will have $\theta < \bar{\rho}(\theta) < \rho$), and all firms apply for leniency. Firms know that they are likely to use the leniency program at some point in the future, and increased leniency lowers their expected fine. This raises the pay-off from collusion, and is what Harrington (2008) calls the "Cartel Amnesty Effect". To summarize, when $\theta < \bar{\rho}(\theta)$, there is a negative "Deviator Amnesty Effect" and a positive "Cartel Amnesty Effect" on collusion. It therefore remains to discuss which effect is the strongest.

Let $V^d = \pi^d + \delta W - \delta\theta F$ be the pay-off from deviating when $\theta < \bar{\rho}(\theta)$, as seen in (5) and discussed in the paragraph above. A marginal decrease in θ therefore increases the pay-off from deviating by $-\dfrac{\partial V^d}{\partial \theta} = \delta F$.[29] This is the "Deviator Amnesty Effect", which *in*creases the pay-off from deviation and therefore *de*creases the incentive to collude. All colluding firms will apply for leniency when collusion breaks down. After applying for leniency, each firm has a probability $\dfrac{1}{n}$ of being granted the fine reduction. The expected pay-off in that period therefore becomes $\pi^{nc} - \delta E[F] = \pi^{nc} - \delta\left[\dfrac{1}{n}\theta F + \dfrac{n-1}{n}F\right]$.[30] A marginal increase in leniency (and therefore a marginal decrease in θ) increases this pay-off by $\delta\dfrac{1}{n}F$. This is the "Cartel Amnesty Affect", which *in*creases the pay-off from collusion, thereby increasing the incentive to collude.

As these two effects have opposite impacts on collusion, we can now compare the sizes of the marginal effects in order to determine the overall effect of the leniency program. It is clear that the "Deviator Amnesty Effect" is the largest, as $\delta F > \delta^{x+1}\dfrac{1}{n}F$, assuming that collusion breaks down in x periods from now. This holds even if firms are infinitely patient,[31] in which case we have $F > \dfrac{1}{n}F$. We therefore find that increased

[27]Recall that it is irrational for firms to apply for leniency if they are colluding, as they could receive a strictly higher pay-off by deviating. This is because $\pi^d > \pi^c$.

[28]$\rho\ \epsilon[0,1]$ and has a cdf $G(\rho)$ over the interval. As $\bar{\rho}(\theta) < 1$, it implies that $1 - G(\bar{\rho}(\theta)) > 0$. For all values $\rho > \bar{\rho}(\theta)$, collusion breaks down and is not incentive compatible.

[29]I choose to assess a marginal effect from *reducing* θ, as this corresponds to *increasing* leniency.

[30]As collusion has broken down, they return to the stage Nash, with the period-by-period pay-off π^{nc}. As all firms apply for leniency, discovery of the cartel is certain. Therefore the firms have an expectation of the fine they will pay, where they pay θF if they are granted leniency and the full fine, F, when they are not.

[31]If firms are infinitely patient, they do not discount future periods, no matter how far into the future they are. Mathematically, it means that $\delta = 1$, and firms care equally about this period as they do any other time period.

leniency always decreases collusion when $\theta < \bar{\rho}(\theta)$.

Intuitively, the leniency affects deviation directly (and in the current time period). As the deviating firm is the only one to apply for leniency, its application is granted for sure and it enjoys the full reduction in fines. On the other hand, if a firm waits until collusion breaks down (in some possibly distant time period), all firms simultaneously apply for leniency. Getting a leniency application approved is no longer guaranteed, thus (in expectations) the firm no longer receives the full fine reduction.

3.1.2 When leniency is small initially: $\bar{\rho}(\theta) < \theta$

Now we have that $\min\{\bar{\rho}(\theta), \theta\} = \bar{\rho}(\theta)$ in (5). Therefore, the pay-off from deviation now becomes $V^d = \pi^d + \delta W - \rho\theta F$, which is unaffected by a change in θ. Leniency is too low for firms to apply, as they would rather face the risk of being detected (which yields a lower expected punishment). A marginal change in leniency does not affect this choice, and therefore does not affect the pay-off from deviation. In this case, there is no "Deviator Amnesty Effect". As previously discussed, an increase in leniency always increases the collusive pay-off, thereby having a positive effect on the cartel-rate. With no change in the deviation-pay-off and the "Cartel Amnesty Effect" being the only one of the two forces at play, an increase in leniency will increase collusion.

However, there is a third effect, which becomes relevant in this situation. The expected value of future collusion ($E[V|\bar{\rho}(\theta), \theta]$) depends on how large the expected fine will be.[32] To calculate this expectation, recall the integral from equation (2).[33] In the previous section, it was always the case that $\theta < \rho$ whenever collusion would break down in the future. As all firms "automatically" would apply for collusion, a change in θ had a straightforward effect on the expected fine, namely, the expected fine would be lowered by a factor of $\frac{1}{n}$.

Now that $\bar{\rho}(\theta) < \theta$, this is no longer the case, and there are two possible ways in which collusion can break down.[34] It may be the case that the realization of ρ in the break-down-period is smaller than θ, as is the case for all $\rho \,\epsilon (\bar{\rho}(\theta), \theta]$. Collusion would still break down, but no firm applies for leniency, thereby choosing the lowest

[32]Note that this is the marginal case where $\rho = \bar{\rho}(\theta)$. However, the expected value of *future* collusion is an expectation over *all future contingencies*, that is, an expectation over all possible values that the realization of ρ may take.

[33]There, we integrated the pay-off for each equilibrium strategy over the interval of ρ where each strategy is applicable (weighted by the cdf $G(\rho)$).

[34]Clearly, whenever collusion breaks down, we always have $\bar{\rho}(\theta) < \rho$, as before.

expected fine, ρF. Finally, there is also the interval $\rho\ \epsilon(\theta, 1]$, where collusion breaks down and firms apply for leniency. Recall that all firms apply for leniency when $\theta < \rho$, however the expected penalty paid is $\theta^E F$ (as opposed to θF). We already established that $\theta^E > \theta$ for $n \geq 2$. This means that there is a jump in the expected fine at $\rho = \theta$.

In the contingency where $\rho \simeq \theta$, we are in a situation where firms stop colluding and do *not* apply for leniency. If the CA now chooses to slightly lower θ to $\theta' = \theta - \varepsilon$, all firms strictly prefer to apply for leniency. However, due to symmetry of the game, this now raises their expected penalty to $\theta^E F$. This is what Harrington (2008) refers to as the "Race to the Courthouse Effect", effectively raising expected fines and thus reducing the pay-off to collusion when leniency is increased. The size of the increase in expected fines is approximately $\theta^E F - \theta F = \left(\frac{n-1}{n}\right) F\,(1-\theta)$, using that $\theta' \simeq \theta$ when ε is small.[35]

In figure 3, we can see the "Race to the Courthouse Effect". I have depicted the expected penalties faced by the firms. A reduction in θ causes the discontinuity to shift further left, thereby increasing the number of contingencies where firms face a high penalty. The expected fine jumps from the lower "x" to the higher "x", reducing the expected pay-off of continued collusion. At the same time, we can also see the "Cartel Amnesty Effect", as $\theta^E(\theta') < \theta^E(\theta)$. The change in θ causes the expected penalties to shift down in the contingencies where firms apply for leniency (i.e. when $\theta < \rho$). This can be seen by the vertical shift from $\theta^E(\theta)$ to $\theta^E(\theta')$, and it increases the expected pay-off from collusion.

