Proportional Restraints and the Patent System[*]

Erik Hovenkamp[†] Jorge Lemus[‡]

June 22, 2017

Abstract

Patent settlements between rivals must necessarily restrain competition to some extent, but antitrust has not yet developed a clear set of boundaries on what is allowed. Further complicating matters, these settlements forestall a ruling on whether the patent is valid and enforceable. Ideally antitrust would engender "proportionality" between patent quality (the likelihood of validity) and the extent to which competition is diminished. We show that antitrust can accomplish this by simply policing the particular *manner* in which competition is restrained—i.e. the type of restraint being used—and by prohibiting certain side-deals designed to subvert proportionality (such as reverse payments). We show that different restraints vary considerably in the extent to which bargaining outcomes line up with the expected result of litigation. This allows us to infer the degree of proportionality from the type of restraint being used, making it unnecessary to estimate patent quality or price effects.

[PRELIMINARY AND INCOMPLETE. DO NOT CITE]

[*] This is a preliminary and incomplete version. A final version will be available soon. Please contact the authors if you want to see the most up to date version of the paper. Please do not circulate.

[†] Harvard Law School.

[‡] University of Illinois at Urbana Champaign, Department of Economics.

1 Introduction

Patent protection is binary: every claimed invention is entitled to either twenty years of protection or no protection at all. And yet the underlying criteria for patentability—in particular, the novelty and non-obviousness of an invention—are clearly not binary; they take values along a continuum. This contrast would seem to create a problematically coarse system of rewarding inventors. It simply cannot be that every invention warrants either two decades of monopoly or no protection at all. Ideally, the patent system would elicit a certain "proportionality" between (1) the novelty and non-obviousness of a patented invention and (2) the extent to which the patent displaces competition.

In principle, we could try to achieve proportionality at the Patent Office by instructing examiners to assign each invention a customized patent term based on its apparent quality. In practice, however, this would be impossibly difficult to perform reliably; the much simpler binary system is already difficult enough. As such, we argue that the best way to achieve proportionality is not to tweak the patent system itself. Instead, we propose a set of antitrust policies designed to elicit proportionality through markets and private contracting, and which are realistically implementable in practice.

In broad outline, our arguments are as follows: patent settlements must necessarily restrain competition to some extent if they are to line up with the expected result of litigation, which will exclude competing users with positive probability. There are many kinds of restraints that firms can use to soften competition, such as price or output restraints, delayed entry, territorial limitations, stock acquisitions, or plain old royalties. It is usually easy to identify what restraint firms are using in a given settlement, although it is generally quite difficult to produce a dollar-estimate a settlement's static welfare effects.

We show that restraints vary widely in the extent to which bargaining outcomes elicit competitive effects that line up with the expected result of litigation, where the latter is a function of patent quality (the likelihood that the patent will be held valid[1]). Some restraints always allow the firms to capture disproportionately large profits; some have the opposite problem, and are not sufficiently accommodating to permit a mutually-acceptable settlement; and some restraints are just right. Based on this analysis, we

[1] A patent is valid if it meets the legal criteria for patentability. An invalid patent is thus one that was granted erroneously.

submit that antitrust's proper role is to promote proportionality by prohibiting restraints that unreasonably subvert it—as well as certain side-deals designed specifically for that purpose—and to otherwise let firms settle on whatever terms they like. In this way, we rely not on courts' or agencies' estimates of patent quality, but on those of the firms themselves.

To motivate our proposal, it is helpful to begin by explaining how antitrust could hope to achieve proportionality though mostly-passive oversight of private settlement. Variation in patent term is not the only possible way to align a patent's market impact with the quality of the invention. An alternative method is to continue giving a fixed term to everyone, but ensure that a patent's market footprint is persistently smaller or larger throughout this term, depending on the quality of the patentee's contribution. It is in this way that our proposed antitrust framework achieves proportionality. To do so, it relies on a critical but often overlooked feature of the patent system: it allows private parties to challenge patents as invalid. This creates a mechanism by which third party firms may credibly demand some access to the patented technology, although such use may be contractually restrained.

Many litigated patents are held invalid. But like most legal claims, a large majority of patent challenges will settle, and this forestalls a determination of the patent's validity. We view such widespread settlement not as as *inherently* problematic, but rather as something that needs well-conceived oversight in order to produce efficient results. In fact, even if a prospective challenger thinks a patent is probably invalid, litigation to judgment is not necessarily the first-best solution. It might be preferable (and certainly much cheaper) for the parties to reach a settlement that does very little to disrupt competition. This ensures that the patentee's reward is not zero, but low enough to match the quality of his contribution.

We focus on *horizontal* settlements, or those in which the patentee and challenger(s) are competing firms. It is easy to see how challenge threats could work to achieve proportionality through private bargaining. If a patent's quality is very high, rivals do not have a credible threat to challenge, and hence the patentee can fully exclude competitors without having to rely on uncertain litigation. If the patent is of intermediate quality, then the challengers have a credible threat to challenge, but a victory is hardly certain, so they will agree to some significant restraint on their use of the invention—for example, by agreeing to delay their use until several years into the future. Finally, if the patent seems quite weak, then the challengers have substantial leverage, and can credibly demand to use the invention with only a minor restraint—for example, a small royalty. If firms always behaved in this way, we would have a direct positive relationship between patent quality and market impact.

The critical problem is that competing firms always have a joint-interest in dampening competition, whether or not this result is justified by a high quality patent. Thus the firms' preference is to preserve monopoly rents in all cases, and relegate bargaining to the distribution of total profits, which then becomes the only thing on which patent quality has any influence. Thus, when settlements are reached in this way, there is a complete absence of proportionality. As such, if we are to rely on challenge threats to achieve proportionality, we will need some effective—and realistically implementable —antitrust rules to channel the firms' negotiations in an efficient direction.

For a clear example of firms striking deals that undermine proportionality, consider the case of ”pay-for-delay” settlements, also known as “reverse payment” settlements, which are known principally for their prevalence in pharmaceutical markets. In such a settlement, an incumbent monopolist is selling a patent-protected, brand-name drug. Its patent is challenged as invalid by one or more prospective entrants, which are generic drug producers. In a pay-for-delay settlement, the patentee pays the challengers to stop attacking the patent and stay out of the market for some period of time, but no later than the date of patent expiration (any longer and the deal would be transparently illegal). Total profits are maximized when the challengers stay out for the full remainder of the patent term, and this is true regardless of patent quality. Debate over these settlements came to a head in 2013, when the Supreme Court's *Actavis* decision held that pay-for-delay settlements may violate the antitrust laws.[2]

Although the *Actavis* decision helped to lay some antitrust groundwork, it was just a first step in assessing the antitrust implications of one particular kind of restraint—delayed entry by challengers—and of a particular kind of “side-deal” that firms can incorporate into their settlement—reverse payments, i.e. payments from patentees to challengers. The broader application of antitrust to the universe of horizontal patent settlements remains highly unresolved.

In principle antitrust could police horizontal patent settlements by (a) estimating the

[2] *F.T.C. v. Actavis* (2013).

extent to which the settlement has restrained competition relative to the (counterfactual) case of invalidation; (b) estimating the likelihood that the patent would be held valid, if litigated to judgment; and (c) comparing these estimates and deciding if the former seems reasonable in relation to the latter. But, for a number of reasons, this would be difficult if not impossible to implement in most cases. Even if the end results were reliable, the high costs would likely have a strong chilling effect on enforcement.

Our proposal is much simpler and requires neither an estimate of the settlement's impact on competition nor an appraisal of patent quality. We propose that antitrust inquires should focus less on the *extent* to which competition is restrained than on the particular *manner* in which competition is restrained. We present a general model of horizontal patent settlements that subsumes essentially all possible ways a patentee could restrain its rival-challengers. Despite its generality, the model is highly tractable, and can easily be applied to evaluate how a given kind of restraint performs within a given model of competition.

Our tool for evaluating a restraint's proportionality is what we call the restraint's "exclusion rate." As the "magnitude" of the restraint (e.g. the size of a royalty or length of a delay period) is increased, the exclusion rate describes the ratio of (1) the rate at which the challengers' profits are falling; to (2) the rate at which industry profits are rising. A lower exclusion rate means that industry profits can be made higher without further injuring the challengers. As such, restraints that systematically over-suppress competition are characterized by slow exclusion rates. By contrast, if the exclusion rate is too high, the firms will be unable to reach a mutually-acceptable settlement, as the challenger's profits fall to quickly to reach a mutually-acceptable agreement. Ideally, firms would stick to the intermediate class of restraints, which are viable and also generate competitive effects that are commensurate with the expected outcome of litigation.

A given kind of restraint (e.g. a royalty) performs differently within different environments—and its viability as a settlement device may similarly vary. But such variation is not so large as to prevent us from assessing its general propensity to elicit proportional settlements. For example, some restraints are always monotonic in the sense that the challengers always want the magnitude of the restraint (e.g. the size of a royalty) to be as small as possible. But other kinds of restraints (such as output caps) are often nonmonotonic, meaning that a collection of challengers might want to be restrained beyond the level they can credibly demand, so long as it applies to all of them. Indeed, in some cases challengers may prefer some positive level of restraint to the case in which they are not restrained at all, as the benefits of softening competition *between challengers* may outweigh the cost of putting the patentee at a competitive advantage.

