Trade Associations: Why Not Cartels?☆

David K. Levine[1], Andrea Mattozzi[2], Salvatore Modica[3]

Abstract

The relevance of special interests lobbying in modern democracies can hardly be questioned. But if large trade associations can overcome the free riding problem in order to form effective lobbies, why do they not also threaten market competition by forming equally effective cartels? We argue that the key to understanding the difference lies in supply elasticity. The group discipline which works in the case of lobbying can be effective in sustaining a cartel only if increasing output is sufficiently costly - otherwise the incentive to deviate is too great. The theory helps organizing a number of stylized facts within a common framework.

Keywords: cartels, labor unions, lobbying, monitoring costs, self organizing groups, special interests.

☆First Version: February 17, 2018. We would like to thank Giacomo Calzolari, Rohan Dutta, Daniel Garcia and seminar participants at the Pamplona Jornadas de Economia Industrial, the Bocconi Theory Meeting in Capri, the SAET meeting in Faro, the Fudan University Theory Workshop, the SHUFE Theory Workshop, the SJTU Theory Workshop, the LAMES meetings in Buenos Aires, and the University of the Pacific in Lima. We gratefully acknowledge support from the EUI Research Council.

Email addresses: `david@dklevine.com` (David K. Levine), `andrea.mattozzi@eui.eu` (Andrea Mattozzi), `salvatore.modica@unipa.it` (Salvatore Modica)

[1]Department of Economics and RSCAS European University Institute and Department of Economics WUSTL

[2]Department of Economics European University Institute

[3]Università di Palermo, Italy, Department of Economics, Business and Statistics (SEAS)

1. Introduction

The objective of this paper is to contend with three stylized facts: (1) large lobbying organizations exist and are effective; (2) large commercial cartels are less common and appear difficult to sustain; and (3) nevertheless trade unions that both lobby and cartelize do form and are effective. We reconcile this apparent puzzle with a simple common theory.

Our interest is in large organizations that face a free rider problem. In standard industrial organization models of collusion through repetition, cooperation can be sustained using simple trigger strategies. Furthermore, as shown by Pecorino (1998) in a repeated tariff lobbying game, the critical value of the discount factor above which cooperation can be sustained does not necessarily rise with the number of firms. This argument, however, hinges on assuming perfect information: a threat to disband a cartel when its rules are not followed cannot be effective with many firms in the presence of small noise.[4] We provide an alternative model in which organizations use peer punishments to overcome free riding. In both types of models the possibility of successful collusion depends on the incentives to deviate. The greater this incentive is, the greater the punishments needed to induce compliance and with imperfect monitoring greater punishments are more costly.

Our contention is that since the incentive to deviate in lobbying organizations is naturally limited to the gain from failing to contribute, collusion is relatively easy. In a market setting, however, large cartels are more difficult to sustain because deviation is potentially far more profitable if production can be ramped up on a large scale to take advantage of a gap between price and marginal cost. The ability to take advantage of this opportunity crucially depends on the elasticity of supply. If marginal cost rises rapidly with output then the gain from deviating is limited and collusion can be sustained. This - we argue - is the case for trade unions but not the case for manufacturing firms or farms.

[4]See Green (1982), Sabourian (1990), Levine and Pesendorfer (1995), Fudenberg, Levine and Pesendorfer (1998), Al Najjar and Smorodinsky (2001), and Pai, Roth and Ullman (2014).

We emphasize that in the case of small cartels the situation is different. Small cartels are similar to lobbying organizations in that there is a natural restriction on increasing output: individual demand is downward sloping and firms are not so eager to produce large amounts. Hence small cartels are easier to sustain. It should be clear that we do not reject the traditional "small cartel" theory in which collusive arrangements are enforced by threat of future price retaliation. As we have indicated, it is for large cartels that these types of punishments are not useful. Our key point is that small cartels - those traditionally studied in the industrial organization literature - are different than large cartels. In particular we argue that supply elasticity plays a much less significant role in the case of small cartels than it does for large cartels.

Going back to the stylized facts mentioned above: that large lobbying organizations exist and are effective is the core of much of the political economy of farm subsidies and trade restrictions. Certainly we observe that special business interests such as farmers or the chamber of commerce for example are effective at lobbying government for subsidies and for entry and trade restrictions. These organizations, which we refer to as *trade associations*, are small as a share of the economy but often quite large in absolute size. Because of their large absolute size they face a substantial free rider problem in raising resources for lobbying: this is well documented by Olson (1965) and his successors. Still, they are able to overcome this free rider problem to be effective at lobbying.[5] In the case of farming, for example, agriculture represents slightly more than 1% of U.S. GDP but there are more than 2 million farms, and they command around 0.5% of GDP in subsidies. In Japan the GDP share is similar, there are over 3 million farms and subsidies exceed 1% of GDP.[6] Notice, however, that while firms benefit from lobbying, they would also benefit from collusion in the

[5] See, for example, Grossman and Helpman (2001).

[6] The share of agriculture in value added is from `http://data.worldbank.org/indicator/NV.AGR.TOTL.ZS`. Total agricultural support as a percent of GDP is from
`http://stats.oecd.org/viewhtml.aspx?QueryId=70971&vh=0000&vf=0&l&il=&lang=en`. Data on the number of farms is from Lowder, Skoet and Raney (2016). In general, farmers are very effective at overcoming the free rider problem to lobby for farm subsidies, see for example Acemoglu and Robinson (2001).

form of an output-restricting cartel. The free riding problem appears similar: produce more and reap extra profits in the cartel case, do not contribute to the lobbying effort in the lobbying case.

This raises the puzzle we want to address: if trade associations are so effective at overcoming the free rider problem in order to lobby, why are they not equally effective at overcoming the free rider problem of sustaining a cartel? But also: how come trade unions, which also are output-restricting cartels, do typically exist? That large cartels are uncommon and difficult to form is well understood in the industrial organization literature where focus is on concentration ratios and industries with few firms - but this does not explain either the ubiquity of lobbying organizations nor the success of trade unions.

In order to understand when trade associations are successful at lobbying and at cartelization we need a theory of how they overcome free rider problems. We know from the work of Ostrom (1990) and her successors how this can be achieved: groups can self-organize to overcome the free rider problem and provide public goods through peer monitoring and social punishments such as ostracism. Formal theories of this type originate in the work of Kandori (1992) on repeated games with many players and have been specialized to the study of organizations by Levine and Modica (2016) and Dutta, Levine and Modica (2018). The basic idea is that groups choose norms consisting of a target behavior for the group members and individual penalties for failing to meet the target; these norms are endogenously chosen in order to advance group interests. Specifically the group designs a mechanism to promote group interests subject to incentive constraints for individual group members, and it provides incentives in the form of punishments for group members who fail to adhere to the norm. Here we build on this theory to compare the public goods problem of lobbying to that of cartelization.

Our theory helps organizing a number of stylized facts. First: we observe trade associations that lobby but do not cartelize, but rarely ones that cartelize but do not lobby. This is because the greater incentive to deviate makes cartelization less attractive than lobbying. Some trade associations both lobby and cartelize - most notably trade unions. In this case individual members are tightly constrained in how much they can increase the number of hours they work; that is, in this case marginal cost is inelastic for individual workers - more so than in typical production settings. The theory says that elastic marginal cost works against cartelization, while inelastic marginal cost works in favor of cartelization - that is, as it is indeed the case, we should not see diffuse large production cartels, but we should see trade unions cartelize.

The rest of the paper is organized as follows. In the next section we present the model. In Sections 3 and 4 we consider the cases of lobbying and large cartels, respectively. In Section 5 we argue that small cartels are different and in fact similar to lobbying organizations. Section 6 deals with the efficiency of punishment and fines. We discuss a number of alternative explanations to our motivating puzzle, which appear not to be particularly compelling, in Section 7. Section 8 concludes.

2. The Model

We study a trade association made up of a continuum of members with unit mass. Members produce output x, which given a unit mass of members is both per capita and aggregate. The gross social value of the output is $Vh(x)$ where $V > 0$ and $h(\cdot)$ is a smooth function described below.