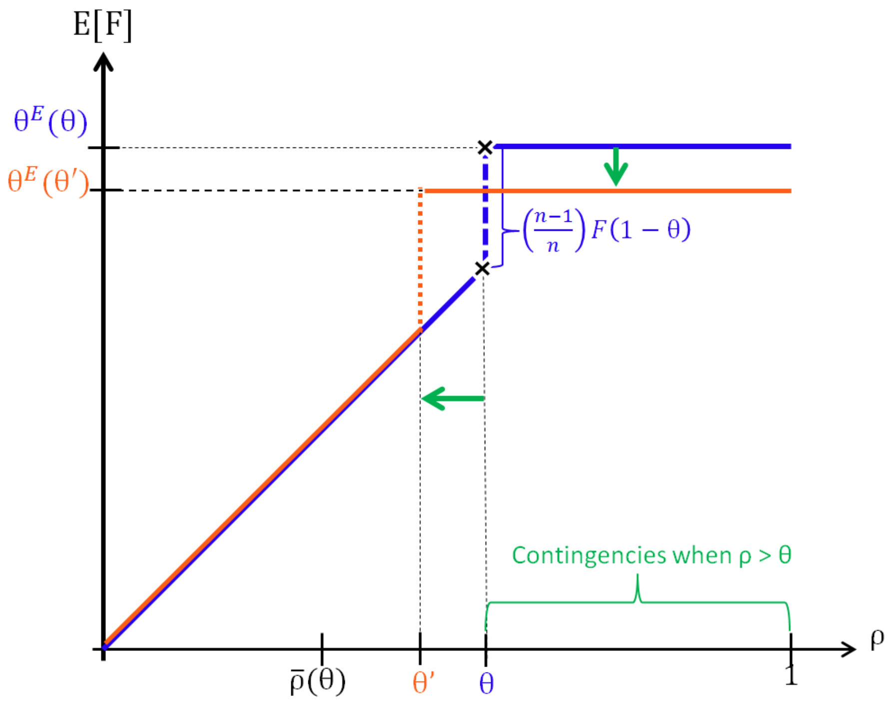
As we have two opposite effects on the collusive pay-off, the question remains whether more leniency increases or decreases collusion when $\bar{\rho}(\theta) < \theta$. Harrington (2008) provides us with the necessary condition on the hazard rate, for which $\bar{\rho}(\theta)$ is increasing in θ.[36] If $\frac{G'(\rho)}{1-G(\rho)} > \frac{1}{(n-1)(1-\rho)} \forall \rho\ \epsilon[0, 1)$, then θ^* exists and $\theta^* = 0$ (Harrington 2008, p.220). Furthermore, for this condition to be satisfied, it is sufficient that $G''(\rho) < 0\ \forall \rho\ \epsilon[0, 1]$ (Harrington 2008, p.230).[37] $G''(\rho)$ describes how the density function of ρ (the so-called pdf) changes. A decreasing density function means that the larger the value of ρ, the less likely the observation becomes. It is noteworthy that

[35]This follows from $\left(\frac{n-1+\theta-\varepsilon}{n}\right) F - \theta F \simeq \left(\frac{n-1+\theta}{n} - \theta\right) F = \left(\frac{n-1-(n-1)\theta}{n}\right) F = \left(\frac{n-1}{n}\right) F(1-\theta)$.

[36]More leniency means a lower θ, and thus a lower $\bar{\rho}(\theta)$. Therefore, more leniency would then lead to less collusion, as a lower cut-off ($\bar{\rho}(\theta)$) implies less contingencies in which collusion can be maintained.

[37]For the full mathematical proofs of these two conditions, I refer to the appendix in Harrington (2008).

Figure 3: The "Race to the Courthouse Effect" and the jump in expected penalties



a strictly decreasing density function is not a common distribution function. Even distributions that are extremely left-skewed would still be *in*creasing in the lowest range of ρ. Harrington (2008) argues that as ρ is the probability of discovery and successful prosecution (when no firm applies for leniency), high success-rates are increasingly harder to accomplish. Given the covert nature of collusion, it is not implausible that this in fact the is case. Harrington (2008) also numerically explores whether less than full optimal leniency could be optimal.[38] However, he concludes that in all reasonable cases, full leniency is optimal or very close to optimal.

3.2 Further leniency considerations

We have now established that full leniency is optimal for the first firm to come forward. In the following section, we will stay within the framework of the model, and assess in which situations leniency applications should be granted. When the CA has opened an investigation, it will collect some evidence against the cartel. However, a conviction is not guaranteed, and inside information from a cartel member can help ensure a successful prosecution. On the other hand, as the CA already has collected information

[38]He uses several specifications of $G(\rho)$, including a symmetric triangular distribution. In the latter case, $G''(\rho) > 0$ for $\rho \, \epsilon [0, 0.5]$, yet $\theta^* = 0$ is still the unique optimum when a collusive equilibrium exists.

on its own, it is not evident that the inside information always would be crucial to the case. We therefore need to find the optimal conditions for when to accept a leniency application. This will be the discussed in 3.2.1, and we will start the section by assessing the option of granting leniency to more than one firm.

Our model limits the CA's choices to the degree of leniency. In 3.2.2 we therefore attempt to look beyond the simplifications of this model, and discuss other CA policy considerations. In reality, the CA may choose (or at least influence) the size of the fine F, the resources spent on investigations and their prosecution efforts. Including these choices, as well as the fact that CAs generally face resource constraints, is the main focus of the section. It then becomes important how the leniency program impacts overall collusion in a range of industries.

3.2.1 Conditions for granting leniency

In the model we have presented here, it is assumed that leniency is only granted if no other firm has applied previously. It means that only the first firm to come forward receives leniency. However, this does not correspond well with EU practice, where also a second firm may receive some leniency (Harrington, 2008). Also in Norway, several firms may receive leniency. Only the first firm will be granted *full* leniency, but partial leniency of 30-50 % may be offered to the second firm to come forward, and leniency of 20-30 % can be offered to all subsequent firms contributing evidence in an investigation (Konkurranseloven, 2004).

A possibility of several firms receiving leniency can be represented in our model by turning the penalty reduction θ into a vector, where $\theta = (\theta_1, \theta_2, ..., \theta_n)$ and $\theta_j \epsilon [0, 1]$ $\forall j = 1, 2, ..., n$. $\theta_1 F$ is the fine paid by the first firm and $\theta_j F$ is the fine paid by the jth firm to come forward. In the case where only the first firm to come forward receives (full) leniency, we therefore have that $\theta = (0, 1, 1, ..., 1)$, which is in fact optimal. To see this, note that the pay-off from deviating (from the collusive arrangement) is only affected by θ_1. Therefore, the "Deviator Amnesty Effect" is unaffected by θ_j for $j = 2, 3, ..., n$.[39] All elements of the vector θ will, however, affect the collusive pay-off in (2). Setting $\theta_j < 1$ for $j = 2, 3, ..., n$, i.e. granting some degree of leniency to subsequent firms to come forward, will *reduce* the expected punishments and *increase* the collusive pay-off. It therefore increases the "Cartel Amnesty Effect" and makes

[39] This is because the deviating firm would be the first to come forward, therefore receiving the penalty $\theta_1 F$. The penalties faced by other firms do not affect the pay-off from deviation.

collusion easier to maintain. In our model, it is therefore rational to grant full leniency to the first firm to come forward, and no leniency to all others.