The fact that a given kind of restraint may not be viable under all market conditions highlights why it is important for antitrust to be mindful of context-specific factors and market structure. This allows us to assess whether an alternative restraint (namely one that tends to be less proportional) is reasonably necessary within the relevant competitive environment. As an example, royalties are generally proportional *when they are viable*, but they are not viable when there is aggressive price competition between the parties, as challengers may be unable to turn a sufficient profit when they are at a significant cost disadvantage. This likely explains why we rarely see royalties employed in pharmaceutical settlements between brand-name drug makers and their generic competitors: price competition is simply too volatile for the firms to predict how a given cost distortion will ultimately influence competition and the distribution of profits (even if we game theorists can readily answer such questions within a given competitive environment). It is thus not surprising that these firms tend to rely on restraints that diminish competition in more "controllable" ways, such as by delaying generic entry until a pre-determined date.

1.1 Related Literature

Patent litigation disputes are settled before final judgment most of the time. Meurer (1989), Daughety and Reinganum (2005), Bessen and Meurer (2006), among many others,[3] have studied the incentives to settle a (patent) lawsuit. They model settlement as a bargaining game where the disagreement payoff of each party is computed as the expected payoff from going to litigation. Our use follows this setting by modeling firms bargaining under the threat of litigation.

Modeling settlement payoffs in patent disputes simply as a reduced form payoff has become controversial. Early work by Blair and Cotter (2002) questions whether some

[3] Somaya (2003) presents a framework where there are multiple stages at which firm can negotiate a settlement and then performs an empirical analysis.

types of settlements are illegal, specifically pay-for-delay. Since then, the literature on pay-for-delay has grown substantially including the work of Willig and Bigelow (2004), Hemphill (2006), and more recently, Olson and Wendling (2013).

But pay-for-delay is only one out of many ways in which firms can settle. For a given class of settlement agreements, a different strand of the literature examines the optimal choice of settlement terms within that class of settlements, often from the perspective of the patent holder. For example, Amir et al. (2014) that examines different licensing mechanisms and provide conditions such that the licensor chooses a per-unit royalty contract. In the same spirit, Lemley and Shapiro (2013) investigate how firms should settle patent disputes when there are "FRAND commitments." However, none of these papers ask the question of which kind of settlement arrangements are more harmful for competition or consumers.

The closest paper to these idea is by Shapiro (2003) where it is explicitly acknowledged that firms can settle in different ways and antitrust authorities should limit some of them. Although the question is closely related to ours, the framework we present, our results, and policy recommendations are quite different to Shapiro's work. The main difference is that we propose a way to evaluate settlements that does not require an antitrust authority to actively search for information about patent quality. Even more, our framework allow us to evaluate the effect on competition of *any* settlement arrangement in an oligopoly game.

2 Model

The market has $N = n + 1$ firms ($N \geq 2$), indexed by $i \in I \equiv \{0, 1, ..., n\}$. Firm 0 is the patent holder, and its remaining patent term is normalized to $T = 1$. Each firm $j \in J \equiv \{1, ..., n\}$ is a competitor who wants to challenge firm 0's patent to try and enter the market; we refer to these firms as the "challengers." If all n challengers compete and are unrestrained, then there is a symmetric N-firm equilibrium in which each $i \in I$ earns flow profit $\pi^N > 0$. This is the "unrestrained equilibrium." If the patentee operates a monopolist for the full patent term, then it earns a flow profit π^m, and each challenger earn a profit of zero (unless it receives a payment from the patent holder). As this reflects, all payoffs are expressed as profits earned over the patent term, since there will always be unrestrained N-firm competition after the patent expires. Throughout the paper we assume $\pi^m > N\pi^N$, and thus total profits are maximized by monopoly.

Any firm $j \in J$ can challenge the patentee's patent if it does not secure a satisfactory settlement. Any such challenge succeeds with probability $1-\theta$, where $\theta \in [0,1]$ is commonly known by all firms. We thus regard θ as a metric for patent quality. If a challenge is successful, then the patent is invalidated, resulting in the unrestrained equilibrium. If instead the patentee wins in court, then it will retain a monopoly for the full term.[4] Litigation costs are $c \geq 0$ for all firms. In a lawsuit between firms 0 and j, their expected payoffs from litigation are, respectively:

$$\begin{aligned} L_0^c(\theta) &= \theta\pi^m + (1-\theta)\pi^N - c, && (1) \\ L_j^c(\theta) &= (1-\theta)\pi^N - c. && (2) \end{aligned}$$

We define the industry payoff from litigation as $L_I^c(\theta) = \sum_{i \in I} L_i^c(\theta)$. Because we assume that all challengers are identical, L_j^c is independent of the identity of the challenger. We assume that $c < \pi^N$, so that $L_j^c(\theta) \geq 0$ for sufficiently low θ. For a given litigation cost c, the challenger has a credible threat to litigate if $L_j^c(\theta) \geq 0$ and the patentee is willing to defend its patent in court as long as $L_0^c(\theta) \geq \pi^N$. All parties have credible litigation threats if

$$\theta_L \equiv \frac{c}{\pi^m - \pi^N} \leq \theta \leq \theta_H \equiv \frac{\pi^N - c}{\pi^N} \tag{3}$$

Figure 1 illustrates the possible scenarios conditional on the quality of the patent. When $\theta \leq \theta_L$, it is too expensive for the patentee to protect its patent through litigation so there is free entry. When $\theta > \theta_H$, challengers do not have a credible litigation threat, so the patentee preserves its monopoly. When $\theta \in [\theta_L, \theta_H]$, both parties have a credible litigation threat and there is bargaining over a settlement.

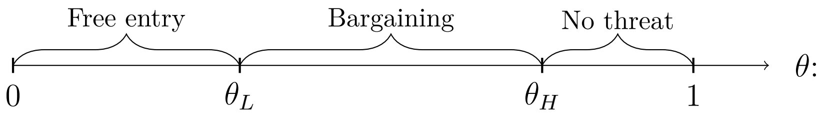

Figure 1: Possible outcomes conditional on patent quality.

[4] This is a simplifying assumption. All we really need for antitrust relevance is that the patent, if upheld by a court, will materially suppress competition in the relevant market.

When $c = 0$, we have $\theta_H = 1$ and $\theta_L = 0$, so bargaining over a settlement is feasible, regardless of the patent quality. When $c > 0$, however, Figure 1 shows that not every patent dispute will lead to litigation, because either the patent holder will not fight back (when $\theta < \theta_L$) or no challenger will have an incentive to challenge the patent in court (when $\theta > \theta_H$). It is immediate to see that in either of these two cases the there is a restrain on competition that is not contingent on the patent quality: When $\theta < \theta_L$, there is not enough restrain on competition (free entry) relative to the patent quality, whereas when $\theta > \theta_H$ there is too much restrain on competition (full exclusion) relative to the patent quality. In the extreme case, when $c \geq \pi^N$, no challenger has incentives to challenge a patent regardless of its quality, and the patent holder will preserve its monopoly position.

The restraint of competition that is introduced by large litigation costs, although it is interesting, it is not the main focus of our paper. We want to understand how different restraints affect the competition when those restraints are the outcome of a settlement. For this reason, we will focus our analysis on the case $\theta \leq \theta_H$ where the restraint on competition does not come from the cost of litigation, but from the settlement agreement reached by the firms reach to avoid litigation.[5]

2.1 Restrictive Settlements

To avoid costly litigation, the patentee can strike settlements with the challengers. Each challengers must receive a deal that is preferable to litigation, or else it will challenge the patent. Since challengers are symmetric, we focus on settlements in which they all receive exactly the same deal. Total expected payoffs from litigation are generally larger than those in the unrestrained equilibrium, and the patentee's payoff is generally larger than those of each challenger. Thus, in order to be unanimously preferred to litigation, a settlement must generally (a) suppress competition relative to the unrestrained equilibrium, and (b) award a larger share of industry profits to the patentee than to each challenger. Following the first condition, any settlement will generate industry profits that are a convex combination of $N\pi^N$ and π^m. Thus, every settlement corresponds to a unique value of $\sigma \in [0, 1]$ such that industry profits are

[5] In the case $\theta < \theta_L$, the patent holder does not have incentives to litigate to defend its patent. However, a settlement may still be feasible because the challengers may want to be restrained. We discuss this case later on in the paper.

given by

$$\begin{aligned}\Pi(\sigma) &= \sigma\pi^m + (1-\sigma)N\pi^N, && (4)\\ &= N\pi^N + \sigma\mu, && (5)\end{aligned}$$

where $\mu \equiv \pi^m - N\pi^N > 0$ denotes the difference in total industry profits from preserving a monopoly versus engaging in competition. As such, we can think of σ as capturing the *magnitude* of any given restraint on competition. For example, σ could represent the size of a royalty or the length of the delay of rival's entry date. In general, any restraint can be used to generate any value of σ by properly adjusting its magnitude. But then what distinguishes one restraint from another? The answer is that they differ in how industry profits are *distributed* among firms at a given level of σ. Thus, we represent a restraint with a function that, for each magnitude σ, describes the share of industry profits allocated to each challenger. Given that we treat challenger symmetrically, we only need to specify a number and not a vector.[6]

Definition 1. *A* ***restraint*** *is a function $\phi(\sigma) : [0,1] \to [0,1]$ that is weakly decreasing with $\phi(0) = 1/N$ and $\phi(1) = 0$. The call the argument σ the* ***magnitude*** *of the restraint.*

A restraint $\phi(\cdot)$ indicates how profits are distributed among the firms at each value of σ. When $\sigma = 0$, firms are not restrained at all and they freely enter the market and compete generating a total industry profits equal $N\pi^N$. With a symmetric restrain, each firm gets the same fraction $1/N$ of the total industry profit. Thus, the condition $\phi(0) = 1/N$ ensures consistency with the unrestrained equilibrium, which gives each firm a common payoff of π^N. That ϕ is decreasing reflects that this restraint constrains only the challengers, not the patentee, thus leaving challengers at a comparative disadvantage. At the maximal magnitude $\sigma = 1$, the challengers are fully excluded, and the patentee earns a full monopoly profit. We assume throughout that ϕ is differentiable on $(0,1)$, although this can be relaxed to a continuity assumption in much of what follows.