We are interested in comparing lobbying activity and production activity. In the lobbying case, the social value is a public good, the group expends effort x in lobbying and $Vh(x)$ represents subsidies, or favorable laws, obtained through lobbying. The case of production applies to a market where a group of firms hold a monopoly over a good to be sold to competitive buyers. In this case $Vh(x)$ is value to the consumers and the marginal value $Vh'(x)$ is the demand price $p(x)$. Letting $r(x) = h'(x) + xh''(x)$, the marginal revenue is $Vr(x)$. We assume that social value is increasing and concave in x, and that marginal revenue is decreasing. Specifically, on the demand side we assume the following:

Assumption. *There is a satiation level $\overline{x} > 0$ such that $h(x) = \overline{x}$ for all $x \geq \overline{x}$. For $x \in [0, \overline{x}]$ the function $h(x)$ is smooth with $h'(x) \geq 0, h''(x) < 0$ with the former inequality strict for $x < \overline{x}$. We also assume $r'(x) < 0$. Finally*

$Vh'(0) > 1$.[7]

Notice that in both the lobbying case and in the production case V, which we refer to as the *value,* represents a measure of the utility to an individual from consuming the good.[8]

Output - or lobbying effort - is produced at constant marginal cost normalized to 1 up to a basic capacity constraint also normalized to 1. Production greater than 1 is feasible but has a greater marginal cost. For simplicity we assume that above basic capacity marginal cost increases linearly so that for $x > 1$ marginal cost is equal to $1 + \sigma(x - 1)$, where σ denotes the reciprocal of supply elasticity. The cost function for each member is therefore $C(x) = x + (\sigma/2)(\max\{0, x - 1\})^2$. In other words cost is piecewise linear with marginal costs becoming elastic only above a production level of one. This is a generalization of the idea of constant marginal cost with a capacity constraint: after the basic capacity is reached costs may rise rapidly, but production is still possible. This simplification enables us to vary a single parameter σ controlling the extent to which capacity constraints bind and get clean comparative static results.

In our setting the group chooses a *norm* for its members, which consists of a target level of output $\xi \geq 0$, and an individual punishment $P \geq 0$ for deviators. Specifically, each group member chooses an output level x and, while output cannot be perfectly monitored, all group members receive a public noisy binary signal of whether $x = \xi$ is generated, that is, whether the norm was adhered to or not.[9] The signal is either "good, followed the norm" or "bad, violated the norm." If the norm is followed, that is, $x = \xi$, then a bad signal is generated with probability $\pi > 0$. If the norm was violated, that is $x \neq \xi$, the bad signal

[7]Other than assuming concavity, which is relatively standard, we require that $\overline{x}$ is finite albeit can be very large. Furthermore, we assume that marginal revenue is decreasing for uniqueness of a solution to the monopoly problem.

[8]In particular increasing the size of the market in the production case is not equivalent to increasing V. In both cases an increased quality of the good or a higher price of substitutes would increase V.

[9]We show in Section 6 that our results are robust to relaxing the assumption that all group members observe the signal.

is generated with a higher probability $\Pi > \pi > 0$. The ratio $\pi/(\Pi - \pi) \equiv \theta > 0$ is the *monitoring difficulty*.[10] When the bad signal is generated the individual is sanctioned by the group and suffers a utility loss of P. For simplicity we assume that this punishment has no costs or benefits to group members other than the one punished.[11] The norm is *incentive compatible* if all members find it individually optimal to follow it given that the others are doing so. The group is assumed to design an incentive compatible mechanism (ξ, P) that maximizes the common utility of the members.

As a benchmark we define the *social optimum* χ as the norm that maximizes social value minus aggregate cost, that is $Vh(x) - C(x)$. This objective function is continuous and concave and, given our previous assumption ($Vh'(0) > 1$ and $Vh'(\overline{x}) = 0$), the maximum is given by the unique solution to

$$Vh'(\chi) = C'(\chi) = 1 + \sigma \max\{0, \chi - 1\}.$$

In the case of production this means price equal marginal cost - it is the competitive equilibrium. In the cartel case the social optimum/competitive equilibrium is the worst possible outcome for the cartel yielding as it does zero profits. The lobbying case is the reverse: the social optimum is the best possible outcome for the cartel optimally balancing the costs and benefits of the public good.

3. Lobbying

With lobbying the social value of output is a public good for the group, so each group member receives $Vh(x)$ where x is aggregate output. If a member follows the norm by contributing ξ she receives a utility $Vh(\xi) - C(\xi) - \pi P$. If she

[10]Note that θ can be very small even when the detection rate Π is very small, as long as π is even smaller. However, as it will be clear in the next section, incentive compatibility requires that when Π is very small the punishments P must be very large. If there is a limit on the worst possible punishments available then Π is naturally bounded below.

[11]In practice punishments may involve exclusion from group benefits or fines and this may impact group members other than the one punished. In Section 6 we discuss the nature of punishment and show that our results are robust to this simplification. Note that in this model deviations are always punished: this is unlike Bernheim and Madsen (2017) in whose model it may be advantageous not to punish some deviations.

deviates from the norm the best deviation is to produce 0 and not contribute to the public good at all, resulting in utility $Vh(\xi) - \Pi P$. Deviating is not optimal if and only if $C(\xi) - (\Pi - \pi)P \leq 0$. Group utility is decreasing in P so for the optimal mechanism, the least incentive compatible punishment $P = C(\xi)/(\Pi - \pi)$ should be used. We can then write the utility of a group member, which is the group objective function, as

$$U(\xi) = Vh(\xi) - (1 + \theta)C(\xi).$$

Observe first that imperfect monitoring - $\theta > 0$ - prevents the group from reaching full efficiency: since $U'(\chi) = -\theta C(\chi) < 0$ the optimal norm $\hat{\xi}$ will always be strictly smaller than χ.

Our first result characterizes the optimal norm $\hat{\xi}$ in the case of lobbying.

Theorem 1. *For $\overline{\theta} = Vh'(0) - 1 > 0$ the optimal norm $\hat{\xi}$ satisfies*

1. For $\theta < \overline{\theta}$ we have $\hat{\xi} > 0$ with $\lim_{\theta \to 0} \hat{\xi} = \chi$ (success),

2. For $\theta \geq \overline{\theta}$ we have $\hat{\xi} = 0$ (failure),

3. $\lim_{V \to \infty} \hat{\xi} = \chi$ (success),

4. For $\theta < \overline{\theta}$ there is a bound $\underline{\xi} > 0$ independent of σ such that $\hat{\xi} > \underline{\xi}$ (no failure).

In the Appendix we prove precise bounds which imply the results stated above. These results are not intended to be surprising but to serve as a benchmark for the case of cartels. The first two results indicate that if monitoring difficult θ is low then the public good is produced, and the first best is approached (success) as θ goes to zero, while if monitoring difficulty is too high then the public good is not produced (failure). The third result shows that as the value V of the public good increases, so the incentive to produce it increases, the first best is approached. The second and fourth result show that the inverse supply elasticity σ has limited relevance. If monitoring cost is high (2) shows that no public good is produced regardless of σ , while (4) shows that if monitoring cost is low public good output is bounded away from zero by a bound that is independent of σ. In the case of cartels we shall see that σ plays a more central role.

4. Cartels

We now study the trade association holding a monopoly over a good to be sold to competitive buyers. In this market context $Vh(x) - C(x)$ is usually called "total surplus". The marginal value is the demand price of the good: $p(x) = Vh'(x)$, and the social optimum, characterized by $p(\chi) = C'(\chi)$, is the competitive equilibrium. $Vh(x) - p(x)x$ is the consumers' surplus, and firms are interested in the producer surplus $W(x) = p(x)x - C(x)$. If the cartel does not form equilibrium is competitive. On the other hand no $\xi > \chi$ would be enforced by the firms since group members would be worse off than at χ. Therefore we restrict attention to norms $\xi \leq \chi$. Observe that in this range $p(\xi) \geq 1$.

If industry norm is ξ the profits of a group member who adheres to the norm is $W(\xi) - \pi P$. What is the best thing to do if violating the norm? The key point is that the answer is *not* to produce χ but it is to produce *more* than that. Indeed, since there are a continuum of members each member is a price taker. Hence, given the price $p(\xi) \geq 1$, the profit maximizing output is the highest for which marginal cost does not exceed that price. Denoting this by $\hat{x} = \hat{x}(\xi)$ we then have $\hat{x} \geq \max\{1, \chi\}$ characterized by the equality $p(\xi) = 1 + \sigma(\hat{x} - 1)$; thus $\hat{x}(\xi) = 1 + (p(\xi) - 1)/\sigma$. The profit from this plan is $p(\xi)\hat{x} - C(\hat{x}) - \Pi P$, where $p(\xi)\hat{x} - C(\hat{x}) \geq W(\xi)$ with equality only for $\xi = \chi$.[12]

Equating payoffs from adhering and violating the norm gives the least incentive compatible punishment P, given by $(\Pi - \pi)P = p(\xi)\hat{x} - C(\hat{x}) - W(\xi)$ so that $\pi P = \theta\,[p(\xi)\hat{x} - C(\hat{x}) - W(\xi)]$. Notice that $P \geq 0$ with equality only for $\xi = \chi$; also, for $\xi < \chi$ the incentive compatible P is higher the lower ξ is. Most importantly, P increases with σ (through $\hat{x}$). The expected utility from norm ξ is

$$U(\xi) = W(\xi) - \pi P = W(\xi) - \theta\left(p(\xi)\hat{x} - C(\hat{x}) - W(\xi)\right).$$

[12]The proof of this is immediate. If $\xi < \chi$ then $p(\xi) > C'(\chi) \geq C'(\xi)$ so $p(\xi)\hat{x} - C(\hat{x}) > p(\xi)\xi - C(\xi) = W(\xi)$. If $\xi = \chi$ the deviation is to $\hat{x} = \max\{\chi, 1\}$; so if $\chi \geq 1$ the equality is immediate. If $\chi < 1$ then $p(\chi) = 1 = C'(x)$ for all $\chi \leq x \leq 1 = \hat{x}$ so $p(\chi)\hat{x} - C(\hat{x}) = p(\chi)\chi - C(\chi)$.