The model in chapter 2 assumes that prosecution is certain ($\rho = 1$) when a firm applies for (and is granted) leniency. In reality, the evidence presented by the firm may raise ρ and be enough for the CA to secure search warrants, however prosecution is not guaranteed. By allowing subsequent firms to receive partial leniency for further evidence, the CA can strengthen its case and raise the probability of convicting other cartel members. As this aspect is not incorporated in our model, we are unable to assess the overall effect of offering leniency to more than one firm (as practiced in the EU and Norway). It remains an open question if some degree of leniency for several firms could be optimal. This is an area in need of further research, as simple leniency models are unable to answer the question at hand.

In reality, it is hard to achieve certainty in prosecution. A natural question then arises about the requirements for getting a leniency application approved. The Norwegian Competition Law (Konkurranseloven, 2004) states that the evidence put forward by the firm must be sufficient for the CA to either issue a search warrant or to prove collusion. In either case, the CA itself must not have sufficient evidence at the time of application. This is equivalent to the condition in the leniency program of the European Competition Network (ECN), which is applicable in the jurisdiction of the European Commission (European Competition Network, 2012). The conditions for leniency in the US are also similar. As stated in the Department of Justice's Corporate Leniency Program, it is required that "the Division does not have evidence against the company that is likely to result in a sustainable conviction" (US Department of Justice, 2008, p. 5).

The question now remains: what constitutes sufficient evidence? Note that this is a question of how much *of their own* evidence is enough for the CA in order to reject a leniency application. As ρ is the probability of successful prosecution without leniency, we can consider this a measure of the CA's evidence against the cartel. We can therefore find a cut-off value $\hat{\rho}$ *chosen by the CA*. If $\rho < \hat{\rho}$, it implies that the CA's case is weak and it therefore accepts the leniency application. Conversely, if the CA already has collected "sufficient evidence" at the time of the leniency application, i.e. when $\rho > \hat{\rho}$, the CA should *not* grant leniency. This implies that the CA now has two choices; the degree of leniency (θ) and the cut-off for when to offer leniency ($\hat{\rho}$). We therefore need to characterize the optimal choice of $\hat{\rho}$, namely $\hat{\rho}^*$.

Harrington (2008, Theorem 5) finds that:

$$\hat{\rho}^* = \frac{n - 1 + \theta^*}{n}$$

and that it is optimal to award leniency if and only if $\rho < \hat{\rho}$. If we additionally assume that the density function of ρ is non-increasing (as in section 3.1.2), he shows that the equilibrium choices of the CA exist and are:

$$(\theta^*, \hat{\rho}^*) = \left(0, \frac{n-1}{n}\right)$$

This tells us that the CA should always grant leniency when the probability of conviction is below 50%.[40] It also implies that the leniency program should not be too strict when granting leniency, especially when larger numbers of firms are involved in collusion. For example, when a cartel consists of 10 firms, it is optimal to grant leniency as long as the probability of conviction is below 90%. This is caused by the "Race to the Courthouse Effect", which becomes stronger as the number of firms increases.[41]

If there are limits to how much evidence a single firm can produce, we again return to the question of whether it could be optimal to allow a second firm to come forward (and possibly also subsequent firms) to receive partial leniency. It is a limitation of the model that it is unable to assess the effects of the European leniency programs, where more than one firm can be granted some degree of leniency. According to the model presented here, it is *not* optimal to offer leniency to more than one firm. The structure of the American leniency program would therefore seem optimal.

Furthermore, we must keep in mind that the cut-off level of evidence ($\hat{\rho}^*$) is chosen to destabilize cartels most efficiently. This does not take into account the revenue the CA can collect though fines. Motta and Polo (2003) include a budget constraint for the CA in their pioneering leniency model. They argue that, if there were not budgetary constraints, it would be optimal for the CA to put more resources into investigations. Thereby, they could increase the probability of prosecution and no leniency program

[40] A cartel must consist of at least two firms. If $n = 2$, then $\hat{\rho}^* = \frac{1}{2} = 50\%$.

[41] Recall that the jump in expected penalties is $\left(\frac{n-1}{n}\right)(1-\theta)F$. We can find the change in the jump's magnitude by: $\frac{\partial}{\partial n}\left[\frac{n-1}{n}(1-\theta)F\right] = (1-\theta)F\frac{n-(n-1)}{n^2} = (1-\theta)F\frac{1}{n^2} > 0$. We can therefore see that the jump in expected penalties is strictly increasing in the number of firms.

would be offered. They point out that a leniency program is only optimal as a second best option, when the CA has limited resources. The welfare loss caused by the leniency program (e.g. through lost revenues from fines) needs to be weighed against the welfare gain from deterring collusion (or causing it to desist).

3.2.2 Resources of the Competition Authority

In reality, a CA has more choices than the degree of leniency alone. It can influence how much resources are put into investigations and prosecution. This would then affect ω and $G(\rho)$ from our model in chapter 2. Furthermore, it can also choose (or at least influence) the size of fines paid by convicted cartel members. Finally, it has budgetary restrictions and therefore cares about how much resources it spends during investigations and prosecutions. It must weigh the costs and effects of different efforts against each other. The loss of resources from a leniency program (as the CA collects less revenue through fines) must be weighed against the welfare benefits from cartel reduction. On the other hand, the leniency program may also save resources, as it makes prosecution shorter and easier. Motta and Polo (2003) find that overall, a leniency program is optimal when the CA has limited resources, and that full leniency should be granted (as does Harrington (2008)).

It is unrealistic to assume that the amount of resources spent on prosecution does not affect investigation efforts. Furthermore, it seems unreasonable that a cartel member applies for leniency when the cartel is well-functioning. In our model, we incorporated this aspect in the sense that no firm applies for leniency when it colludes. Collusion stops due to a stochastic realization of ρ, which then causes collusion to desist. The CA was able to influence the cut-off ρ, which causes collusion to break down. In this sense, the break-down of the cartel still occurs prior to a leniency application. If the prosecution of leniency cases is costly, the question remains if it could be more efficient to spend more resources on discovering functioning cartels and less on prosecuting those which have broken down on their own.[42]

Harrington and Chang (2014) present a model to assess this resource trade-off. They introduce stochastic industry profits and allow for re-cartelization. Therefore, there no longer are grim-trigger strategies such as in the model in chapter 2. If the realization of one period's profit becomes too high, deviation becomes so profitable

[42] If the goal of a CA is to minimize the losses caused by collusion, efficiency can be considered achieving the lowest possible cartel-rate.

that the cartel breaks down. Cartels may also break down if the CA discovers and successfully prosecutes the cartel members. In this way, there arises a life- and death-cycle for cartels. The authors also allow for differences in cartel frequencies between industries, by allowing cartel stability to vary across industries.[43]

In this model, the CA also has more policy choices. They can still choose θ, the degree of leniency. Due to previous results, it is assumed throughout the model that only the first firm receives leniency, and that there is full leniency (i.e. $\theta = 1$). However, the CA may also choose the size of the fine F, which now is represented as a fraction of the profit increase.[44] Finally, the CA can also affect the probability of discovery and prosecution, denoted by σ. It is still assumed that leniency cases lead to prosecution for sure. Therefore, σ can be considered the non-leniency efforts of the CA (i.e. the efforts of discovering and prosecuting active cartels). σ is affected by an exogenous probability of discovery, the fraction of firms investigated and the probability of winning a case. The latter will be adversely affected by a large caseload, which is the number of non-leniency cases and (to a somewhat lesser extent) the number of leniency cases. To sum up, the main policy choices of the CA are the fine paid by cartel members, the fraction of firms investigated and its own caseload.