In lieu of lump-sum transfers, a settlement is fully characterized by specifying a restraint $\phi(\cdot)$ and a magnitude $\sigma \in [0,1]$. We thus write settlement payoffs of a restrained of

[6] A monopolist may well benefit from asymmetric restrains. In this version of the paper, we focus on symmetric restrains.

magnitude σ, conditional on restrain ϕ as

$$\begin{aligned} S_0(\sigma|\phi) &= [1 - n\phi(\sigma)]\Pi(\sigma) && (6) \\ S_j(\sigma|\phi) &= \phi(\sigma)\Pi(\sigma) && (7) \end{aligned}$$

It is easy to see that $S_0(\cdot|\phi)$ is strictly increasing, meaning that a larger magnitude of the restraint allows the patentee to capture more of the industry surplus. The payoff of challengers, however, it is not always strictly decreasing in σ, meaning that larger magnitude of the restraint could increase a challenger's payoff. Whether or not this is possible will depend on the shape of the restraint.

Definition 2. *Restraint ϕ is* ***monotonic*** *if $S_j(\sigma|\phi)$ is monotonically decreasing in σ.*

As we will see in detail in a later section, some restraints generate a non-monotonic payoff for the challengers. In particular, when the restrain is such that $S_j(\cdot)$ takes an inverted-U shape, challengers *want* to be restrained (at least a little). This happens, for example, when the restraint serves as a profitable commitment device for challengers, or when it makes the patentee's behavior much more accommodating.

Definition 3. *A restraint $\phi(\cdot)$ is* **non-viable** *if there exists a value of θ such that for all $\sigma \in (0, 1)$*

$$S_i(\sigma|\phi) < L_i^c(\theta), \quad \text{for some } i \in I.$$

A non-viable restraint is one that violates the incentive compatibility constraint, regardless on the magnitude of the restraint, for at least one value of θ and at least one of the firms. That is, one of the firms is strictly better off by going to litigation than accepting any settlement for at least one value of θ. Hence, non-viable restraint is one that would not resolve a patent dispute through a settlement agreement for all θ.

Definition 4. *Given a restraint $\phi(\cdot)$, σ is a* **proportional settlement** *if $\sigma = \theta$*

A proportional settlement is one that restrains competition in the same way that litigation would do if litigation costs were zero, this is, $\Pi(\theta) = L_I^0(\theta)$. In contrast, $\sigma > \theta$ ($\sigma < \theta$), the settlement restrains competition too much (too little) relative to the outcome that results from litigation, if litigation costs were zero.

Notice that with positive litigation costs, the total industry surplus from litigation is $L_I^c(\theta) = L_I^0(\theta) - (N + 1)c$. Thus, by avoiding litigation firms increase their aggregate surplus by $(N+1)c$. To isolate the effect of the settlement on the restrain in competition, and separate it from the surplus created by bargaining over the savings on litigation costs, we say the the settlement is proportional when $\Pi(\theta) = L_I^0(\theta)$ rather than $\Pi(\theta) L_I^0(\theta) - 2c$.

As we show later on in the paper, many restraints $\phi(\cdot)$ allow the firms to reach mutually-acceptable settlements with $\sigma > \theta$, even if litigation is costless. To discern whether a given restraint allows for proportional settlements, we need to know the relation between (1) the rate at which S_j falls in σ; and (2) the rate at which industry profits rise in σ.

Definition 5. *The* marginal exclusion rate *of restraint $\phi(\cdot)$ at σ is given by*

$$\psi(\sigma|\phi) = -\frac{\dfrac{\partial S_j(\sigma|\phi)}{\partial \sigma}}{\dfrac{\partial \Pi(\sigma)}{\partial \sigma}},$$

and the average exclusion rate *of $\phi(\cdot)$ at σ is given by*

$$\overline{\psi}(\sigma|\phi) = -\frac{S_j(\sigma|\phi) - S_j(0|\phi)}{\Pi(\sigma) - \Pi(0)}.$$

These exclusion rates measure the relation of the speed at which challengers' payoff decline versus the speed at which total industry profits raises when the magnitude of the constrain increases. The slower the restraint's exclusion rate, the larger industry profits can be without breaking the challengers' individual rationality condition ($S_j \geq L_j^c$). As such, restraints that elicit disproportionately large profits are characterized by slow exclusion rates.[7]

We can use our setting to study a determine the particular restraint associated to an oligopoly environment. For example, suppose we are examining royalties in Cournot oligopoly (as we will do in detail later on). Let $\alpha \in [0, \bar{\alpha}]$ denote the (symmetric) royalty rate imposed on the challengers. The equilibrium payoff of each firm is denoted by $u_i^*(\alpha)$ and the equilibrium industry profits is denoted by $U^*(\alpha) = \sum_i u_i(\alpha)$. The key

[7] Note that $\psi(\cdot|\phi_A) < \psi(\cdot|\phi_B)$ implies $\overline{\psi}(\cdot|\phi_A) < \overline{\psi}(\cdot|\phi_B)$, but not the converse.

assumption is that the equation

$$u_i^*(\alpha) = S_i(\sigma|\phi) \tag{8}$$

defines an implicit function $\alpha(\sigma)$. In that case, we obtain

$$\frac{\partial S_i(\sigma|\phi)}{\partial\sigma} = \frac{\partial u_i(\alpha(\sigma))}{\partial\sigma} = \frac{\partial u_i(\alpha(\sigma))}{\partial\alpha} \cdot \frac{\partial\alpha(\sigma))}{\partial\sigma}, \quad \text{for all } i \in I.$$

From these equations plus the fact that $\sum_{i\in I} S_j(\sigma|\phi) = \Pi(\sigma)$, we can back out the restrain $\phi(\cdot)$ associated with this particular oligopoly environment. In the next section, we present several examples of settlement agreements and we back out their respective restraints $\phi(\cdot)$.

2.2 Proportional Restraints

Some restraints naturally elicit proportional settlements when firms negotiate in the absence of litigation costs. We are interested in this kind of restraints because firms would jointly appropriate exactly the same profits that they would do if litigation was costless. Formally,

Definition 6. *A restraint ϕ is **perfectly proportional** if, for any θ, it satisfies*

$$S_i(\sigma|\phi) \geq L_i^0(\theta) \quad \text{for all } i \in I \quad \Longleftrightarrow \quad \sigma = \theta$$

A perfectly proportional restraint is one that, in an environment where litigation costs are zero, the only incentive-compatible agreements is one such the magnitude of the restraint $\sigma = \theta$. As we mentioned before, when litigation costs are positive, firms negotiate over the costs savings created by avoiding litigation, which would enlarge the set of incentive compatible agreements.[8] However, litigation costs only introduces rent allocation that has nothing to do with the restrain of competition, but with the firms' bargaining power. For this reason, given that our interest is by how much competition is restrained by the way firms settle, our definition of perfect proportionality allow us

[8] Even more, depending the firm firms' bargaining power, an incentive compatible agreement could involve $S_i(\sigma|\phi) < L_i^0(\theta)$ for some firm $i \in I$. For each θ, the measure of the bargaining core is increasing in c. When $c \to 0$ the bargaining core converges to $L^0(\theta)$.

to separate the effects of bargaining over savings of litigation costs and the extent to the restraint of competition.

There are different kinds of restraints that are perfectly proportional in the way described above, as illustrated in the two examples below. But all such restraints are represented by the same restraint function, ϕ^*. This is formalized in the proposition below.