We show in the Appendix that the optimal norm $\hat{\xi}$ lies between the monopoly output, denoted by μ, and the competitive outcome: $\mu < \hat{\xi} \leq \chi$.[13] Closeness of $\hat{\xi}$ to μ measures success of the cartel; closeness to χ makes the cartel ineffective. The strict inequality $\mu < \hat{\xi}$ holds since monitoring costs prevent the cartel ever to reduce output down to monopoly level. Indeed, at $\xi = \mu$ we have $U'(\xi) = W'(\xi) - \pi P'(\xi) = -\pi P'(\xi) > 0$, that is the marginal loss in profit from increasing output above monopoly level is zero while the marginal reduction in monitoring cost is strictly positive. The main properties of the optimal norm are stated in the following result, proved in Appendix.[14]

Theorem 2. [Main Theorem] *The optimal norm $\hat{\xi}$ satisfies*

1. $\lim_{\theta \to 0} \hat{\xi} = \mu$ *(success),*
2. $\lim_{\theta \to \infty} \hat{\xi} = \chi$ *(failure),*
3. $\lim_{V \to \infty} \hat{\xi} = \chi$ *(failure),*
4. $\lim_{\sigma \to 0} \hat{\xi} = \chi$ *(failure).*

The main insight here is that while monitoring difficulty goes as in the lobbying case (small θ favors the cartel, large θ hinders it), the crucial difference between the cartel and public good case is the central role σ plays. If the marginal cost of exceeding basic capacity rises slowly the optimal norm is close to competition - the cartel is not effective. The reason is that the temptation to cheat on the cartel is too great: in the face of a price above marginal cost it is cheap to increase output and reap a large profit. The cost of providing incentives not to take advantage of this is high: large and costly punishments must be used. The trade association does not find it in its best interest to do this. The contrast with the lobbying case is significant: Theorem 1 shows that when the value is large relative to monitoring difficulty, then lobbying activity - that is active collusion among group members - does not depend on the elasticity of supply.

---

[13]Since we have assumed decreasing marginal revenue and weakly increasing marginal cost, μ is uniquely characterized by $W'(\mu) = 0$. Moreover $\mu < \chi$, since concavity of h implies that marginal revenue is lower than price and hence, at χ, it is also lower than marginal cost.

[14]In the Appendix we also show that when $\chi \geq 1$ an analogous result holds for payoffs (see point 3 in Theorem 5).

The result that with low σ the optimal norm is close to competition holds substance if it is also the case that competition is not close to monopoly. In the appendix we show that this is indeed the case by establishing a lower bound on $\chi - \mu$ which is larger the lower is σ (Theorem 6).

The other interesting point is that the value V works in the opposite way than in the lobbying case. A large value pushes the optimal norm towards χ in both cases. In the public good case, this is desirable for the trade association. In the case of a cartel, however, it means failure. Why does not large value lead to more collusion? The answer is that on the one hand a reduction in output raises price and hence industry profits by order V. On the other hand it raises the per unit incentive to deviate by the same amount but also raises the amount a firm wants to deviate by order V, meaning that the incentive to deviate goes up by roughly V^2. Hence as value increases the incentive to deviate goes up more than profits and so the trade association optimally restricts output less. In short, if output elasticity is high (small σ) it is difficult for a trade association to self-organize, that is, to set up a cartel. For this to be possible one has to have low θ (as in lobbying), moderate V (unlike in lobbying), and high σ (no analogue in lobbying).

A last important remark is in order. We have not dealt here with the possibility of entry. Notice that the considerations with entry are similar to those discussed here: in the case of lobbying a modest level of entrants who do not belong to the organization will have a modest effect on its effectiveness. In the case of cartelization the effect of a modest level of entrants who do not belong to the organization will depend on how capacity constrained the entrants are: that is how large is the elasticity of supply. If σ is large then the entrants will not have a big impact on the price set by the cartel, otherwise they will.

5. Small Cartels Are Different

The theory we have presented is a theory of large cartels: cartels where individual firms are sufficiently small that their individual output has no impact on prices. Existing theory - and practice - has focused on small cartels. Here we present a brief overview of that theory, not to prove new results but to give a clear idea how it differs from the theory of large cartels. In the present context small cartels are cartels with a fixed finite number N of members so that if x is the output of an individual firm and the other firms adhere to the quota ξ then per firm output is $\phi = [(N-1)\xi + x]/N$. There are two dimensions in which the theory is different, which we discuss next.

Information and Collective Punishment

The first dimension is that information can be used to enforce collective punishments when cartels are small.

If per firm output is observable then in small cartels it is possible to use a collective punishment to punish all firms equally upon deviation. For example, in a dynamic setting this might take the form of a price war as often studied in the literature. With sufficiently large punishment, corresponding to sufficient patience on the part of the firms in the dynamic case, all deviations can be made unprofitable. Hence no firm will wish to deviate, punishment never occurs on the equilibrium path so has no social cost, and any norm is incentive compatible. In particular the cartel will agree on the monopoly solution and the firms will split the monopoly profit.

If the actual output of a firm is a noisy function of intended output, it can be shown that if the noise is sufficiently small (see, for example Green and Porter (1984)) nearly the monopoly profit will be split between the firms. However, with noise this method does not scale to large cartels. As the number of firms increases individual deviations being small can no longer easily be distinguished from the background noise and as a result individual firms have little effect on the probability of punishment and so have every incentive to deviate. Even if the individual firm outputs are observed this is no help.[15] It is for this reason that our analysis of large cartels focuses on the combination of individual signals and individual punishments.

If output is not directly observable, as in our model, the key difference between individual and collective punishment is that in the former case monitor-

[15]See Levine and Pesendorfer (1995) and Fudenberg, Levine and Pesendorfer (1998) for a proof of this.

ing difficulty is clearly independent from the number of firms in the cartel. To the contrary, for collective punishment monitoring difficulty grows unboundedly with the number of firms. Thus we should expect large cartels to face greater difficulty than small cartels because they lose access to effective collective punishments.[16]

Demand and Cost

The second dimension in which our theory differs from the theory of small cartels rests in the fact that in small cartels firms face downward sloping individual demand curves: while with large cartels the only restraint on deviation lies in increasing cost, with small cartels firms limit their deviation because their increased output lowers price. For this reason, cost conditions - σ - plays a much less important role.[17]

The next result, proven in the Appendix, shows that in the case of small cartels regardless of σ - even if it is zero - downward sloping demand implies a limit to how much output firms want to produce when they deviate and hence a limit to how much they can gain by deviating.[18]

[16]Indeed, consider a fixed finite number of firms N and recall that if $x = \xi$ then punishment occurs with probability $\pi > 0$, while if $x \neq \xi$ it occurs with probability $\Pi > \pi > 0$. In the case of individual punishment we define as before $\theta = \pi/(\Pi - \pi)$. In the case of collective punishment, there are different schemes m that may be used for determining how the triggering of the collective punishment depends on the number of bad signals. Let $\pi(m)$ be the probability that punishment occurs when no firm deviates and $\Pi(m)$ the probability of punishment when one firm deviates. We define $\theta(m) = N\pi(m)/(\Pi(m) - \pi(m))$ and $\theta = \min_m \theta(m)$. A key point is that from the bounds given in Levine and Pesendorfer (1995) in the case of collective punishment $\theta \to \infty$ as $N \to \infty$.

[17]That small cartels have a relative advantage over large ones does not imply that bargaining problems are absent. Marshall, Marx and Samkharadze (2019) for example investigate a sample of 22 cartels whose average number of members is less than a dozen and for 9 of them find "evidence of frequent bargaining problems or of ongoing issues with deviations throughout the cartel period." Similar evidence is contained in the study by Genesove and Mullin (2001) of the Sugar Institute, composed of 14 sugar refining firms, which operated legally in the U.S. from 1927 to 1936.