The goal of the CA is to maximize welfare, which in this context is equivalent to minimizing the overall cartel rate. The interesting aspect of this model is how a seemingly successful leniency program actually may raise the overall cartel rate. A leniency program may be perceived as successful if many firms apply for leniency. However, the applicants are members of already broken-down cartels, and the large caseload from these leniency cases may crowd out non-leniency efforts. Therefore, less is done to discover and prosecute active cartels, and more cartels are able to form and/or last longer. Harrington and Chang (2014) find that a leniency is counter-productive if the leniency cases require much resources,[45] and if the penalties are not sufficiently large to deter cartel formation.

These results underpin the conditions for granting leniency. It is necessary that the evidence presented by the leniency applicant is strong enough, in order to significantly

[43]This can be caused by differences in demand elasticities in the industries. The industry heterogeneity affects the frequencies of the cartels, but not the firms' profit streams.

[44]Colluding firms earn higher profits than competing firms, as established earlier. In this model, the CA chooses a fine which is proportional to the gains from collusion. The more firms gain by colluding, the larger the fine they face if discovered.

[45]In this context, it means that they require close to the same amount of resources as non-leniency cases.

reduce the resources spent on the case. The model also puts forth another important consideration for leniency programs. If the penalty fines for collusion are too low, it might be more efficient not to implement the leniency program in the first place.

As Harrington and Chang (2014) include an industry-specific parameter for cartel stability, they are able to assess differences in the effect of the leniency program. They find the following property: "A leniency program generally reduces the range of markets that are able to form cartels. The effect of a leniency program on average cartel duration is decreasing in η [the parameter for cartel stability] so that markets with less stable cartels experience a bigger decline in average cartel duration. This differential can be so significant that a leniency program reduces average cartel duration of relatively unstable cartels and, at the same time, increases average cartel duration of relatively stable cartels." (Harrington and Chang, 2014, p. 23-24).

The model makes a simplifying assumption that η only affects cartel stability, but not the firms' profit streams. It also assumes that the profit increase from collusion is the same fraction in all industries.[46] However, if there is a correlation between cartel stability and the profit increase from collusion, then the cartels that are strengthened by the leniency program are precisely those that cause the most damage. If this is indeed the case, perhaps it is not appropriate to minimize the cartel rate in order to maximize welfare.

Its is noteworthy that all but one of the results in Harrington and Chang (2014) depend on the assumption that there is full leniency, which is offered only to the first firm to come forward. The model is quite complex, and turning the degree of leniency (θ) into a vector could make the calculations intractable. Offering partial leniency to several subsequent firms could however be modeled as reducing the caseload of the leniency case.[47] In this sense, allowing more than one firm to receive leniency could help ensure that the introduction of the leniency program reduces the overall cartel rate. Conversely, allowing more firms to receive (some degree of) leniency reduces the expected fine from colluding. As previously discussed, this raises the pay-off from colluding, which in turn can raise the cartel rate. The effects of allowing more than one firm to receive leniency therefore remain unclear.

This model clearly incorporates more of the policy choices a CA can make, making

[46] As before, this increase refers to the difference between the collusive and the competitive profit.

[47] In the model, a leniency case increases the caseload only by a fraction λ of a non-leniency case. Allowing more than one firm to receive leniency could be modeled as reducing λ beyond the one-firm leniency situation.

it more realistic. However, I find it reasonable to assume that income generated through fines has an impact on the CA's resources. As mentioned in section 1.1, over 90 % of the fines from cartel cases in the US came from leniency cases (Stolt-Nielsen, 2011). Despite requiring resources, leniency programs may at the same time generate resources for a CA. How the resources generated compare to resources spent, is not immediately clear. Furthermore, resources are not only monetary, but also time and efforts of the CA staff. However, we can expect that revenue generated by fines has a dampening effect on the crowding out of resources, making it less severe than assumed in this model.

4 Whistle-blowing

I will base this section on a simple model of whistle-blowing in the leniency context. The model is presented by Harrington (2013), and portrays the more realistic situation when firms have private information. Firms receive private signals about the strength of the CA's case. They therefore have to make a choice based not only on their own signal, but also on whether or not they believe that the other cartel member will apply for leniency. The set-up of the model is not unlike auction models where bidders have affiliated valuations (Klemperer, 1999). In order to assess the effect of private information, a comparison will be made with the baseline case of full information.

As previously discussed, a result of leniency models is that either all or none of the firms apply for leniency. This, however, is quite far from reality, where it is most common that a single firm applies. A proper model of whistle-blowing therefore allows us to better understand the choice firms make. By understanding why firms do or do not apply for leniency, there is also the possibility to influence these choices. In this section, I first present the structure of the model and then turn to a discussion of its results.

4.1 Structure of the model

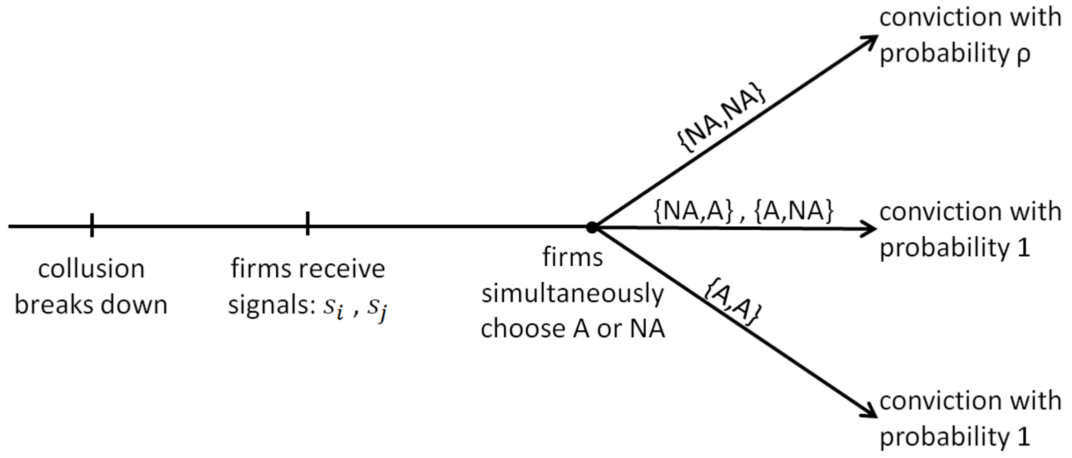
As previously, we assume that firms only apply for leniency when collusion has broken down.[48] In this model, there are two firms, whose only choice is whether or not to apply for leniency. Each firm receives a private signal about $\rho \, \epsilon [0, 1)$, the probability that they will be discovered and convicted. Let $s_i \, \epsilon [\underline{s}, \bar{s}]$ denote the signal that is received by firm i. Conditional on its own signal, firm i assigns probabilities to all possible values of the *other* firm's signal. Let us denote this conditional cdf as $H(s_j|s_i)$.[49] We assume that the signals are informative about ρ, and that they therefore are correlated:

A1 $H(s_j|s_i)$ $(i \neq j)$ is continuously differentiable in s_i and s_j. If $s_i'' > s_i'$, then $H(\cdot|s_i'')$ weakly first-order stochastically dominates $H(\cdot|s_i')$. If firm i receives a high

[48]Recall that it is irrational to continue colluding while applying for leniency. The reason for this is that a firm could increase its profit by deviating. Therefore, it would choose not to collude when it applies for leniency. In that sense, collusion has already broken down when a firm applies for leniency.

[49]This means that $H(s_j|s_i)$ is the probability that the *other* firm's signal is smaller than (or equal to) s_j, *given* that firm i has received the signal s_i.

Figure 4: Time line for the model of whistle-blowing



signal, it is more likely that firm j also received a high signal (compared to if firm i had received a low signal).