Proposition 1. *There exists a unique perfectly proportional restraint function, ϕ^*, and it is defined by*

$$\phi^*(\sigma) = \frac{(1-\sigma)\pi^N}{\Pi(\sigma)} \tag{9}$$

Proof. First, it is easy to check that ϕ^* is in fact a perfectly proportional restraint. If $\sigma = \theta$, by definition $S_i(\sigma|\phi^*) = L_i^0(\sigma)$ for all $i \in I$. If $S_i(\sigma|\phi^*) \geq L_i^0(\theta)$ for all $i \in I$, then $S_0(\sigma|\phi^*) \geq L_0^0(\theta)$ implies $\sigma \geq \theta$ and $S_j(\sigma|\phi^*) \geq L_j^0(\theta)$ implies $\sigma \leq \theta$, so these together implies $\sigma = \theta$. Second, suppose there exists $\phi^{**} \neq \phi^*$ such that ϕ^{**} is also perfectly proportional. Then there must exist some θ and σ such that $S_i(\theta|\phi^{**}) \geq L_i^0(\theta)$ for all $i \in I$ with strict inequality for some $k \in I$. Then $\sum_{i\in I} S_i(\theta|\phi^{**}) > L_I^0(\theta)$. But this is impossible because, for any ϕ, we have $\sum_{i\in I} S_j(\sigma|\phi) = \Pi(\sigma) = L_I^0(\sigma)$ for any σ. □

Thus, the perfectly proportional restraint is that which uniquely satisfies $S_i(\cdot|\phi^*) = L_i^0(\cdot)$ for all $i \in I$. With a perfectly proportional restraint in a world where litigation costs were zero, each firm would be exactly indifferent between going to litigation or settling according an incentive compatible agreement consisting on a proportional restrain. Even more, the restrain on competition after the settlement is the same that firms would have created on expectation by going to litigation.

Corollary 1. *The marginal exclusion rate of the perfectly proportional restraint $\phi^*(\cdot)$ is constant and equal to*

$$\psi^* = \frac{\pi^N}{\mu}.$$

Comparing the marginal exclusion rate of a restraint with ψ^*, we can see whether there is too much restrain on competition

Corollary 2. *We have:*

$$\psi(\sigma|\phi) \leq \psi^* \text{ for all } \sigma \Rightarrow \phi(\cdot) \geq \phi^*(\cdot)$$

Proof. $\psi(\sigma|\phi) \leq \psi^*$ is equivalent to $S_j'(\sigma|\phi) \geq -\pi^N$. Integrating between 0 and σ, using that $S_j(0|\phi) = \pi^N$, and using the definition of S_j and ϕ^* we get the result. □

2.2.1 Examples of Proportional Restraints

In this section we present examples of commonly used restraints used to resolve patent disputes which are in fact perfectly proportional. We start with an oligopoly model under some restraint and we use equation (8) to back out the associated restraint $\phi(\cdot)$ for each case.

Example 1 (Pure Delay): The remaining time of patent protection is normalized to $T = 1$. The restraint $\phi_{PD}(\cdot)$ consists in that the challengers agree to delay their entry until time $\alpha \in [0, 1]$, but the patentee cannot make reverse payments to persuade challengers to accept a longer delay period. Let $u_i(\alpha)$ denote firm i's settlement payoff as a function of α, and let $U(\alpha) = \sum_i u_i(\alpha)$. We assume there is no inter-temporal discounting. We have

$$\begin{aligned} u_0(\alpha) &= \alpha\pi^m + (1-\alpha)\pi^N \\ u_j(\alpha) &= (1-\alpha)\pi^N \end{aligned}$$

Notice that, by definition $u_i(\alpha) = L_i^0(\alpha)$. Also, notice that $U(\alpha) = \Pi(\alpha)$. Hence, the equation $U(\alpha) = \Pi(\sigma)$ defines the implicit function $\alpha(\sigma) = \sigma$. Given that for all $i \in I$ we then have $S_i(\sigma|\phi_{PD}) = u_i(\sigma) = L_i^0(\sigma)$ and therefore, according to Proposition 1 we have $\phi_{PD} = \phi^*$.

Our model of pure delay does not account for the various statutory complications that arise in the pharmaceutical industry, which is the environment best-known for delayed-entry settlements. To the extent that the Hatch-Waxman Act or other statutes undermine incentives to challenge,[9] settlements will elicit disproportionately large profits.

[9] See for example Olson and Wendling (2013) or Hemphill (2006).

Example 2 (Territorial Restrictions): Consider a territorial restraint that puts limits on where the challengers can operate but places no territorial limitations on the patentee. Suppose there is a continuum of territories $t \in [0,1]$. We can model this territorial restraint as a limit t such that the patentee excludes challengers from operating in $[0,\alpha]$ and places no restrictions on $(\alpha,1]$. This model is therefore equivalent to pure delay in Example 1, and therefore is a perfectly proportional restrain. Of course, if the patentee agreed to stay out of $(\alpha,1]$, then this would not be equivalent to pure delay (it would be ordinary market division), and the results would not be proportional.[10] This reflects a more general point, briefly addressed in a later subsection, which is that allowing settlements to restrain *the patentee* in addition to the challengers will invariably lead to disproportionate results.

Example 3 (Royalties in Linear Cournot): Competition is Cournot with costless production and inverse demand $p = 1 - Q_I$, where $Q_I = \sum_{i \in I} q_i$ is total output. This yields $\pi^N = (N+1)^{-2}$ and $\pi^m = 1/4$. Now consider the following restraint $\phi_C(\cdot)$: firm 0 imposes a per-unit royalty of $\alpha \geq 0$ on each challenger otherwise they cannot produce. In the Appendix we show that as a function of α, profits from the royalty settlement are

$$\begin{aligned} u_0(\alpha) &= \left(\frac{1+n\alpha}{N+1}\right)^2 + \left(\frac{1-2\alpha}{N+1}\right)n\alpha \\ u_j(\alpha) &= \left(\frac{1-2\alpha}{N+1}\right)^2 \\ U(\alpha) \equiv \sum_i u_i(\alpha) &= \left(\frac{1+n\alpha}{N+1}\right)\left(\frac{N-n\alpha}{N+1}\right) \end{aligned}$$

The function $U(\alpha)$ is strictly increasing and concave for the range $\alpha \in [0, 0.5]$ which is the relevant range for the restraint parameter. We define $\alpha(\sigma)$ implicitly as the solution to $U(\alpha(\sigma)) = \Pi(\sigma)$. Then the marginal exclusion rate is

$$\psi(\sigma|\phi) = \frac{-u_j'(\alpha(\sigma))}{U'(\alpha(\sigma))} = \frac{4}{n^2} = \frac{\pi^N}{\mu} = \psi^*$$

Then, by Corollary 2, the restrain associated to royalties in this example is the perfectly proportional restraint $\phi_C = \phi^*$.

[10] Similarly, if the challengers were separated *from each other* within $[\alpha, 1]$, so that each j is the sole challenger in some sub-interval of length $(1-\alpha)/n$, then we would again get disproportionate results.

Royalties are not proportional in all models of competition, however. In some other contexts (including linear ones), royalties can provide excessive profits by inducing an "accommodating shift" in the patentee's reaction function.[11]

2.3 Over-Proportional Restraints

Some restraints always allow for excessive industry profits through settlement. We refer to these as over-proportional restraints. Formally,

Definition 7. *A restraint ϕ is* ***over-proportional*** *if, for any $\theta \in (0,1)$, there exists some $\sigma > \theta$ such that $S_i(\sigma|\phi) \geq L_i^0(\theta)$ for all $i \in I$.*

The idea of restraints that are over-proportional is that all firms agree on a magnitude of restraint that is strictly larger than θ and all firms are weakly better-off.

Given that $S_0(\cdot|\phi)$ is strictly increasing, when $S_j(\cdot|\phi)$ is strictly decreasing the patentee would like to increase σ but the challengers want to decrease it. Therefore, there is a conflict of interest in the preferences of the patentee and the challengers over the magnitude of the restraint. When $\phi(\cdot)$ is over-proportional, the set of parameters σ such that the patentee and the challengers satisfy their incentive constraint is not a singleton which creates a tension over the magnitude of the restraint.

Should we change the indexes here? It would be more natural to label $\widetilde{\sigma}_0(\theta|\phi)$ the most preferred constraint for the patent holder instead of $\widetilde{\sigma}_j(\theta|\phi)$

Definition 8. *The preferred magnitude of restraint for the patent holder is*

$$\widetilde{\sigma}_j(\theta|\phi) = \max\{\sigma \,:\, S_j(\sigma|\phi) \geq L_j^0(\theta)\},$$

whereas the preferred magnitude of restraint a challenger is

$$\widetilde{\sigma}_0(\theta|\phi) = S_0^{-1}(L_0^0(\theta)|\phi).$$

[11] In particular, if the underlying oligopoly model takes price as the choice variable, then the patentee will internalize its impact on its own royalty receipts. This makes it systematically more accommodating, which is reflected in a shift in its reaction function.

With a perfectly proportional restrain, we have $\widetilde{\sigma}_j(\theta|\phi) = \widetilde{\sigma}_0(\theta|\phi) = \theta$. With an over-proportional restraint we have $\min\{\widetilde{\sigma}_0(\theta|\phi), \widetilde{\sigma}_j(\theta|\phi)\} > \theta$. Even more, we have the following result:

Proposition 2. *The following statements are equivalent:*

(i) ϕ *is over-proportional.*

(ii) $\overline{\psi}(\sigma|\phi) < \psi^*$ *for all* $\sigma \in (0, 1)$.

(iii) $\phi(\sigma) > \phi^*(\sigma)$ *for all* $\sigma \in (0, 1)$.