[18]This idea is reminiscent of dynamic oligopoly models with capacity constraints. In Benoit and Krishna (1987) two oligopolistic firms first choose their scales of operation and then compete on prices. If firms cannot adjust the initial capacity choice, they tend to carry excess capacity in equilibrium, which precludes achieving monopoly behavior. The reason is that both firms invest in excess capacity only to use it as threat in case of deviation from collusion. While on the one hand excess capacity makes implicit collusion sustainable, on the other hand

Theorem 3. *In small cartels the optimal norm $\hat{\xi}$ satisfies $\lim_{\theta \to 0} \hat{\xi} = \mu$ uniformly in σ.*

This strongly contrasts with the large cartel case where the only restraint on deviation lies in increasing cost and a small enough σ drives the cartel solution towards competition.

6. Efficiency of Punishment and Fines

We have so far assumed that the punishment P is a cost to the cartel of punishing a bad signal. This would be the case if the punishment involves some form of exclusion. In practice, however, cartels often use fines rather than exclusion: a fine is not a net cost to the cartel, since the cost to the member punished is a gain to the other members. In this case, the cost of punishment to the cartel is ψP where ψ is a measure of the inefficiency of the punishment. In the extreme case of a pure transfer payment $\psi = 0$; in the case that the cost to the member punished is the cost to the cartel - the case we considered earlier - $\psi = 1$. If there are additional costs to other cartel members of exclusion we might also have $\psi > 1$.

A related consideration is that depending on how the costs or benefits of punishments to other cartel members are distributed incentive problem may be created for those providing information and carrying out punishments. For example, a group member might be tempted to falsely report another for violating a quota because they hope to benefit from a fine. In this case the monitors themselves must be monitored so that there are multiple round of monitoring. In fact a strict incentive to punish can be given by punishing those who fail to punish. A simple model of multiple punishment rounds as in Levine and Modica (2016) shows that this amounts to add a multiplicative factor to monitoring costs -

it precludes charging the monopoly price. Davidson and Deneckere (1990) focus on the case in which firms collude in price but not in capacity and show that the level of collusion and the level of excess capacity are both decreasing in the cost of capacity. Finally, Compte, Jenny and Rey (2002) consider the case of asymmetric capacity constraints and show that the effect of this asymmetry on the level of collusion depends on aggregate capacity.

equivalent to increasing the parameter ψ. In other words, ψ may be interpreted as arising from a reduced form of a model with repeated punishment.

As is usual in mechanism design theory, the design problem is separated from the implementation. It may be useful to briefly describe how this system, including punishment rounds, can be decentralized on a peer network as developed, for example, in Levine and Modica (2016). In this implementation it is not necessary that the signal be commonly observed. In that implementation group members are aligned on a circle and each member observes the signal of the clockwise neighbor and is also responsible for issuing a punishment when the signal is bad. In subsequent rounds each member again observes a signal of the clockwise neighbor, now a signal about whether the neighbor "did their duty" by observing and punishing, and again carries out a punishment on a bad signal. In this setting members "do their duty" not because they are indifferent, but because they will be punished by their neighbor in a subsequent round if they fail to do so - and the incentives can easily be made strict.

If we consider the model with general values of ψ, the first observation is that the basic theory changes quantitatively but not qualitatively. That is, we have defined monitoring difficulty as $\theta \equiv \pi/(\Pi - \pi)$. We can incorporate the inefficiency of punishment by defining instead $\theta \equiv \psi\pi/(\Pi - \pi)$ and the earlier results remain valid. We see immediately that a more efficient punishment technology improves possibilities for cartelization. This idea that transfer payments enhance collusion appears in the repeated game literature - it is the basis for the folk theorem with imperfect private information where patience allows future punishments to take the form of efficient transfer payments rather than inefficient pure punishments (see Fudenberg, Levine and Maskin (1994)). In the mechanism design literature, see for example D'Aspremont and Gerard-Varet (1979), the same idea is captured through budget balance. Here transfer payments may be in the form of fines or the buy-backs documented by Harrington and Skrzypacz (2011). However, as Myerson and Satterthwaite (1983) showed not all information technologies allow the first best to be obtained with budget balance so it can be optimal to "burn money", corresponding to socially costly punishments.[19]

Both fines and exclusion seem common in practice. Estimates of the fraction of cartels using exclusion[20] range from 5 to 27% and those using fines[21] from 4 to 64%. These however involve very different samples: Hyytinen, Steen and Toivanen (2019) report that in their sample 27% use exclusion and 15% fines.

One reason to use exclusion is that transfers are not a panacea. First, transfers must ultimately be backed by some other more costly form of punishment - the punishment for refusing a transfer cannot simply be a larger transfer that will also be refused. For example, in Harrington and Skrzypacz (2011) transfers are backed by collective punishment for refusal to pay.[22] Second, transfer payments are generally not 1-1 - that is there are costs and inefficiencies of collecting fines or buybacks - as with taxes - and the value to the recipient will generally be less than the cost to the payer. Finally, transfers introduce wrong incentives for monitoring. Those who receive the transfers have an incentive to make false accusations.[23] Indeed, if any cartel member can secretly plant false evidence - switch a signal to a bad signal - then all will choose to do so and the signal will be useless. All of these considerations lead to the conclusion that while we may get a substantial reduction in ψ by using transfers, it is unlikely to be zero.

There are two key conclusions: first, if transfers are available they are likely to be used, and second, if transfers are available cartelization is more likely. In particular in the case of large cartels, high supply elasticity σ may be offset

[19]Notice that fines are an individual not a collective punishment. As we have observed this is useful in a cartel of any size. This point does not always come across clearly in empirical studies: a good example is Hyytinen, Steen and Toivanen (2018) who study legal cartels in Finland. One piece of evidence they use to show that legally registered cartels are not merely paper entities is the use of fines. On the other hand the theory on which their study is based is that of repeated games and price wars as punishment.

[20]Hay and Kelley (1974), Frass and Greer (1977), Hyytinen, Steen and Toivanen (2019).

[21]Posner (1970), Suslow (2005), Fink et al (2017), Hyytinen, Steen and Toivanen (2019).

[22]In their setting a firm refuses to pay by lying about output.

[23]We see this, for example, when local governments lower speed limits and reduce the length of yellow lights in order to increase revenue from traffic fines. Indeed the problem of false accusations is an ancient one - one element of the code of Hammurabi is punishment for false accusation - see Fudenberg and Levine (2006).

by a low value of ψ. A case in point is that of the Consorzio Grana Padano, that owns the trademark for Grana Padano cheese in Italy. It is a reasonably large group consisting of about 200 producers and collects fees from farmers and monitors quality of cheese production before authorizing the use of the trademark.[24] The Consorzio was fined by the Italian Competition Authority (decision 4352) essentially for imposing fines to members producing too much.[25] The Consorzio has very good monitoring technology (they have inspectors on the floor of the producers) and they believed that their system of fines was legal. Hence in this case θ was quite small both because π is small and because ψ is small. The cartel activities were also very visible and as a result anti-trust action was effective. The general implication for anti-trust authorities is that since fines are easy to observe, illegalizing them should be an effective means of reducing cartelization.

In conclusion, cost elasticity is less likely to inhibit the formation of large cartels when fines can be used as punishment. When membership in the association is crucial it will be relatively easy to enforce fines. Two contexts in which this is true are trademark associations and sports leagues. Since cartelization is useful in these industries so is an exemption from anti-trust law. We do not think it is a coincidence that sports leagues lobby heavily and in many cases successfully, for exemption from anti-trust laws.

7. Discussion

In this section we discuss a number of alternative explanations to our motivating puzzle. First, it could be argued that monitoring is more difficult in a cartel setting than in a public goods setting, hence a collusive arrangement more difficult to enforce. However, it is not immediately obvious that farmers living in a farm community are less able to observe how many fields their neighbors plant than they are to observe whether their neighbors contribute to farm lobbying efforts. In manufacturing monitoring of prices is difficult, and

[24] See https://www.granapadano.it/public/file/201904USA-34282.pdf

[25] See Consorzio Grana Padano, 24 June 2004, n. I569, Bulletin 26/2004, and Siragusa, Beretta and Bay (2019).

perhaps even monitoring of outputs. But monitoring of inputs is not so difficult. If manufacturing firms agreed to limit themselves to one six hour shift a day - in respect of workers rights - that would not only be relatively easy to monitor but would be unlikely to violate anti-trust laws. We should indicate as well that cartel practices, as documented, for example, in Marshall, Marx and Samkharadze (2019) indicate that cartels are relatively effective in monitoring and auditing each others practices.