Based on the signals the firms have received, they simultaneously make a choice of whether or not to apply for leniency. Based on signal s_i, firm i has the strategy $\phi(s_i)$, which results in the choice of either "Apply" (A) or "Not Apply" (NA) (for leniency). If a firm applies for leniency, it will pay a punishment θF, where $\theta \, \epsilon [0, 1)$. The other firm will be convicted for sure, and has to pay the full penalty F. If both firms apply simultaneously, they have an equal chance of being granted leniency. In expectation, they will therefore each pay a fine of $\left(\dfrac{1+\theta}{2}\right) F$.[50] Finally, if none of the firms apply for leniency, they will pay the full fine if discovered and no fine if not. The time line for the model is depicted in figure 4. When making its choice, a firm does not know the true value of ρ. It must therefore form an expectation $E[\rho|s_i]$, which is based on its signal s_i. We assume that:

A2 $E[\rho|s_i] : [\underline{s}, \bar{s}] \to (0, 1)$ is continuously differentiable and increasing in s_i.

This assumption entails that the expectation of ρ (which lies between zero and one) gets larger when the signal s_i is increased. As the signal is an indication of how strong the CA's case is, it seems intuitive that a larger signal leads to a higher expectation of (the probability of) conviction. Note that this is the firms' expectation of ρ when they have private signals, as the expectation is formed only based on their own signal.

As a baseline comparison to the situation with incomplete information, we may consider the case with public signals. The structure and timing of the game are as

[50]This is because the expected fine is $\dfrac{1}{2} \cdot \theta F + \dfrac{1}{2} \cdot F = \dfrac{1}{2}(\theta + 1)F = \left(\dfrac{1+\theta}{2}\right) F$.

described above and depicted in figure 4. The only difference is that the signals s_i and s_j now are common knowledge. The firms' expectation of ρ is therefore based on *both* signals, i.e. we have $E[\rho|s_i, s_j]$. Their strategy in the full information case, $\psi(s_i, s_j)$, is also based on the signals of both firms. We assume that:

A3 $E[\rho|s_i, s_j] : [\underline{s}, \bar{s}] \rightarrow (0, 1)$ is continuously differentiable, responds symmetrically to s_i and s_j, and is increasing in s_i and s_j.

This is the firms' expectation of ρ when there are public signals. The difference compared to assumption 2 is that firms now can utilize the information about ρ from both firms' signals.

4.2 Equilibria of the model

We now turn to the Bayesian Nash equilibria of the game. We will start by assessing the baseline case of public signals, in order to compare it to the solution of the game with incomplete information.

4.2.1 Baseline case with public signals

As with our previous model, it is optimal for firm i to apply for leniency if firm j does.[51] Therefore, either both or none of the firms apply for leniency in equilibrium. If firm j does not apply for leniency, it is optimal for firm i to do the same if and only if the following incentive compatibility constraint (ICC) holds:

$$\theta F \geq E[\rho|s_i, s_j] F$$
$$\theta \geq E[\rho|s_i, s_j]$$

$$E[\rho|s_i, s_j] - \theta \leq 0 \tag{6}$$

The ICC in (6) states that the expected punishment from not applying is smaller than the expected punishment from applying (and receiving leniency for sure).

We therefore get the following set of Bayesian Nash equilibria of the game:

$$\psi(s_i, s_j) = \begin{cases} \text{Apply (A)} & , (s_i, s_j) \notin \Omega \\ \text{Not Apply (NA)} & , (s_i, s_j) \in \Omega \end{cases}$$

[51] By also applying, firm i will lower its expected penalty from F (for sure) to $\frac{1+\theta}{2}F$.

where

$$E[\rho|s_i, s_j] - \theta \leq 0 \, , \forall(s_i, s_j) \in \Omega$$

Ω is the set of signals that satisfy the ICC in equation (6). In words; for all combinations of signals, where the expected probability of conviction is lower than the degree of leniency (θ), the firms do not apply for leniency. For the remaining combinations of signals, both firms apply.

When $E[\rho|s_i, s_j] = \theta$, the firms do not apply for leniency. However, with a slight increase in one of the signals, both firms apply. As with our previous model, this causes a jump in expected penalties; now from $E[\rho|s_i, s_j]F = \theta F$ to $\left(\dfrac{1+\theta}{2}\right)F$. As a result, whenever $\theta < E[\rho|s_i, s_j] < \left(\dfrac{1+\theta}{2}\right)$, both firms apply for leniency. At the same time, they have *higher* expected penalties than they would if they did not apply. [52]

4.2.2 The case of incomplete information

In the situation with incomplete information, a firm bases its decision solely on its own signal. We must therefore find a cut-off value, x, for the signal. When the signal is larger than the cut-off, the probability of conviction is likely to be high. Therefore, the firm applies for leniency. If the signal is below the cut-off, the firm does not apply. We can express the firms' symmetric strategy as:

$$\phi(s_i) = \begin{cases} \text{Apply (A)} & , s_i \in (x, \bar{s}] \\ \text{Not Apply (NA)} & , s_i \in [\underline{s}, x] \end{cases}$$

Incentive compatibility

We can now find the set of x for which $\phi(s_i)$ constitutes a Bayesian Nash equilibrium.[53] As in 4.2.1, we need to assess the incentive compatibility constraint for NA, in order to determine the set of signals for which both firms choosing NA is an equilibrium:

$$\text{Expected penalty from playing NA} \leq \text{Expected penalty from playing A}$$

[52] This is because it is not incentive compatible for them to choose NA, as each of them has an incentive to deviate. By applying for leniency, the deviating firm would get leniency for sure. As $\theta F < E[\rho|s_i, s_j]F$, it could lower its expected punishment by deviating. Thus, it is not an equilibrium for the firms to choose NA.

[53] Note that as there is a set of x's, there is also a set of Bayesian Nash equilibria. Harrington (2013) briefly discusses the Pareto ranking of the equilibria, and concludes that firms prefer the equilibrium with the highest cut-off x.

If firm i does not apply for leniency, one of three contingencies may occur. Firstly, firm j may apply for leniency, and firm i has to pay the fine F. Secondly, firm j may also choose not to apply, but the firms are discovered and prosecuted by the CA. They both have to pay F. Finally, firm j may choose not to apply, the firms may not be discovered and no fines are paid. We therefore get the following expected penalty from playing NA:

$$\underbrace{[1-H(x|s_i)]}_{\text{Prob.that firm } j \text{ applies}} \cdot F + H(x|s_i) \cdot \underbrace{E[\rho|s_i, \{s_j \leq x\}]}_{\text{Exp.prob. of conviction}} \cdot F + \underbrace{H(x|s_i)}_{\text{Prob. that firm } j \text{ does not apply}} \cdot 0 \quad (7)$$

As $H(x|s_i)$ is the probability that firm j receives a signal smaller than the cut-off x, it is also the probability that firm j does not apply for leniency. Note that this probability is conditional on firm i's signal. The expected probability of conviction does not only depend on s_i, but it is also necessary that firm j does not apply for leniency. The expectation is therefore also conditional on s_j being under the cut-off.