(iv) $\theta < \widetilde{\sigma}_0(\theta|\phi) < \widetilde{\sigma}_j(\theta|\phi)$ *for all* $\theta \in (0, 1)$.

Condition (*iv*) highlights a particularly important property of over-proportional restraints: unless litigation costs are sufficiently high, the parties will never settle on $\sigma = \theta$, because the patentee would rather litigate. An over-proportional restraint will usually[12] generate a non-singleton bargaining core of σ-values (this is always true when $c = 0$), but this tells us that $\sigma = \theta$ will not lie in the core unless c is sufficiently large. Even if litigation costs are high, the analysis also suggests that, if the patentee has more bargaining power than challengers (in the sense that challengers will settle for payoffs between $L_j^c(\theta)$ and $L_j^0(\theta)$), then settlements will not be proportional. The next example illustrates a settlement that is monotonic, but over-proportional.

Example 3 (Output Cap in Linear Cournot Duopoly): Consider the same linear Cournot duopoly from the last example, but with $N = 2$, so that $\pi^N = 1/9$. Firm 0 imposes a cap on firm 1's output. This is binding iff the cap is below 1/3, so the firms agree on some $\alpha \in [0, \frac{1}{3}]$ such that firm 1 is constrained to set $q_1 = \frac{1}{3} - \alpha$. This results

[12] If $c > 0$ and θ is sufficiently close to either 0 or 1, then either the patentee or challengers will lack a credible threat to litigate, and this can lead the bargaining core to be a singleton, since the other side can then dictate its preferred terms.

in the following payoffs as a function of α:

$$\begin{aligned} u_0(\alpha) &= \left(\frac{2+3\alpha}{6}\right)^2 \\ u_1(\alpha) &= \left(\frac{2+3\alpha}{6}\right)\left(\frac{2-6\alpha}{6}\right) \\ U(\alpha) &= \left(\frac{2+3\alpha}{6}\right)\left(\frac{4-3\alpha}{6}\right) \end{aligned}$$

Define α_σ by $U(\alpha_\sigma) = \Pi(\sigma)$. The average exclusion rate is:

$$\overline{\psi}(\sigma|\phi) = \frac{u_1(0) - u_1(\alpha_\sigma)}{U(\alpha_\sigma) - U(0)} = \frac{6 + 18\alpha_\sigma}{6 - 9\alpha_\sigma}$$

which is strictly below $\psi^* = 4$ for all $\alpha < \frac{1}{3}$. Hence the output restraint is over-proportional.

We can use conditions resembling those in Proposition 2 in order to rank different restraints in terms of how "restrictive" they are, i.e. how far their settlements will tend to deviate from proportional levels. Formally:

Definition 9. *ϕ_A is* ***more restrictive*** *than ϕ_B if $\phi_A \neq \phi_B$ and any of the following conditions are satisfied:*

(i) $\overline{\psi}(\cdot|\phi_A) \leq \overline{\psi}(\cdot|\phi_B)$.

(ii) $\phi_A \geq \phi_B$.

(iii) $\widetilde{\sigma}_i(\cdot|\phi_A) \geq \widetilde{\sigma}_i(\cdot|\phi_B)$ *for all* $i \in I$.

Thus an over-proportional settlement is one that is strictly more restrictive than ϕ^*.

Proposition 3. *If $S_j(\cdot|\phi)$ is strictly concave, then ϕ is over-proportional.*

Proof. Note that all restraints satisfy $S_j(0|\phi) = \pi^N$ and $S_j(1|\phi) = 0$, implying $\overline{\psi}(1|\phi) = \psi^*$. Strict concavity implies $\overline{\psi}(\cdot|\phi)$ is strictly increasing. Then, given $\overline{\psi}(1|\phi) = \psi^*$, this implies $\overline{\psi}(\sigma|\alpha) < \psi^*$ for all $\sigma \in (0, 1)$. Then, by Proposition 2, ϕ is over-proportional. □

Note, however, that this proposition deals with the behavior of S_j with respect to σ, which is generally not identical to the behavior of u_j with respect to α.

2.4 Nonmonotonic Restraints

Strictly monotonic restraints have a socially desirable property: lower values of θ will always lead to less restrictive settlements.[13] So even when the restraint is over-proportional, this kind of restraint is not 'too bad' as long as it is strictly monotonic. As such, strictly monotonic restraints are the next best option after perfect proportionality and they may even provide a close approximation. But some restraints are nonmonotonic, at least under certain market conditions, and these have fewer redeeming qualities. When a restraint is nonmonotonic, the challengers sometimes prefer larger σ-values to lower ones, so that the latter will never be selected through bargaining, no matter the value of θ. Of particular interest are those nonmonotonic restraints that can give challengers more profits than the unrestrained equilibrium.

Definition 10. *A restraint ϕ is* ***accommodating*** *if* $\max_\sigma S_j(\sigma|\phi) > \pi^N$.

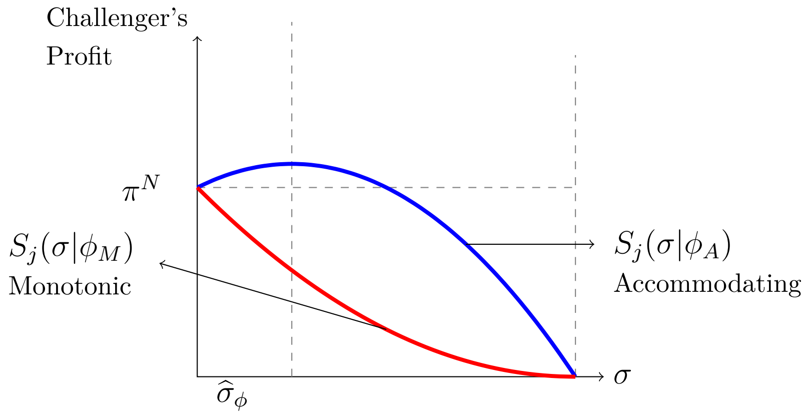

Figure 2: CHANGE THIS: ADD THE PERFECT RESTRAINT (WHICH GIVES LINEAR PAYOFF) + MONOTONIC BUT OVER-PROPORTIONAL RESTRAINT (WHICH IS DECREASING BUT CONCAVE AND ABOVE THE PERFECT RESTRAINT + ACCOMMODATING RESTRAINT, WHICH IS AS SHOWN IN CURRENT GRAPH. THE RED LINE IN CURRENT GRAPH IS ACTUALLY A NON-VIABLE RESTRAINT...

[13] More precisely, as a function of θ, the bargaining core is strictly decreasing in the strong set order.

Figure 2 shows two classes of restraints. ϕ_A is an accommodating restraint. In this case, when the magnitude of the restraint is sufficiently low, settling on those magnitudes is Pareto inferior. Both challengers and the patentee are strictly better off by increasing the magnitude of the restraint. In Figure 2, every $\sigma < \widehat{\sigma}_\phi$ is Pareto-dominated by, for example, $\widehat{\sigma}_\phi$.

A restraint can be accommodating if it allows the challengers to engage in profitable commitment, or if it makes the patentee's behavior much less aggressive. Output caps and price floors (more generally, restraints that truncate the action set) can be accommodating, as can some other kinds of restraints. By contrast, royalties will be monotonic (and thus non-accommodating) unless the underlying model of oligopoly has heroically strong strategic effects. Restraints that merely limit when or where challengers can enter – as with pure delay or territorial restraints – are always monotonic.

In all examples of accommodating restraints that we have identified, $S_j(\cdot|\phi)$ is strictly concave, giving it an inverted-U shape. The previous proposition thus ensures that these restraints are also over-proportional.[14]

Consider the example of output caps (which are akin to price floors). If $n = 1$, an output cap can be nonmonotonic only for "Stackelberg reasons." This requires that, in the underlying duopoly model, an optimal commitment involves less competitive behavior than in the Nash equilibrium.[15] This explains why the output cap in Example 3 was not accommodating: a Cournot duopolist wants to commit to a *higher* output level, not a lower one. But when there are multiple challengers, there is a simpler reason why a restraint can be accommodating. A market becomes more competitive as the number of firms increases, but an output restraint can counteract this by making the set of challengers produce a joint-output that corresponds to a smaller set of unrestrained competitors. This can benefit the challengers themselves. This is illustrated in the example below, which is just the Cournot game from Example 3 extended to the case of $n \geq 3$.

[14] If S_j is not strictly concave, then $\overline{\psi}(\cdot|\phi)$ is not necessarily strictly increasing, and we cannot rule out the possibility that the restraint it will elicit proportional settlements when θ is close to 1.

[15] For example, if the underlying model of oligopoly involves price competition with linear demand and $n = 1$, then it is easy to show that a single challenger would like to commit to a binding output cap, even if $\theta = 0$.