Second, it is certainly true that public policy and anti-trust law play a role in inhibiting cartel formation but are not directed against lobbying. Our view is that while legal restrictions may be part of the answer it is unlikely to be the entire answer. In practice most anti-trust activity is directed against small cartels: for example the average number of firms in a cartel pursued by the European Commission is 7.61 (see Ordóñez-De-Haro, Borrell and Jiménez (2018)). As we have observed input restrictions are not so likely to run afoul of anti-trust laws - manufacturing firms can hide collusion as concern over workers rights. In a similar way if farmers got together and talked about colluding to reduce output this would be legally problematic. But if they get together - as they do - to discuss best farming practices and agree that a number of fields should be left fallow, that less fertilizers and less intensive farming is a better practice - and this could be successfully enforced as it is in the case of contributions to lobbying efforts - it seems unlikely it would run afoul of anti-trust policy. Indeed, most governments encourage farmers to discuss and adapt best farming practices - often even subsidize them to do so. In fact, even when cartels are legal, the existing empirical evidence seems to support the idea that large cartels are not very common. For example Haucap, Heimeshoff and Schultz (2010) documents that the median number of members for legal cartels authorized by the German Federal Cartel Office (FCO) between 1958 and 2004 was four and since the median number of members of illegal cartels in the same period was five, there is little evidence that the size distribution of legal and illegal cartels is much different.[26]

[26]In particular Haucap, Heimeshoff and Schultz (2010) comment: "the difference between the number of members in legal (average:10, median: 4, modus: 2) and in illegal cartels

Third, it must be observed that some industries with a large number of "firms" do indeed successfully restrict output. Workers are often able to exploit their monopsony power, especially in a labor union setting. Even without a labor union an informal agreement not to "work too hard" is common. Since the demand for effort is downward sloping, workers as a group can take advantage of their monopsony power by reducing effort - and indeed they often do exactly that. Furthermore, while nowadays we think of labor unions as encouraged by and supported by government this has not always been true. In the early 20th Century labor unions were actively discouraged by governments - and indeed union members were sometimes murdered as happened, for example, in 1927 at the Columbine mine massacre (Zieger (1994)). More recently the Solidarity Union in Poland operated in a hostile political environment. Nevertheless these unions were effective in restricting input.

In the model here lobbying and cartelization are viewed as independent activities. In practice this may not be the case and models that examine the competition for resources and complementarity between the two activities are an interesting line for future research. Here, we may informally observe that successful lobbying alleviates the need for cartels. This could be the case of price support or output limitation schemes that we sometimes see in agriculture. It is not true for direct subsidies which are also common in agriculture. However, the most common form of successful lobbying are restrictions on competition - illegalizing domestic competitors, for example in the case of taxis, or more commonly limitations on foreign competition.[27] Success at these types of lobbying

(average: 9, median: 5, modus: 4) is surprisingly small. Since legal cartels are (a) more stable as they can be enforced in the courts and (b) easier to organize as they do not have to be kept secret, the number of members in legal cartels should be higher than in illegal cartels. However, from our data we cannot observe a significant difference." Dick (1996) studies a few large cartels made up of small firms that arose under the Webb-Pomerene Act. These were different than the type of cartels studied here, not engaging in restrictive trading practices, but rather joint marketing. We should note that joint marketing does pose a free-rider problem, but one very similar to the lobbying problem.

[27]Other examples can be found in the advertising restrictions historically imposed by several professional service organizations, see for example Benham (1972), or in the anticompetitive effects of occupational licensing, see for example Pagliero (2011) for recent empirical evidence on US lawyers. In his survey on occupational licensing Kleiner (2000) mentions "[...] dentist's

raises rather than lowers the incentive for cartelization by increasing monopoly power. Moreover, it should not escape attention that lobbying is endogenous: those organizations that we find would most easily be able to cartelize - labor unions, sports leagues - should lobby for legal cartelization. Organizations such as manufacturing firms that would find it difficult to cartelize should lobby instead for restrictions on competition. This seems to be the case.

8. Conclusion

In practice trade associations can both lobby and form cartels and must allocate resources between the two. Our goal here is to understand a simpler and more conceptual point: what is the difference between the free rider problem in lobbying and in sustaining a cartel? We have used the same model for both lobbying and cartel formation in order to isolate the effect of market organization. There is no reason to presume that the technology for producing resources to be used for lobbying is the same as market production technology. What our results show, however, is that this is probably not an important reason why lobbying is so much more common than cartel formation. In particular, the value of the prize or market plays little role for cartels. While it plays an important role in lobbying (with more valuable prizes likely to elicit greater lobbying effort), there is reason to think that the size of the prize is limited in practice by opposition from those who have to pay the subsidy (Levine and Modica (2017)). Rather, our results direct attention to two key variables: the difficulty of monitoring and the elasticity of supply.

The difficulty of monitoring plays a key role in both lobbying and cartels: if monitoring is difficult then public goods such as lobbying and cartels will not be provided by trade associations. It may be that there are important differences in these costs between lobbying efforts and cost of sustaining a cartel - but it is neither obvious nor evident that this is the case. The second key variable is supply elasticity. We find this relatively unimportant in lobbying but crucial for

organizations have attempted creative political and legal maneuvers to block licensed dental hygienists from opening independent shops without the supervision of a licensed dentist".

cartels. If it is relatively cheap to increase output, incentives to cheat on a cartel are great and cartel formation will be inhibited. If - by contrast - it is difficult to increase output beyond basic capacity, the cartel problem is relatively similar to the lobbying problem.

We can illustrate our results by contrasting three industries:

1. Manufacturing firms: it is relatively easy for manufacturers to observe each others activities but firms can easily expand in size by hiring more inputs.

2. Plant workers: it is relatively easy for workers on a factory floor to observe each others effort but workers are physically limited in how much they can increase individual output.

3. Hair dressers: like plant workers hair dressers are physically limited in how much they can increase individual output, but they are diffused in many locations and cannot easily monitor each other. Here we view hair dressers as representative of a class of service workers who are diffused to many locations.

The theory then predicts the pattern given in the table below: manufacturers should be effective at lobbying but not cartelization, plant workers at both, and hair dressers at neither.

industry	monitoring cost	supply elasticity	lobbying	cartel
manufacturing	low	high	yes	no
plant workers	low	low	yes	yes
hair dressers	high	low	no	no

That manufacturers are good at lobbying and better at lobbying than forming cartels is perhaps not so controversial, as is success of unionized workers in both lobbying and cartel operations. What about hair dressers and similar service workers? The U.S. Bureau of Labor Statistics report unionization by different occupational categories: in 2017-2018 only 6.6% of "Personal care and service" workers were unionized, in contrast to 20.2% of "Construction and extraction" workers. Also similar to plant workers, school teachers are heavily concentrated in particular locations, and "Education, training, and library" workers have a 37.2% unionization rate. Further with respect to lobbying, if we examine the lobbying records[28] of the large U.S. state of California among the top ten we find teachers, various business organizations, and one service employee organization, the California State Council of Service Employees, but this according to their website[29] represents highly concentrated and unionized service workers, not more diffuse groups such as hair dressers. So generally speaking we find the facts in accord with the theory.[30]

[28]https://prd.cdn.sos.ca.gov/lobreport2005/Lobbyist_Report_2005.pdf

[29]http://seiuca.org/about/

[30]In practice there is a continuum of characteristics of service workers ranging from doctors, who participate in common training and work together in hospitals, and hence are more like manufacturing workers and are relatively effective at lobbying and cartelization and hairdressers on the other end. Other workers such as optometrists fall some where in between with corresponding intermediate effectiveness as indicated in Benham (1972).

References

Acemoglu, D. and J. Robinson (2001): "Inefficient Redistribution," *American Political Science Review* 95: 649-661.

Al-Najjar, N. I. and R. Smorodinsky (2001): "Large Nonanonymous Repeated Games," *Games and Economic Behavior* 37: 26-39.

Benham, L. (1972): "The Effect of Advertising on the Price of Eyeglasses," *Journal of Law and Economics* (15): 337-52.

Benoit, J.P. and V. Krishna (1987): "Dynamic Duopoly: Prices and Quantities," *Review of Economic Studies* 54(1): 23-35.

Bernheim, D. B. and E. Madsen (2017): "Price Cutting and Business Stealing in Imperfect Cartels," *American Economic Review* 107: 387–424.

Compte, O., F. Jenny and P. Rey (2002): "Capacity Constraints, Mergers and Collusion," *European Economic Review* 46(1): 1-29.