If firm i applies for leniency, conviction is certain. However, it is not certain whether or not firm j also applies. If firm i is the only one to apply, it receives leniency and must pay θF. If both firms apply, they face an expected penalty of $\left(\frac{1+\theta}{2}\right) F$. We therefore get the following expected penalty from playing A:

$$\underbrace{H(x|s_i)}_{\text{Prob.that firm } j \text{ does not apply}} \cdot \theta F + \underbrace{[1-H(x|s_i)]}_{\text{Prob.that firm } j \text{ applies}} \cdot \left(\frac{1+\theta}{2}\right) F \quad (8)$$

We can insert these expected penalties in the incentive compatibility constraint. If the expression in (7) is less than or equal to that in (8), firm i does not apply for leniency. The **ICC for NA** is therefore:

$$[1-H(x|s_i)]F + H(x|s_i)E[\rho|s_i, \{s_j \leq x\}]F \leq H(x|s_i)\theta F + [1-H(x|s_i)]\left(\frac{1+\theta}{2}\right) F \quad (9)$$

Conversely, if the expression in (7) is strictly larger than that in (8), firm i applies for leniency. The **ICC for A** is therefore:

$$[1-H(x|s_i)]F + H(x|s_i)E[\rho|s_i, \{s_j \leq x\}]F > H(x|s_i)\theta F + [1-H(x|s_i)]\left(\frac{1+\theta}{2}\right) F \quad (10)$$

Rewriting either of the ICCs, we can define:

$$\Delta(s_i, x) \equiv E[\rho|s_i, \{s_j \leq x\}] - \theta + \left(\frac{1-\theta}{2}\right)\left[\frac{1-H(x|s_i)}{H(x|s_i)}\right]$$

Hence, if $\Delta(s_i, x) > 0$, then firm i applies for leniency. If $\Delta(s_i, x) \leq 0$, then firm i does not apply.

Bayesian Nash equilibrium

Let us now find the conditions for $\phi(s_i)$ to be a Bayesian Nash equilibrium:

1. It must be the case that firm i has no incentive to deviate (i.e. to play NA) when $s_i > x$. We must therefore have that $\Delta(s_i, x) > 0$ for all $s_i > x$. As $\Delta(s_i, x)$ is strictly increasing in the firm's signal,[54] this holds if $\Delta(x, x) \geq 0$.

2. It must be the case that firm i has no incentive to deviate (i.e. to play A) when $s_i \leq x$. We must therefore have that $\Delta(s_i, x) \leq 0$ for all $s_i \leq x$. As $\Delta(s_i, x)$ is strictly increasing in the firm's signal, this holds if $\Delta(x, x) \leq 0$.

As both of these conditions must hold for $\phi(s_i)$ to be a Bayesian Nash equilibrium, we must have $\Delta(x, x) = 0$ (such that x is an equilibrium cut-off). In other words, the ICC for not applying in (9), must hold with equality when the signal received, s_i, is the cut-off value x.

To summarize, we have the following possibilities for equilibrium cut-offs:

- If $\Delta(x, x) = 0$, then x is an equilibrium cut-off.
  When firm i receives the cut-off value x as a signal (i.e. when $s_i = x$), the expected penalty from choosing A must equal the expected penalty from choosing NA.

- $x = \underline{s}$ is an equilibrium cut-off.
  It means that firms always apply for leniency, regardless of their signal. Therefore, only condition 1 above must be satisfied, i.e. $\Delta(x, x) \geq 0$ or equivalently that (10) holds. To see this, note that $H(\underline{s}|s_i) = 0$. (10) then yields: $F + 0 > 0 + \left(\frac{1+\theta}{2}\right) F$. As $\theta < 1$, this condition always holds for $x = \underline{s}$. As

[54] $$\frac{\partial \Delta(s_i, x)}{\partial s_i} = \underbrace{\frac{\partial E[\rho|s_i, s_j \leq x]}{\partial s_i}}_{> 0 \text{ by assumption 2}} - 0 + \left(\frac{1-\theta}{2}\right)\left[\underbrace{\frac{\partial H(x|s_i)}{\partial s_i}}_{\geq 0 \text{ by assumption 1}} \frac{1}{(H(x|s_i))^2}\right] > 0$$

previously discussed, it is always optimal to apply for leniency when the other firm does. Therefore, if one firm always applies for leniency, it is optimal for the other firm to do the same.

- If $E[\rho|\bar{s}] \leq \theta$, then $x = \bar{s}$ is an equilibrium cut-off.
  If the expectation of ρ when firm i receives the highest signal possible still is smaller than θ, never applying for leniency is an equilibrium for the firms.
  First note that $H(\bar{s}|s_i) = 1$. The ICC for NA in (9) therefore yields:
  $0 + E[\rho|s_i, s_j \leq \bar{s}]F \leq \theta F + 0 \Rightarrow E[\rho|s_i] \leq \theta$
  Given that firm j never applies for leniency, it is optimal for firm i to do the same if (9) holds regardless of the signal firm i receives.

- Furthermore, if $E[\rho|\bar{s}] < \theta$ (and $\theta < E[\bar{s}), \bar{s}]$), then $\Delta(\bar{s}, \bar{s}) < 0$. Note that, compared to the previous point, this is now a *strict* inequality. As $\Delta(\underline{s}, \underline{s}) > 0$, by continuity there must therefore exist some $x = s'$, $s' \in [\underline{s}, \bar{s}]$, such that $\Delta(s', s') = 0$. Thus, there are at least three equilibrium cut-offs: $x \in \{\underline{s}, s', \bar{s}\}$.

4.3 The role of information

Let us recall the incentive compatibility constraint for not applying when firms have public signals in (6). Assuming that firm j does not apply, firm i prefers to *apply* if and only if:

$$E[\rho|s_i, s_j] - \theta > 0 \tag{11}$$

In the case of private signals, we can find the equivalent condition by rewriting the ICC for A in (10):

$$E[\rho|s_i, \{s_j \leq x\}] - \theta > -\left(\frac{1-\theta}{2}\right)\left[\frac{1 - H(x|s_i)}{H(x|s_i)}\right] \tag{12}$$

The left hand sides of the two equations can be referred to as the "Prosecution Effects" for each of the two information cases. They are affected by the probability of conviction when no firm applies, compared to how lenient the leniency program is. As conviction becomes more likely (the expectation of ρ increases), the conditions in (11) and (12) become more easily satisfied. We can interpret this as whistle blowing becoming more attractive. Note that the LHS expressions are not the same. In (11), the expectation of ρ is formed by observing both firms' signals, as they are public knowledge. In (12), on the other hand, the other firm's signal is unknown. The expectation must therefore be formed conditional on one's own signal. Additionally, the other firm's signal must be low enough for it not to apply for leniency.

The right hand sides of the expressions constitute what Harrington (2013) refers to as the "Pre-emption Effect". When there is no private information, this effect is non-existent, as seen in (11). However, when firms have private information about the probability of conviction, they do not know for sure what their opponent will do. Their choices of whistle blowing therefore depend on the likelihood that their opponent will apply before them. When it is more likely that the opponent will apply for leniency, the pre-emption effect becomes stronger.[55] Therefore the incentive for whistle blowing increases. It is noteworthy that the pre-emption effect also becomes stronger the more lenient the leniency program is (the smaller θ is). As the fine reduction becomes bigger, the more a firm has to gain by applying before its opponent.