Example 4 (Output Cap in Linear Cournot with $n \geq 3$): Consider again the Cournot model with inverse demand $1 - \sum q_i$ and costless production, but now we let $n \geq 3$. Analogous to Example 3, the firms agree on a value $\alpha \in [0, \frac{1}{n+2}]$ such that each j is constrained to set $q_j = \frac{1}{n+2} - \alpha$. As a function of α resulting equilibrium gives challengers a profit of

$$\begin{aligned} u_j(\alpha) &= \left(\frac{1-(n+2)\alpha}{n+2}\right)\left(\frac{2+n(n+2)\alpha}{2(n+2)}\right) \\ \implies \quad u_j'(\alpha) &= \frac{1}{2(n+2)^2}\Bigg[n^2 - 4 - 2n(n+2)\alpha\Bigg] \end{aligned}$$

Thus $n \geq 3$ implies that $u_j(\alpha)$ is strictly increasing at low α-values, and ultimately takes an inverted-U shape.

Generalizing from this example, it is intuitively clear that if an accommodating restraint is just a truncation of the challengers' strategy set, then the challengers' favorite σ-level corresponds to a Stackelberg-duopoly equilibrium (with the challengers jointly acting as leader). Even if a restraint does not shrink the challengers' strategy set, it can be accommodating if it makes the patentee behave much less aggressively. This is illustrated in the example below, which contemplates a transaction in which the patentee is compensated by acquiring a portion of stock in each challenger.

Example 5 (Stock Acquisition): Consider the same Cournot model as in the last example, but instead of an output cap, the patentee acquires a fraction α of each challenger's outstanding stock. Thus firm 0's payoff function is $(1 - Q_I)(q_0 + \alpha Q_J)$ and each j's payoff function is $(1 - \alpha)(1 - Q_I)q_j$, where $Q_I \equiv \sum_i q_i$ and $Q_J = \sum_{j \in J} q_j$. As a function of α, the equilibrium gives each challenger a profit of

$$\begin{aligned} u_j(\alpha) &= \frac{1-\alpha}{(n+2-n\alpha)^2} \\ \implies \quad u_j'(\alpha) &= \frac{n^2 - 4 - 2n^2\alpha + n^2\alpha^2}{(n+2-n\alpha)^4} \end{aligned}$$

Thus, each challenger's profits are strictly concave and, if $n \geq 3$, they are initially increasing in α. Note that if $n = 1$ or $n = 2$, then this restraint is not accommodating, but it is still over-proportional.

2.5 Transfers and Reverse Payment

We have so far focused on settlements with a restraint alone, but in practice the firms may include lump sum transfers. A common example is licensing financed through a two-part tariff, with all challengers paying an upfront fee in addition to a running royalty. Conditional on σ, transfers do not affect industry profits, but a transfer does affect bargaining over σ-values. To illustrate, we define $\widetilde{\sigma}_i^+$ as an analogue to $\widetilde{\sigma}_i$ for the case where settlement also includes a transfer payment τ (which can be negative) from each challenger to the patentee. Explicitly:

$$\begin{aligned} \widetilde{\sigma}_0^+(\theta, \tau|\phi) &= S_0^{-1}(L_0^0(\theta) - n\tau|\phi) \\ \widetilde{\sigma}_j^+(\theta, \tau|\phi) &= \max\Big\{\, \sigma \mid S_j(\sigma|\phi) - \tau \geq L_j^0(\theta)\Big\} \end{aligned}$$

Intuitively, a larger value of τ means that challengers are not as willing to accept as strong a restraint, but also that the patentee is willing to accept a weaker restraint. The following result is immediate:

Proposition 4. *For any ϕ and $\theta \in (0,1)$, $\widetilde{\sigma}_0^+(\theta, \tau|\phi)$ and $\widetilde{\sigma}_j^+(\theta, \tau|\phi)$ are strictly decreasing in τ whenever they are interior.*

[**Are we assuming $S_j' < 0$ here?**]

Intuitively, royalties are a little like the combination of an output cap and a transfer. In fact, unpacking this point helps to clarify that output caps are always more restrictive than royalties.

Remark 2 (Royalties are Less Restrictive than Output Caps): A royalty increases industry profits only by diminishing challengers' output; beyond that, the payments just reallocate profits from challengers to the patentee. As such, output caps must be more restrictive than royalties within any oligopoly environment. To illustrate, let ϕ_r be a royalty restraint function. Then there exists an output cap restraint function ϕ_{oc} with the following property: for any σ, there exists some transfer $\tau_\sigma > 0$

such that

$$\phi_r(\sigma)\Pi(\sigma) = \phi_{oc}(\sigma)\Pi(\sigma) - \tau_\sigma$$

This implies that σ_r lies below ϕ_{oc}, so that ϕ_{oc} is more restrictive than ϕ_r.

2.5.1 Reverse Payments

A reverse payment is a negative transfer, $\tau < 0$. Following Proposition 4, a negative transfer shifts the bargaining core to higher σ-levels. This enhances total profits and undermines proportionality. In fact, the condition $\pi^m > N\pi^N$ ensures that, if transfers were allowed to be arbitrarily negative, then the firms would always set $\sigma = 1$ and adjust τ so that each i earns at least $L_i^c(\theta)$. Thus, the firms' bargaining reduces to simply choosing the size of the reverse payment. This completely eliminates the desired positive correlation between θ and σ. And, importantly, this is true regardless of what kind of restraint is accompanied by the reverse payment.

Proposition 5. *Suppose that firms can set $\tau < 0$. Then for any ϕ, every settlement (σ, τ) with $\sigma < 1$ is Pareto-dominated.*

This clarifies the function of (and impetus for) reverse payments: to subvert proportionality. A reverse payment can accompany any kind of restraint, and it always enables the firms to capture disproportionately large profits. As such, it is a mistake regard the terms "reverse payment" and "pay-for-delay" as synonymous, as is common in the literature. First, as we have shown, the underlying restraint in pay for delay (delayed entry) can be highly proportional when it is not augmented by reverse payment, so it would be a mistake to assume that restraints on the timing of entry are *inherently* problematic. Second, a reverse payment could just as easily accompany any other kind of restraint, and the results will still be bad. Third, a reverse payment is not itself a restraint; it is just a lump sum payment, while pay for delay is the combination of an entry restraint and a payment.

2.6 All Firms Restrained

The last subsection identifies reverse payment as something that systematically undermines proportionality, suggesting that such payments are highly problematic. But this is not the only thing firms might rely on to overcome proportionality. Just as problematic are settlements that, in addition to restraining challengers, impose a countervailing restraint on the patentee. This provides the basic ingredient that supports a cartel, which is that all firms in the market are restrained. Once this is satisfied, all of the firms are happy to elevate the magnitude of the restraint, for this increases total profits without interfering with the distribution. In our model, this would be like a restraint function that satisfies $\sigma(\sigma) = 1/N$ for all σ, ensuring that S_j is monotonically increasing in σ.

For example, it is well known in the economic literature on patent licensing that rivals can indirectly cartelize a market by charging each other royalties that are fine-tuned to induce cartel price levels. But when only one of them pays a royalty, the settlement will be proportional, as illustrated in Example 2. Similarly, we noted earlier that a territorial restraint is proportional if it constrains where the challengers can operate but lets the patentee operate everywhere, whereas the restraint turns into ordinary market division if the patentee is itself constrained to stay out of the challengers' turf. Indeed it is easy to see that we can take any restraint and impose a countervailing version on the patentee, and the result will always be that challengers are subsequently happy to agree to higher values of σ.

In some sense, a countervailing restraint is like a reverse payment in that it is a side-deal intended to persuade the challengers to accept a higher level of σ. The difference is that, while a reverse payment is just cash, a countervailing restraint is effectively a promise by the patentee that it will behave less competitively that it would otherwise.

In practice, many horizontal settlements involve several patentees, each wanting to use the others patents. In this case would expect that the patents are themselves complementary, although the firms sell competing products. It is well understood that patent pooling is an efficient way to license complementary, separately-held patents, since it eliminates a double marginalization problem. Alternatively, the firms might just cross-license with each other. But in any case, problems will arise if competing patent holders construct an arrangement in which all of them are restrained in parallel, at least if they jointly maintain a dominant market position.

Nor are countervailing restraints necessary in an agreement between rival patent holders, for the scenario must take one of two mutually-exclusive forms: either (1) a subset of firms contributes more to the others than its constituents receive in turn; (2) each firm contributes about the same amount. In case (1), the subset of firms who contribute more need not be restrained (although they may impose restraints on the others), for they are already providing more value to the others than they are getting back. In case (2), no firms need be restrained, for the exchange of technology rights does not leave anyone at a comparative disadvantage.

2.7 Non-Viability

If a restraint's exclusion rate is too fast, then it may be *non-viable* in the sense that, for some values of θ, no σ-level is mutually-preferred to litigation. This means that $\widetilde{\sigma}_j(\theta|\phi) < \widetilde{\sigma}_0(\theta|\phi)$.[16] An important example is that, if price competition is aggressive – that is, if cross-price elasticity is quite high – then royalties may not be viable, because challengers' profits will fall too quickly in relation to the growth in industry profits. More generally, a royalty inherently leave challengers at a cost disadvantage, leaving them vulnerable to being undercut by the patentee,[17] who has lower costs. As a result, it is easy to see that royalties are non-viable if competition is Bertrand with homogeneous products.