Davidson, C. and R. Deneckere (1990): "Excess Capacity and Collusion," *International Economic Review* 31(3): 521-541.

d'Aspremont, C. and L. Gérard-Varet (1979): "Incentives and Incomplete Information," *Journal of Public Economics* 11: 25-45.

Dick, A. R. (1996): "When are Cartels Stable Contracts?" *Journal of Law and Economics* 39: 241-283.

Dutta, R., D. K. Levine and S. Modica (2018): "Peer Monitoring, Ostracism and the Internalization of Social Norms," EUI.

Fink, N., Schmidt-Dengler, P., Stahl, K. and Zulehner, C. (2017): "Registered Cartels in Austria: Overview and Governance Structure," *European Journal of Law and Economics*, 44: 385-422.

Frass, A. and Greer, D. (1977): "Market Structure and Price Collusion: an Empirical Analysis," *Journal of Industrial Economic*s, 26: 21-44.

Fudenberg, D. and D. K. Levine (2006): "Superstition and Rational Learning," *American Economic Review* 96 (3): 630-651.

Fudenberg, D., D.K. Levine and E. Maskin (1994): "The Folk Theorem with Imperfect Public Information," *Econometrica*, 62(5): 997-1039.

Fudenberg, D., D.K. Levine and W. Pesendorfer (1998): "When are Nonanonymous Players Negligible?" *Journal of Economic Theory*, 79: 46-71.

Genesove, D. and W. P. Mullin (2001): "Rules, Communication, and Collusion: Narrative Evidence from the Sugar Institute Case," *American Economic Review* 91: 379-405.

Green, E. J. (1982): "Noncooperative Price Taking in Large Dynamic Markets," In *Noncooperative Approaches to the Theory of Perfect Competition*: 37-64.

Green, E. J. and R. H. Porter (1984): "Noncooperative Collusion Under Imperfect Price Information," *Econometrica* 52(1): 87-100.

Grossman, G. M., and E. Helpman (2001): *Special interest politics*, MIT press.

Harrington, J. E., and A. Skrzypacz (2011): "Private Monitoring and Communication in Cartels: Explaining Recent Collusive Practices," A*merican Economic Review,* 101: 2425-2449.

Hay, G. and Kelley, D. (1974): "An Empirical Survey of Price Fixing Conspiracies," *Journal of Law and Economics*, 17: 13-38.

Hyytinen, A., F. Steen, and O. Toivanen (2018): "Cartels Uncovered," *American Economic Journal: Microeconomics* 10: 190-222.

Hyytinen, A., Steen, F., and Toivanen, O. (2019): "An Anatomy of Cartel Contracts," *Economic Journal* (129): 2155-2191.

Haucap, J., U. Heimeshoff and L. M. Schultz (2010): "Legal and Illegal Cartels in Germany Between 1958 and 2004," No 08, DICE Discussion Papers, University of Düsseldorf.

Kandori, M. (1992): "Social Norms and Community Enforcement," *The Review of Economic Studies* 59(1): 63-80.

Kleiner, M. M. (2000): "Occupational licensing," *Journal of Economic Perspectives* 14: 189–202

Levine, D. K.and S. Modica (2016): "Peer Discipline and Incentives within Groups," *Journal of Economic Behavior and Organization* 123: 19-30

Levine, D. K.and S. Modica (2017): "Size, Fungibility, and the Strength of Lobbying Organizations", *European Journal of Political Economy* 49: 71-83

Levine, D. K. and W. Pesendorfer (1995): "When are Agents Negligible?" *American Economic Review.*

Lowder, S. K. , J. Skoet, and T. Raney (2016): "The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide," *World Development* 87.

Marshall, R. C., L. M. Marx and L. Samkharadze (2019): "Monopolization Conduct by Cartels", Working Paper.

Myerson, Roger B., and Mark A. Satterthwaite (1983): "Efficient Mechanisms for Bilateral Trading," *Journal of Economic Theory* 29: 265-281.

Olson Jr., M. (1965): *The Logic of collective action: public goods and the theory of groups*, Harvard Economic Studies.

Ordóñez-De-Haro, J., R. Borrell and J. L. Jiménez (2018): "The European Commission's Fight Against Cartels (1962 – 2014): A Retrospective and Forensic Analysis," *Journal of Common Market Studies* 56, 1087-1107.

Ostrom, E. (1990): *Governing the commons: the evolution of institutions for collective action*, Cambridge university press.

Pagliero, M. (2011): "What is the Objective of Professional Licensing? Evidence from the US Market for Lawyers," *International Journal of Industrial Organization* 29: 473-483.

Pai, M. M., A. Roth and J. Ullman (2014): "An Anti-folk Theorem for Large Repeated Games with Imperfect Monitoring," arXiv preprint arXiv:1402.2801.

Pecorino, P. (1998): "Is There a Free-rider Problem in Lobbying? Endogenous Tariffs, Trigger Strategies, and the Number of Firms," *American Economic Review* 88: 652-60.

Posner, R. (1970): "A Statistical Study of Antitrust Enforcement," *Journal of Law and Economics*, 13: 365-419.

Sabourian, H. (1990): "Anonymous Repeated Games with a Large Number of Players and Random Outcomes," *Journal of Economic Theory*, 51(1), 92-110.

Siragusa, M., M. Beretta, and M. Bay (2009): *Competition Law in Italy: The first 20 years of law and practice*, Cleary Gottleib.

Suslow, V. Y. (2005): "Cartel Inter-war International Contract Duration: Empirical Evidence from Cartels," *Industrial and Corporate Change*, 14: 705-44.

Zieger, R. H. (1994): *American workers, American unions*, Johns Hopkins Univ. Press.

Appendix

For the formal analysis in the sequel we observe that the maintained continuity assumption enables us to define bounds which are used in the proofs: for $x \in [0, \overline{x}]$ we have $0 < \underline{\eta} \leq |h''(x)| \leq \overline{\eta}$ and $0 < \underline{\rho} \leq |r'(x)| \leq \overline{\rho}$. These bounds depend only on h. Furthermore, as a last piece of notation, the H_i's that we will use below are positive constants that depend only on h (and not on V, σ and θ).

Proof of Theorem 1

Theorem 1 is implied by the following

Theorem 4. *There are positive constants H_1, H_2 depending only on h such that the unique optimal norm $\hat{\xi}$ satisfies*

1. $\hat{\xi} = 0$ *if and only if* $Vh'(0) \leq 1 + \theta$
2. $0 < \chi - \hat{\xi} \leq H_1 \min\left\{[\theta/V](1+\sigma), \theta + \sqrt{\theta}\right\}$
3. $\hat{\xi} > \min\{1, H_2(h'(0) - (1+\theta)/V)\}$.

Proof. The objective function U is continuous and strictly concave on the compact set $[0, \overline{x}]$ so an optimal norm exists and it is unique and may be characterized by the derivative $U'(\xi) = Vh'(\xi) - (1+\theta)C'(\xi)$.

In the interior either $\xi > 1$ or $Vh'(\xi) = 1 + \theta$. In the latter case we use Taylor's theorem $h'(\xi) \geq h'(0) - \overline{\eta}\xi$ to get result 3 with $H_2 = 1/\overline{\eta}$.

At $\xi = 0$ we have $C'(0) = 1$ so the necessary and sufficient for a corner solution in which the optimal norm is $\hat{\xi} = 0$ is $Vh'(0) - (1+\theta) \leq 0$, which is point 1 in the statement.

Suppose $Vh'(0) - (1+\theta) > 0$ so the solution is interior. For $\xi \geq \chi$ we have $U'(\xi) = Vh'(\xi) - (1+\theta)C'(\xi) \leq Vh'(\chi) - (1+\theta)C'(\chi) = -\theta C'(\chi) < 0$ so the optimum norm satisfies $\xi < \chi$. Set $z = \chi - \xi > 0$. We have

$$U'(\xi) = Vh'(\chi - z) - (1+\theta)(1 + \sigma \max\{0, \chi - z - 1\}).$$

Using $Vh'(\chi) = 1 + \sigma \max\{0, (\chi - 1)\}$ we may write this as

$$U'(\xi) = V[h'(\chi - z) - h'(\chi)] + \sigma(\max\{0, \chi - 1\} - \max\{0, \chi - z - 1\}) + \\ - \theta(1 + \sigma \max\{0, \chi - z - 1\}) \geq V\underline{\eta} z - \theta(1 + \sigma \max\{0, \chi - z - 1\}).$$