However, as the LHSs of the two expressions cannot be easily compared, it is not clear whether or not private information leads to more or less leniency applications (and convictions). An important result of the model is that $x = \underline{s}$ is the *unique* Bayesian Nash equilibrium when leniency is sufficiently large (i.e. θ is sufficiently low) and firms have private information (Harrington, 2013, Theorem 1, p.13). If a large enough fraction of the fines is waived through the leniency program, firms always apply for leniency (regardless of their signal). Even if $E[\rho|s_i] < \theta$,[56] the incentive compatibility constraint for applying may still hold. This is because of the pre-emption effect. Even if the prosecution effect is weak, the risk that the opponent may apply for leniency first causes both firms to apply for sure. It is therefore not needed that there is full leniency (i.e. $\theta = 0$) for this equilibrium to exist and be unique.[57]

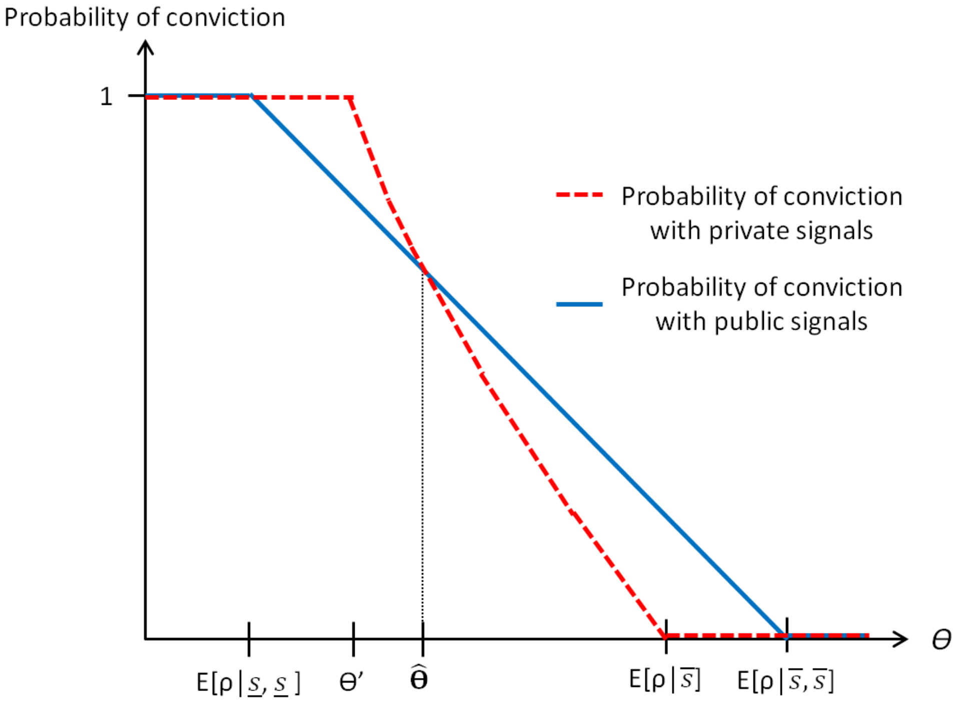
The existence of this unique equilibrium, however, does not explain how the private information case compares to one with public signals. Harrington (2013, Theorem 2)

[55] $[1 - H(x|s_i)]$ is the probability that the opponent applies for leniency. As the fraction $\frac{1 - H(x|s_i)}{H(x|s_i)}$ is increasing in $1 - H(x|s_i)$, the LHS of (12) becomes smaller (i.e. more negative). The ICC for applying therefore becomes easier to satisfy.

[56] This implies that the expected punishment from applying for leniency is actually higher than the expected punishment from not applying. Thus, the prosecution effect alone is not strong enough to ensure the firm's leniency application.

[57] In his proof, Harrington (2013) constitutes what value of θ is "sufficiently low", such that always applying is the unique equilibrium. For the ICC in (12) to hold, it is equivalent to state that $\Delta(x, x) > 0$. By setting $\Delta(x, x) = 0$ and solving for θ, this constitutes the maximum value θ', such that the ICC for A holds. More precisely, if $\theta < \theta'$, then $x = \underline{s}$ is the unique Bayesian Nash equilibrium.

Figure 5: Probability of conviction with private and public signals



shows that, if leniency is sufficiently generous,[58] then leniency usage and conviction are (weakly) higher with private signals. Moreover, if leniency at the same time is not full (and $E[\rho|\ \underline{s}\ ,\ \underline{s}] < \theta$), then the probability of conviction is strictly higher with private signals. This is because there will be contingencies with public signal, where the firms choose not to apply. When signals are private, however, firms still apply in every contingency (and the probability of conviction is equal to one).

Finally, Harrington (2013, Theorem 3) shows that, when leniency is weak, the probability of conviction and leniency use is higher with public than with private signals.[59] To summarize, we know that the probability of conviction with private signals is *higher* when θ is sufficiently low, and *lower* when θ is sufficiently high compared to public signals. By assessing the two ICCs in (11) and (12), we can see that a change in θ affects the prosecution effect equally in both equations. However, θ additionally affects the pre-emption effect in (12). Therefore, I argue that a change in the degree of leniency (θ) has a stronger effect on the probability of conviction

[58]By sufficiently generous, I mean that the condition in theorem 1 (Harrington, 2013) is satisfied. Therefore, firms always apply for leniency when signals are private.

[59]Specifically, he shows that this is the case when $E[\rho|\bar{s}] < \theta < E[\rho|\bar{s}, \bar{s}]$. The condition implies that whistle-blowing never occurs when there are private signals, as $E[\rho|\bar{s}] < \theta$. However, when signals are public, there are contingencies when firms apply for leniency. This occurs when $\theta < E[\rho|s_i, s_j]$.

when firms have private information.[60] Finally, this implies that there will typically be a unique threshold, $\hat{\theta}$, such that for all $\theta \leq \hat{\theta}$, the probability of conviction is (weakly) higher when firms have private signals. I have illustrated this threshold in figure 5. We may therefore conclude that, if leniency is sufficiently large ($\theta < \hat{\theta}$), then the probability of conviction is higher when firms have private information. We can also get the more realistic result that not all firms apply for leniency simultaneously, because of private information. At the same time, the private information creates an incentive to pre-empt the opponents application, thus raising the probability of whistle-blowing and conviction.

As long as private actors are able to raise lawsuits against former cartel members, it can be argued that there is no full leniency (Harrington, 2013). This is because private claims can be very costly for firms and need to be incorporated in expected punishments. According to the model, full leniency would lead to all firms applying simultaneously (despite of incomplete information). The possibility of private lawsuits can therefore be part of the explanation for why not all firms simultaneously rush to apply for leniency. We have established that, if leniency is sufficiently large, private information between firms leads to more leniency applications than the models in chapters 2 and 3 suggest. Incorporating the information problem presented in this section into leniency models seems to be a natural next step. Finally, this model of whistle-blowing also suggests that it is in the CA's best interest to keep firms' information incomplete. Assuming that leniency is sufficiently generous, a CA can use the pre-emption effect to amplify the incentive for whistle-blowing. The uncertainty concerning the opponent's choice will generate more leniency applications and thus more convictions.

[60]By moving all terms to the LHS of (12), we can assess the effect of a change in θ on the ICC: $\frac{\partial ICC}{\partial \theta} = -1 - \frac{1 - H(x|s_i)}{H(x|s_i)} < -1$. Even though the threshold x may change as θ increases, the probability H will still lie between 0 and 1. As a result, the condition $\frac{\partial ICC}{\partial \theta} < -1$ holds.

5 Conclusion

This thesis has discussed corporate leniency programs and their impact on collusion. On one hand, waiving the fines of a (former) cartel member reduces the expected punishment from colluding. This may induce more illegal behaviour. On the other hand, the leniency program introduces an incentive to cheat on the collusive arrangement. Thereby, it can cause cartels to break down. Additionally, if only the first firm to apply receives amnesty, firms may "race to the courthouse" in order to apply for leniency first. As only one firm is granted leniency, this raises the expected punishment from collusion, making it less profitable. The main conclusion of this model (and the main body of work on leniency programs) is that leniency programs reduce collusion.