This is likely why we very rarely see royalties used in horizontal settlements in pharmaceutical markets. Instead, the firms tend to rely on delayed entry (which may or may not be accompanied by a reverse payment). In these settlements, the patentee sells an expensive brand-name drug, while the challengers want to sell a generic version. The brand-name drug and the generic versions are essentially fungible, so price competition can be quite aggressive. Empirical evidence suggests that prices remain significantly above cost when just one or two generics have entered, but they eventually fall substan-

[16] By expanding the bargaining core, positive litigation costs might be able to overcome this, provided they are sufficiently large.

[17] In principle the firms could write a contract saying the patentee will not undercut the challengers, but this would be a countervailing restraint on the patentee. And, as we have already argued, this will ultimatley lead to disproportionate results, for it will influence the choice of the royalty rate.

tially as generic entry expands (Olson and Wendling, 2013). The initial maintenance of high prices likely represents some degree of "conscious parallelism" or tacit collusion, which is ultimately supported by each firm's knowledge that, if it cuts price, the others can easily follow. If the challengers must pay high royalties, this condition is not present: the challengers could not match the patentee's price cut, and the patentee knows it. It is thus unsurprising that such firms do not like to rely on royalties.

More generally, there may be market-specific reasons a given kind of restraint may not be feasible or realistic within a particular case. By analogy, we know that, in a vertical licensing relationship, joint profits are maximized by financing the deal with lump sums alone, as this avoid double marginalization. But in practice the large majority of agreements involve royalties. The point is that a contract may be easy to theorize but very challenging to implement in practice. For example, we know that pay-for-delay cases arise in a dense fog of complicated (and sometimes ill-conceived) statutes and regulations. These complications, which this paper has largely ignored, must be factored in to discern whether a particular kind of restraint is reasonably viable within a given market.

3 Incentives to Invent

REVISIONS BY ERIK:

The patentability laws assign θ-levels to inventions, with the goal being to strike the optimal balance between static competition and the rate of innovation. Perfectly proportional restraints will ensure that this optimal balance is preserved through settlement. In this section we show two further points. First, if a restraint is over-proportional but monotonic, then the incentive to innovate increases by a bit, but competition is systematically lower, relative to the ideal balance of competition and innovation. This is not the first-best, but at least it has the property that innovation increases somewhat. By contrast, if firms are allowed to rely on accommodating restraints—or to use reverse payments or countervailing restraints—then the results are much worse: both competitiveness and the rate of innovation decline.

Let $D(\theta)$ denote the cost of develop an invention whose patent has quality θ.

We assume that D is differentiable, strictly increasing, and strictly convex with $D(0) = 0$. In reality we would of course expect that a higher θ also accrues larger profits, but we assume this profit-effect of θ is invariant in the restraint chosen, and thus we can ignore it when doing comparative statics on the type of restraint being used.

To examine incentives to innovate, we first need to specify the mapping between θ and σ for a given restraint, which depends on bargaining power. Given θ, ϕ, and a bargaining parameter $\beta \in [0, 1]$, the function $\widehat{\sigma}_\beta(\theta|\phi)$ gives the value of θ reached through bargaining. This is defined explicitly by

$$\widehat{\sigma}_\beta(\theta|\phi) = \beta\widetilde{\sigma}_0(\theta|\phi) + (1 - \beta)\widetilde{\sigma}_j(\theta|\phi) \tag{10}$$

We thus interpret β as the patentee's bargaining power, while each challenger's bargaining power is $1 - \beta$. With this, we can easily model the patentee's ex ante decision problem of deciding how much to invest in innovation, conditional on ϕ and β. It is:

$$\max_\theta \;\; S_0(\widehat{\sigma}_\beta(\theta|\phi)|\phi) - D(\theta)$$

and thus an interior solution is characterized by the first order condition

$$S_0'(\widehat{\sigma}_\beta(\theta|\phi)|\phi)\widehat{\sigma}_\beta'(\theta|\phi) = D'(\theta) \tag{11}$$

We can thus focus mainly on the lefthand side of the above equation—the patentee's marginal profits as a function of θ—to discern the strength of the incentive to innovate. To form a baseline, we begin by determining the incentive to innovate when a restraint is perfectly proportional.

Part 1: perfectly proportional restraint.

We have $\widehat{\sigma}_\beta(\theta|\phi^*) = \theta$, and it's easy to show that $S_0'(\widehat{\sigma}_\beta(\theta|\phi)|\phi)\widehat{\sigma}_\beta'(\theta|\phi) = \pi^m - \pi^n$. Denote the resulting choice of θ by θ^*.

Part 2: a strictly monotonic but over-proportional restraint.

Let ϕ be strictly monotonic and over-proportional. We assume that $S_j(\cdot|\phi)$ is strictly concave, implying ψ is strictly increasing. By definition we have

$$S_0(\widetilde{\sigma}_0(\theta|\phi)|\phi) = L_0^0(\theta) = \pi^N + (\pi^m - \pi^N)\theta \tag{12}$$

$$S_j(\widetilde{\sigma}_j(\theta|\phi)|\phi) = L_j^0(\theta) = (1-\theta)\pi^N \tag{13}$$

Differentiating with respect to θ and rearranging yields

$$\widetilde{\sigma}_0'(\theta|\phi) = \frac{\pi^m - \pi^N}{S_0'(\widetilde{\sigma}_0|\phi)} = \frac{\pi^m - \pi^N}{\mu[1 + n\psi(\widetilde{\sigma}_0|\phi)]} \tag{14}$$

$$\widetilde{\sigma}_j'(\theta|\phi) = \frac{-\pi^N}{S_j'(\widetilde{\sigma}_j|\phi)} = \frac{\pi^N}{\mu\psi(\widetilde{\sigma}_j)} \tag{15}$$

And obviously $\widehat{\sigma}_\beta' = \beta\widetilde{\sigma}_0' + (1-\beta)\widetilde{\sigma}_j'$. Then, using $S_0'(\sigma|\phi) = \mu[1 + n\psi(\sigma|\phi)]$, we obtain

$$S_0'(\widehat{\sigma}_\beta(\theta|\phi)|\phi)\widehat{\sigma}_\beta'(\theta) = \beta(\pi^m - \pi^N)\left(\frac{1 + n\psi(\widehat{\sigma}_\beta)}{1 + n\psi(\widetilde{\sigma}_0)}\right) + (1-\beta)\pi^N\left(\frac{1 + n\psi(\widehat{\sigma}_\beta)}{\psi(\widetilde{\sigma}_j)}\right) \tag{16}$$

which seems to be strictly larger than it would be under ϕ^* (in which case we'd have $\widetilde{\sigma}_0 = \widetilde{\sigma}_j = \widehat{\sigma}_\beta$), given that ψ is strictly increasing. This is clearly true at low values of θ. At the high range we just need to make sure that ψ is not too large (i.e. that S_j isn't falling *too* fast).

Part 3: Accommodating restraints, reverse payments, and countervailing restraints

For reverse payments and countervailing restraints, we have already shown that we'll have $\sigma = 1$ for all θ, although the challengers will get a larger piece of the total profit (π^m) when θ is larger. Show that patentee's payoff function is linear with a left-intercept strictly above π^N. It must therefore have a lower slope than it would under ϕ^*. That means incentive to innovate goes down strictly.

For accommodating, just show there is a huge range of pareto-dominated σ-levels. So the incentive to innovate must be lower if θ^*. But it could be that incentive to invent is stronger when θ^* is large (since S_0 is rising very rapidly); I'm not totally sure right now.

In this section we study how different restraints shape innovation incentives. Suppose that to obtain a patent of quality θ a firm must pay a cost $c(\theta)$ where $c(\theta)$ is increasing and convex.

Perfectly Proportional Restraints

Suppose that only proportional restraints are allowed. For a patent of quality θ, the magnitude of the settlement is $\sigma = \theta$ and also we know that the unique perfectly proportional settlement is $\phi^*(\theta) = \frac{(1-\theta)\pi^N}{\Pi(\theta)}$. Therefore, under a perfectly proportional restraint a patent holder expects to receive

$$\begin{aligned} S_0(\theta|\phi^*) &= (1 - n\phi^*(\theta))\Pi(\theta), \\ &= \pi^N + \theta(\pi^m - \pi^N). \end{aligned}$$

With a perfectly proportional restraint, the marginal incentive to invest is given by $\pi^m - \pi^N$. Then, the optimal level of quality under a proportional restraint satisfy

$$c'(\theta) = \pi^m - \pi^N.$$

In general, suppose firms agree on a settlement under restraint ϕ. The set of feasible agreements is

$$\Omega(\theta) = \{\sigma \in [0,1] \ : \ S_i(\sigma|\phi) \geq L_i^0, \text{ for all } i \in I\}$$

Consider the best agreement for the patent holder in $\Omega(\theta)$, denoted by $\hat{\sigma}(\theta)$. Then, the patent holder gets

$$S_0(\hat{\sigma}(\theta)|\phi) = [1 - n\phi(\hat{\sigma}(\theta))]\Pi(\hat{\sigma}(\theta))$$

Notice that we can write

$$\frac{\partial S_0(\sigma|\phi)}{\partial \sigma} = \mu[1 + n\psi(\sigma|\phi)]$$

Therefore, the incentive to invest for restraint ϕ is smaller than the incentive to invest with a perfectly proportional restraint if

$$\mu[1 + n\psi(\hat{\sigma}(\theta)|\phi)]\hat{\sigma}'(\theta) < \mu[1 + n\psi^*] \Leftrightarrow [1 + n\psi(\hat{\sigma}(\theta)|\phi)]\hat{\sigma}'(\theta) < [1 + n\psi^*]$$

With over-proportional restraints, we have that $\psi(\hat{\sigma}(\theta)) < \phi^*$, so a sufficient condition for this inequality to hold is $\sigma'(\theta) \leq 1$.