There are two cases: $z \geq \chi - 1$ and $z \leq \chi - 1$. In the former case we have $U'(\xi) \geq V\underline{\eta} z - \theta$ so that the necessary and sufficient condition for the social optimum $U'(\xi) = 0$ requires $V\underline{\eta} z - \theta \leq 0$ or $z \leq (1/\underline{\eta}) \cdot (\theta/V)$. If $z \leq \chi - 1$ (so that in particular $\chi > 1$) then we have $U'(\xi) \geq V\underline{\eta} z - \theta(1 + \sigma\overline{x})$ so that $z \leq (1/\underline{\eta})(1 + \sigma\overline{x})\theta/V$. From $Vh'(\chi) = 1 + \sigma(\chi - 1)$ we have $Vh'(1) \geq \sigma(\chi - 1)$ so that $\chi - 1 \leq Vh'(1)/\sigma < Vh'(0)/\sigma$. Hence

$$\begin{aligned} z &\leq \max\left\{\frac{1}{\underline{\eta}}\frac{\theta}{V}, \min\left\{\frac{1}{\underline{\eta}}\frac{\theta}{V}(1 + \sigma\overline{x}), \frac{Vh'(0)}{\sigma}\right\}\right\} \\ &\leq \min\left\{\frac{1}{\underline{\eta}}\frac{\theta}{V}(1 + \sigma\overline{x}), \frac{1}{\underline{\eta}}\frac{\theta}{V} + \frac{Vh'(0)}{\sigma}\right\} \\ &\leq \frac{1}{\underline{\eta}}\frac{\theta}{V}(1 + \sigma\overline{x}) \leq \left[\frac{1}{\underline{\eta}}\max\{1, \overline{x}\}\right](1 + \sigma)\frac{\theta}{V} \end{aligned}$$

giving the first half of condition in point 2.

For the second half, starting from the first inequality above we can write

$$\begin{aligned} z &\leq \left[\frac{1}{\underline{\eta}}\overline{x} + h'(0)\right] \min\left\{\frac{(1 + \sigma)\theta}{V}, \frac{\theta}{V} + \frac{V}{\sigma}\right\} \\ &= \left[\frac{1}{\underline{\eta}}\overline{x} + h'(0)\right] \left[\frac{\theta}{V} + \sqrt{\theta}\min\left\{\frac{\sigma\sqrt{\theta}}{V}, \frac{V}{\sigma\sqrt{\theta}}\right\}\right] \\ &\leq \left[\frac{1}{\underline{\eta}}\overline{x} + h'(0)\right] \left(\frac{\theta}{V} + \sqrt{\theta}\right) \leq \left[\left(\frac{1}{\underline{\eta}}\overline{x} + h'(0)\right)\max\{1, h'(0)\}\right]\left(\theta + \sqrt{\theta}\right), \end{aligned}$$

which ends the proof letting

$$H_1 = \min\left\{(1/\underline{\eta})\max\{1, \overline{x}\}, \left((1/\underline{\eta})\overline{x} + h'(0)\right)\max\{1, h'(0)\}\right\}.$$

□

Proof of Theorem 2

We show the following, which implies the result stated in the text.

Theorem 5. *There are H_3, H_4 such that the optimal norm $\hat{\xi}$ satisfies*

1. $\mu < \hat{\xi} \leq \chi$, with second inequality strict if $\chi > \theta/(1+\theta)\,(= \pi/\Pi)$

2. $\hat{\xi} - \mu \leq H_3\theta\,(1 + [V/\sigma])$ and $\chi - \hat{\xi} \leq H_4/\,(\theta(1 + [V/\sigma]))$

Payoff differences are similarly bounded:

3. Assuming $\chi \geq 1$, $[W(\mu) - U(\hat{\xi})]/W(\mu)$ has the same bounds as $\hat{\xi} - \mu$, and $[U(\hat{\xi}) - W(\chi)]/W(\mu)$ has the bounds of $\chi - \hat{\xi}$.

Proof. We need to compute the derivative of $U(\xi) = (1+\theta)W(\xi) - \theta\,[p(\xi)\hat{x}(\xi) - C(\hat{x}(\xi))]$. Since $\hat{x}$ is characterized by $p(\xi) = C'(\hat{x})$ the derivative of the second term is just $\theta p'(\xi)\hat{x}$. After substituting the expressions of $C'(\xi)$ and $\hat{x}$ we then get

$$U'(\xi) = (1+\theta)\,[p(\xi) + p'(\xi)\xi - C'(\xi)] - \theta p'(\xi)\hat{x}$$

First we show that the optimal norm satisfies $\mu < \xi \leq \chi$ with second inequality strict when $(1+\theta)\chi > \theta$. For $\xi \leq \mu$ we have $U'(\xi) \geq -\theta p'(\xi)\hat{x}$ so the optimum satisfies $\xi > \mu$. For $\xi > \chi$ lowering ξ increases profits and relaxes the incentive constraint, so certainly $\xi \leq \chi$. Moreover when $\xi = \chi$ then $p = C'$ and $\hat{x} = \max\{1, \chi\}$ so $U'(\chi) = [(1+\theta)\chi - \theta\max\{1,\chi\}]\,p'(\chi)$, which is strictly negative for $(1+\theta)\chi > \theta$. This proves the point 1.

To get bounds on the norms we start from U'. Recall that $p(\xi) = Vh'(\xi)$, $p(\xi) + p'(\xi)\xi = Vr(\xi)$, $C'(\xi) = 1 + \sigma\max\{0, \xi - 1\}$ and $\hat{x} = 1 + (p(\xi) - 1)/\sigma$. Therefore we may write

$$U'(\xi) = (1+\theta)\,[Vr(\xi) - (1 + \sigma\max\{0, \xi - 1\})] - \theta Vh\text{"}(\xi)\,[1 + (Vh'(\xi) - 1)/\sigma]\,.$$

We now take $z = \xi - \mu$ and look for an upper bound on U'. We have

$$\begin{aligned} U'(\xi) = &\; (1+\theta)\,[Vr(\mu + z) - (1 + \sigma\max\{0, \mu + z - 1\})] + \\ &\qquad -\theta Vh''(\xi)\,(1 + (Vh'(\xi) - 1)/\sigma) \\ \leq &\; -(1+\theta)V\underline{\rho}z + \theta V\overline{\eta}\,(1 + (Vh'(0) - 1)/\sigma)\,. \end{aligned}$$

This is negative for

$$z > \frac{\theta}{1+\theta} \frac{\overline{\eta}\,(1 + (Vh'(0) - 1)/\sigma)}{\underline{\rho}}$$

so

$$\begin{aligned}
\xi - \mu &\leq \frac{\theta}{1+\theta} \frac{\overline{\eta}\,(1 + (Vh'(0) - 1)/\sigma)}{\underline{\rho}} \\
&\leq \left[\frac{\max\{h'(0), 1\}\overline{\eta}}{\underline{\rho}}\right] \theta\,(1 + [V/\sigma]) = H_3\theta\,(1 + [V/\sigma])\,.
\end{aligned}$$

Next take $z = \chi - \xi$ and look for a lower bound on $U'(\xi)$. From $Vh'(\chi) = 1 + \sigma \max\{0, \chi - 1\}$ we have

$$\begin{aligned}
U'(\xi) &\geq (1+\theta)\left[V\underline{\eta}z + Vh''(\xi)(\chi - z)\right] - \theta Vh''(\xi)\left(\chi + V\underline{\eta}z/\sigma\right) \\
&= (1+\theta)\left[V\underline{\eta}z - Vh''(\xi)z\right] + (1+\theta)Vh''(\xi)\chi - \theta Vh''(\xi)\left(\chi + V\underline{\eta}z/\sigma\right) \\
&\geq 2(1+\theta)V\underline{\eta}z - V\overline{\eta}\overline{x} + \theta V^2\underline{\eta}^2 z/\sigma.
\end{aligned}$$

This is positive for

$$z > \frac{\overline{x}}{2(1+\theta) + \theta V\underline{\eta}/\sigma}$$

so

$$\chi - \xi \leq \frac{\overline{x}}{2(1+\theta) + \theta V\underline{\eta}/\sigma} \leq \left[\frac{\overline{x}}{\min\{2, \underline{\eta}\}}\right] \frac{1}{\theta(1 + V/\sigma)} = \frac{H_4}{\theta(1 + V/\sigma)}.$$

In order to bound payoff differences, first we compute a lower bound on monopoly profits: $W(\mu) \geq (Vh'(1/2) - 1)\,(1/2)$, and the assumption $\chi \geq 1$ is equivalent to $Vh'(1) \geq 1$, so that $W(\mu) \geq (Vh'(1/2) - Vh'(1))\,(1/2) = (h'(1/2) - h'(1))\,V/2 \equiv \underline{W}$.