The leniency model presented here finds that it is optimal to offer full leniency, i.e. to waive all fines. However, this amnesty should only be offered to the first firm to come forward. A defense could be made to offer partial leniency to subsequent applicants, if the information presented by the first firm is not sufficient to ensure a conviction. It is a simplifying assumption of the model that conviction is guaranteed with only one informant. This may however be unrealistic, and there is a need to explore the benefits and effects of granting partial leniency to additional firms. It is also noteworthy that additional information can lead to less resources being spent on prosecuting leniency cases. Therefore, more resources would be freed in order to expose undetected cartels. Regardless, I cannot help but wonder whether a leniency program where all cartel members can receive leniency (as in the SAS case in the EU) does more damage than it does good. Removing the possibility of multiple leniency recipients can alleviate this problem. I therefore assert that the EU and Norway ought to grant leniency to the first firm *only*.

As the goal of the CA in the model presented here is to destabilize cartels most efficiently, it also implies that amnesty should be granted even if the CA already has a fairly strong case. The more firms are involved in a cartel, the less strict the CA should become in accepting the application. When there are more cartel members, the incentive to race to the courthouse becomes stronger. By allowing leniency applications, despite already having a strong case, the CA can take advantage of this incentive.

The model I present focuses on the optimal degree of leniency. The fraction of fines waived is therefore modeled as the CA's only policy choice. In reality, however, the CA can also influence the magnitude of fines, investigation efforts and prosecution efforts. More importantly, CAs are generally also resource constrained. They must therefore prioritize between prosecution and investigation efforts. If a leniency program produces a large number of leniency applications, the prosecution of these cases may crowd out investigation efforts. Instead of hunting active cartels, the CA can end up spending a large share of their resources on prosecuting already broken-down cartels. Hence, a seemingly successful leniency program may in reality raise the overall cartel rate.[61] In order to prevent this, sufficiently large penalties are needed (Harrington and Chang, 2014). The question of whether or not to allow for multiple leniency recipients also pertains to the resource discussion. If the information presented by additional firms reduces the cost of prosecution, it might be beneficial to allow more firms to receive (partial) leniency. A final point to be made in the resource discussion is that, while leniency programs require resources, they also generate them. If leniency cases result in a larger number of convictions, a larger amount of fines may also be collected. This can dampen the crowding out of resources.

Models of leniency programs generally result in symmetric equilibria, where either all or none of the firms apply for leniency. However, all members of a cartel simultaneously racing to the courthouse rarely happens in reality. This thesis therefore assesses a simple model of the whistle-blowing choice, where firms have private information about the strength of the CA's case. We find that the probability of conviction is higher when firms have private information, given that the leniency program is sufficiently generous. Firms still have an incentive to avoid prosecution, but additionally want to pre-empt their opponent's leniency application. It is therefore desirable for CAs to maintain the information problem between firms. A natural next step would be to incorporate private information into the general leniency model. This would yield a more realistic model of the firms' choices, and would better allow us to assess the effect on cartel formation.

Given the differences in leniency policies between the US and the EU/Norway, there clearly is a need to explicitly model the costs and benefits of allowing multiple leniency recipients. However, if opening up for several firms to receive leniency results in the practice of *all* cartel members receiving some degree of leniency, the program can hardly be anything but counter-productive. Finally, the long-term effects

[61]The model used to analyze resource constraints in the leniency context, uses a life and death cycle for the cartels. It allows for re-cartelization of industries, whenever it is incentive compatible.

of leniency programs are rarely mentioned. Receiving leniency, while all competitors are sentenced with large fines, could give the firm a competitive advantage. In the long run, this could lead to other problems related to market power and forcing opponents to exit the industry. Such long-term strategic incentives as well as their social costs seems to be a neglected topic in the discussion of leniency models.

References

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

Bloomberg, 2003. Odfjell's CEO to Be Jailed, Pay Fine for Price Fixing (Update5). Published 29.09.2003. Retrieved 29.03.2015.
URL http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aQH5gAn9U.Dc&refer=europe#share

Chen, Z., Rey, P., 2013. On the design of leniency programs. Journal of Law and Economics 56 (4), 917–957.

Dagens Næringsliv, 2014. Krever erstatning fra SAS etter ulovlig prissamarbeid. Published 22.08.2014. Retrieved 29.03.2015.
URL http://www.dn.no/nyheter/naringsliv/2014/08/22/1412/Luftfart/krever-erstatning-fra-sas-etter-ulovlig-prissamarbeid

European Commission, 2010. Press Release. Antitrust: Commission fines 11 air cargo carriers €799 million in price fixing cartel. Published 09.11.2010. Retrieved 29.03.2015.
URL http://europa.eu/rapid/press-release_IP-10-1487_en.htm?locale=en

European Competition Network, 2012. ECN Model Leniency Programme. 2012 revision. Retrieved 17.03.2015.
URL http://ec.europa.eu/competition/ecn/mlp_revised_2012_en.pdf

Foundation for Economic Education, 1997. Breaking Up Antitrust. Published 01.01.1997. Retrieved 28.03.2015.
URL http://fee.org/freeman/detail/breaking-up-antitrust

GOV.UK, 2014. Countries in the EU and EEA. Published 12.11.2014. Retrieved 01.05.2015.
URL https://www.gov.uk/eu-eea

Harrington, J. E., 2008. Optimal Corporate Leniency Programs*. The Journal of Industrial Economics 56 (2), 215–246.

Harrington, J. E., 2013. Corporate leniency programs when firms have private information: The push of prosecution and the pull of pre-emption. The Journal of Industrial Economics 61 (1), 1–27.

Harrington, J. E., Chang, M.-H., 2014. When Can We Expect a Corporate Leniency Program to Result in Fewer Cartels? Available at SSRN 2530545. Forthcoming in Journal of Law and Economics.

Klemperer, P., 1999. Auction theory: A guide to the literature. Journal of economic surveys 13 (3), 227–286.

Konkurranseloven, 2004. Lov om konkurranse mellom foretak og kontroll med foretakssammenslutninger (konkurranseloven). Retrieved 16.03.2015.
URL `https://lovdata.no/dokument/NL/lov/2004-03-05-12`

Konkurransetilsynet, 2014. Anker asfaltsaken. Published 20.03.2014. Retrieved 28.03.2015.
URL
`http://www.konkurransetilsynet.no/no/aktuelt/nyheter/anker-asfaltsaken/?hl=asfalt`

Ludwig Maximilian University of Munich, 2014. Jean Tirole wins 2014 Nobel Prize in Economics. Published 14.10.2003. Retrieved 19.02.2015.
URL `http://www.en.isto.bwl.uni-muenchen.de/news/nobel_tirole/index.html`

Meyer, C. B., 2011. Vennskap og kartell. Published 14.06.2011. Retrieved 29.03.2015.
URL `http://www.konkurransetilsynet.no/no/Aktuelt/Artikler-og-innlegg/Vennskap-og-kartell/`

Motta, M., Polo, M., 2003. Leniency Programs and Cartel Prosecution. International journal of industrial organization 21 (3), 347–379.

NRK, 2011. Slik ble staten svindlet for millioner. Published 25.01.2014. Retrieved 28.03.2015.
URL `http://www.nrk.no/norge/innrommer-prissamarbeid-pa-asfalt-1.7476170`

Spagnolo, G., 2004. Divide et impera: Optimal deterrence mechanisms against cartels and organized crime.

Stolt-Nielsen, E., 2011. Når tale blir gull og taushet koster. Advokatbladet (3), 42–44.

Tirole, J., 1988. The theory of industrial organization. MIT press.

US Department of Justice, 2008. Frequently Asked Questions Regarding the Antitrust Division's Leniency Program and Model Leniency Letters. Retrieved 16.03.2015. URL http://www.justice.gov/atr/public/criminal/239583.htm