We then notice that $W(\mu) - U(\xi) > W(\xi) - U(\xi) = \pi P(\xi) \geq 0$. For the upper utility bound on $W(\mu) - U(\xi)$ observe that the optimal norm satisfies $U(\xi) \geq U(\mu)$ so that $W(\mu) - U(\xi) \leq W(\mu) - U(\mu)$. We have $U(\mu) = W(\mu) - \theta(p(\mu)\hat{x} - C(\hat{x}))$ and $p(\mu)\hat{x} - C(\hat{x}) \leq Vh'(\mu)\hat{x} - \hat{x}$. Combining these with $\hat{x} = 1 + (Vh'(\mu) - 1)/\sigma$ gives $W(\mu) - U(\xi) \leq \theta\,(Vh'(\mu) - 1)\,(1 + (Vh'(\mu) - 1)/\sigma)$. Observe that $Vh'(\mu)-1 \le Vh'(0)$ giving $W(\mu)-U(\xi) \le \theta Vh'(0)\,(1+Vh'(0)/\sigma) \le \theta Vh'(0)\max\{1,h'(0)\}(1+[V/\sigma])$. Dividing by the lower bound on monopoly profits $\underline{W}$ gives the bound in the theorem:

$$\frac{W(\mu)-U(\xi)}{W(\mu)} \le \left[\frac{2h'(0)\max\{1,h'(0)\}}{h'(1/2)-h'(1)}\right]\theta\,(1+[V/\sigma])\,.$$

Next, $U(\xi)-W(\chi)>0$ since $P(\chi)=0$. For the upper utility bound observe that the utility gain from ξ over χ is less than the profit gain because $P(\xi)$ gets larger as ξ goes down. Reducing output by $\chi-\xi$ raises price by no more than $\overline{\eta}V(\chi-\xi)$ and saves at most marginal cost times $\chi-\xi$ and that marginal cost is at most $Vh'(1)$. Hence $U(\xi)-W(\chi)\le V(\overline{\eta x}+h'(1))(\chi-\xi)$, and dividing by $\underline{W}$ we get

$$\frac{U(\xi)-W(\chi)}{W(\mu)} \le 2\frac{\overline{\eta x}+h'(1)}{h'(1/2)-h'(1)})(\chi-\xi).$$

□

Proof of Theorem 3

We show the following, which implies the result stated in the text.

Theorem. *There exists a positive constant H_5 depending only on h such that the optimal norm $\hat{\xi}$ satisfies $\hat{\xi}-\mu \le H_5\theta(N-1)$.*

Proof. First we observe that $\hat{x}$ is bounded independent of σ: indeed we have $\hat{x} \le N\overline{x}$ as the individual firm will never choose industry output to be above the satiation level. Next we look at the objective function

$$U(\xi)=(1+\theta)W(\xi)-\theta\,[p([(N-1)\xi+\hat{x}(\xi)]/N)\hat{x}(\xi)-C(\hat{x}(\xi))]\,.$$

By the envelope theorem the derivative is

$$U'(\xi)=(1+\theta)\,[p(\xi)+p'(\xi)\xi-C'(\xi)]-\theta p'(((N-1)\xi+\hat{x}(\xi))/N)\hat{x}(\xi)(N-1)/N.$$

In the proof of Theorem 2 we showed this results in a bound of the form

$$\hat{\xi}-\mu \le \left[\frac{\max\{h'(0),1\}\overline{\eta}}{\underline{\rho}}\right]\theta\hat{x}(N-1)/N \le \left[\frac{\max\{h'(0),1\}\overline{\eta x}}{\underline{\rho}}\right]\theta(N-1)$$

giving the desired result. $\square$

Competition versus Monopoly

In the text we established bounds on how close the optimal norm is to the competitive and monopoly output. Of course the two could be close to each other. In the next theorem we provide bounds on output and profit differences between competition and monopoly.

Theorem 6. *There exists H_6 such that $\chi - \mu$ and $1 - W(\chi)/W(\mu)$ are both bounded below by*

$$H_6 \min\{\chi, 1\} \left(\frac{1}{1 + [\sigma/V]} + \max\{0, \frac{1}{V} - r(1)\} \right).$$

The case in which $Vh'(0)$ is very close to 1 is uninteresting. In this case χ is close to zero, the market is not very profitable and little is produced regardless of how the industry is organized. Otherwise for the competitive and monopoly solutions to be close together two things must both be true: σ/V must be very large and marginal revenue at aggregate output 1 equal to $Vr(1)$ must be either greater than marginal cost or at least not too much smaller. This corresponds in the limit to the well known case of inelastic supply and marginal revenue above marginal cost at the capacity constraint. In this case there is no difference between monopoly and competition - the monopolist is content to take competitive rents rather than profits. On the other hand when it is easy to increase output above basic capacity, that is, σ is small, then not only is the optimal norm close to the competitive equilibrium in quantity and profit - but the competitive equilibrium is not close to monopoly either in quantity or profit.

Proof. First consider $Vh'(1) \leq 1$ (case $\chi \leq 1$). Marginal revenue at the competitive equilibrium is $Vr(\chi) = p(\chi) + p'(\chi)\chi \leq V(1 - \underline{\eta}\chi)$. Hence from $Vr(\mu) = 1$ we deduce that it must be $\chi - \mu \geq \left[\underline{\eta}/\overline{\rho}\right]\chi$. So the result holds for $Vh'(1) \leq 1$.

For $Vh'(1) > 1$, that is $\chi > 1$ we will show

$$\chi - \mu \geq \frac{\underline{\eta}/2}{\sigma/V + \overline{\rho} + \overline{\eta}} + \frac{\max\{0, 1/V - r(1)\}}{2\overline{\rho}}$$

from which the result follows.

Since $\chi - \mu = \chi - 1 + 1 - \mu$ we bound the two terms. First suppose that $r(1) \leq 1/V$. Then at $x = 1 - [1 - Vr(1)]/V\overline{\rho}$ we have $Vr(x) \leq 1$ so that $x \geq \mu$, or $1 - \mu \geq [1 - Vr(1)]/V\overline{\rho}$. Next we bound $\chi - 1$. Recall that $Vh'(\chi) - (1 + \sigma(\chi - 1)) = 0$. This implies $Vh'(1) - V\overline{\eta}(\chi - 1) - (1 + \sigma(\chi - 1)) \leq 0$, from which

$$\chi - 1 \geq \frac{Vh'(1) - 1}{\sigma + V\overline{\eta}}.$$

For $r(1) > 1/V$ we add together the bounds on $1 - \mu$ and $\chi - 1$ to find

$$\chi - \mu \geq \frac{h'(1) - 1/V}{\sigma/V + \overline{\eta}} + \frac{(1/V) - r(1)}{\overline{\rho}} \geq \frac{1}{2}\frac{h'(1) - 1/V}{\sigma/V + \overline{\eta}} + \frac{(1/V) - r(1)}{\overline{\rho}}$$

$$\geq \frac{1}{2}\frac{h'(1) - r(1)}{\sigma/V + \overline{\eta} + \overline{\rho}} + \frac{1}{2}\frac{(1/V) - r(1)}{\overline{\rho}}.$$

Moreover, $r(1) \leq h'(1) - \underline{\eta}$. This covers the case $r(1) > 1/V$.

Now suppose that $r(1) \leq 1/V$. By the intermediate value theorem for $x \leq y \leq \chi$ we have

$$\begin{aligned} Vr(x) &= Vr(\chi) - Vr'(y)(\chi - x) \leq p(\chi) - V\underline{\eta}\chi + V\overline{\rho}(\chi - x) \\ &\leq 1 + \sigma(\chi - 1) - V\underline{\eta} + V\overline{\rho}(\chi - x) \end{aligned}$$

and $C'(x) \geq 1 + \sigma(1 - x)$. Hence if $1 + \sigma(\chi - 1) - V\underline{\eta} + V\overline{\rho}(\chi - x) - (1 + \sigma(x - 1)) = 0$ then $\mu \leq x$. Solving the equality for x gives the bound

$$\chi - \mu \geq \frac{\underline{\eta}}{\sigma/V + \overline{\rho}} \geq \frac{1}{2}\frac{\underline{\eta}}{\sigma/V + \overline{\eta} + \overline{\rho}}.$$

For the profit bound we have

$$W(\mu) - W(\chi) = \int_{\mu}^{\chi} (-Vr(x) + C'(x))\, dx.$$

Choose $y = \mu + (\chi - \mu)/2$ we have

$$W(\mu) - W(\chi) \geq \int_y^{\chi} (-Vr(x) + C'(x))\, dx.$$

Moreover $-Vr(\mu) + C'(\mu) = 0$ implies for $x \geq y$ that $-Vr(x) + C'(x) \geq V\underline{\rho}(\chi - \mu)/2$ giving the desired bound. □