4 Antitrust Analysis

How should an antitrust authority decide on what kind of settlements to allow and which one to condemn? One possible avenue that an antitrust could purse is what we call the "comparative estimates" approach:

(a) Estimate the extent to which the settlement has diminished competition relative to patent invalidation.

(b) Estimate the probability that the patent is valid.

(c) Ask whether the estimate in (a) seems commensurate with the estimate in (b).

Suppose that a patent dispute is resolved by settlement that considerably diminishes competition and, even more, there is a strong suspicion that that the underlying patent is likely to be invalidated if re-examined. In this scenario, it would seem reasonable for an antitrust authority to intervene. In theory this is a simple and effective way to administer the antitrust laws. In practice, however, this tends to be difficult or impossible in most cases. For example, PTAB conducts patents re-examinations when based on preliminary evidence there is a reasonable likelihood of ex-post invalidation. In Hovenkamp and Lemus (2016), we document that many pharmaceutical firms settle *after* PTAB has determined that invalidation is reasonably likely. Even more, we find evidence of generic firms delaying their entry to the market many years after the settlement. In the language of or framework, we have a situation in which it is very likely that θ is low, however the restrain on competition is very large (long entry delay).

There are several reasons for why the "comparative statics" approach is difficult to implement in practice. First, estimating the impact of the settlement on competition will be very difficult to do reliably, not least because it requires an estimate of the counterfactual unrestrained equilibrium that was avoided by the settlement. Further, a court is unlikely to be willing to render an antitrust judgment based on its own beliefs about the patent's validity. Judges will demand something more objective—such as economic inference—to support a finding that competition is being restrained too far in light of the patent's quality.

Our proposal, instead, avoids the challenges faced by the comparative estimates approach. It requires neither an estimate of patent quality nor an estimate of the settlement's effect on competition. Rather, it instructs courts to look at *how* the settlement restrains competition, and to rely on economic inference to determine whether that settlement structure is likely to produce a reasonably proportional result. Put differently, we should focus less on the *extent* to which competition is restrained than on the particular *manner* in which it is restrained. Expressed in terms of our model's notation, the comparative estimates approach works by estimating θ and σ and asking if they seem reasonably close, while our approach just looks at ϕ and asks whether it is reasonably proportional, given the circumstances.

Thus, the most important element of our proposal is that antitrust should focus on the way competition is restrained and ask whether it is reasonably justified under the circumstances. For example, we have argued that pure delay is highly proportional, and thus does not inherently create an antitrust concern. But as already noted, an important caveat is that such deals often occur under the auspices of the Hatch-Waxman Act, which undermines the incentive to challenge, this and may undermine proportionality. But this is a problem with the Hatch-Waxman Act, not a problem with pure delay. Our analysis also shows that royalties will tend to produce reasonably proportional settlements, although they may be non-viable when cross-price elasticity is high. Territorial restraints are often highly controversial, but so long as such restrictions are not imposed reciprocally on the patentee, the results will be proportional. To be sure the results are proportional, we can look to whether the patentee continues to operate in the locations where the challengers are permitted entry.

A workable and generally acceptable standard is to allow all restraints that are likely to be strictly monotonic in most settings. By contrast, we should be wary of settlements that are accommodating in the sense that the challengers might benefit from being restrained in parallel, at least a little bit. Thus output caps and price restraints raise legitimate antitrust concerns. In such cases, we should apply antitrust's "less restrictive alternatives" doctrine. This standard is applied to practices or agreements that include some procompetitive features, but also include some features that are reasonably likely to disrupt competition to some extent. If an alternative arrangement would have provided the same procompetitive elements but without the anticompetitive ones, then the deal may be condemned. Following this logic, output caps and price restraints seem not to be reasonably justified if there are some less restrictive restraints that are very likely to be viable for them. But we should not rule out the possibility that an output cap may be reasonably justified in certain market conditions.

Aside from looking at the manner in which challengers are restrained, antitrust should look for ancillary phenomena that systematically undermine proportionality. We have highlighted two such phenomena: reverse payments, and countervailing restraints imposed on patentees. In our view, reverse payments in horizontal settlements are a transparent effort to erect a disproportionately large barrier to competition, and should be illegal per se. Countervailing restraints should also be regarded as highly suspicious and, as we have argued, there seems not to be anything special about patent licensing that makes it necessary for all of the firms to be restrained. As such, countervailing restraints should be evaluated using the same machinery antitrust already used to evaluate horizontal restraints.

5 Conclusion

TBA

6 References

Amir, Rabah, David Encaoua, and Yassine Lefouili (2014) "Optimal licensing of uncertain patents in the shadow of litigation," *Games and Economic Behavior*, Vol. 88, pp. 320–338.

Bessen, James E and Michael J Meurer (2006) "Patent litigation with endogenous disputes," *The American economic review*, pp. 77–81.

Blair, Roger D and Thomas F Cotter (2002) "Are Settlements of Patent Disputes Illegal Per Se?" *The Antitrust Bulletin*, Vol. 47, pp. 491–539.

Daughety, Andrew F and Jennifer F Reinganum (2005) "Economic theories of settlement bargaining," *Annu. Rev. Law Soc. Sci.*, Vol. 1, pp. 35–59.

Hemphill, C Scott (2006) "Paying for delay: Pharmaceutical patent settlement as a regulatory design problem," *NYUL Rev.*, Vol. 81, p. 1553.

Hovenkamp, Erik and Jorge Lemus (2016) "Reverse Settlement and Holdup at the Patent Office."

Lemley, Mark A and Carl Shapiro (2013) "A simple approach to setting reasonable royalties for standard-essential patents," *Berkeley Tech. LJ*, Vol. 28, p. 1135.

Meurer, Michael J (1989) "The settlement of patent litigation," *The RAND Journal of Economics*, pp. 77–91.

Olson, Luke M and Brett W Wendling (2013) "Bureau of Economics Federal TRade Commission Washington, DC 20580."

Shapiro, Carl (2003) "Antitrust limits to patent settlements," *RAND Journal of Economics*, pp. 391–411.

Somaya, Deepak (2003) "Strategic determinants of decisions not to settle patent litigation," *Strategic Management Journal*, Vol. 24, pp. 17–38.

Willig, Robert D and John P Bigelow (2004) "Antitrust policy toward agreements that settle patent litigation," *Antitrust Bull.*, Vol. 49, p. 655.

Luke M. Olson and Brett W. Wendling (2013) FTC paper on generic entry

Appendix

Cournot Royalties

Let q_i be the quantity produced by firm $i \in I$. Let $Q = \sum_{i \in I} q_i$ the total quantity produced, so the market price is $P = 1 - Q$. Let α be the royalty rate set by firm 0. In our symmetric setting, firm 0 sets the same royalty rate to every challenger. Then, firm 0 solves

$$\max_{q_0} \left(1 - q_0 - \sum_{j=1}^{n} q_j\right) q_0 + \alpha \sum_{j=1}^{n} q_j$$

The FOC is

$$q_0 = \frac{1 - \sum_{j=1}^{n} q_j}{2}.$$

Firm j on the other hand solves

$$\max_{q_j} \left(1 - q_0 - q_j - \sum_{\ell \neq j}^{n} q_\ell\right) q_0 - \alpha q_j$$

The FOC is

$$q_j = \frac{1 - q_0 - \sum_{\ell \neq j}^{n} q_\ell - \alpha}{2}, \quad \text{for all } j = 1, ..., n.$$

In a symmetric equilibrium, $q_j = q$ for all $j = 1, ..., n$. Therefore, from the FOC of firm 0 and firm j we obtain

$$q_0^* = \frac{1 + (N-1)\alpha}{N+1}, \quad q_j^* = \frac{1 - 2\alpha}{N+1}.$$

Notice that we require for this equilibrium to be valid that $\alpha \leq 0.5$. When $\alpha > 0.5$ firms will not produce and firm 0 obtains monopoly profits. Suppose $\alpha \leq 0.5$. Then,

$$Q^* = q_0^* + (N-1)q_j^* = \frac{N(1-\alpha) + \alpha}{N+1}, \quad P^* = 1 - Q^* = \frac{1 + (N-1)\alpha}{N+1}.$$

From here, we get total industry profits given by

$$\sum_{i=0}^{n} S_i(\alpha|\phi_C) \equiv \sum_{i=0}^{n} \pi_i(\alpha) = P^* Q^* = \frac{(N(1-\alpha) + \alpha)(1 + (N-1)\alpha)}{(N+1)^2}$$

Proof of Proposition 2:

Proof. First, consider the following identities directly from the definitions

$$\begin{aligned} S_0(\sigma|\phi) &= \pi^N + \mu\sigma[1 + n\overline{\psi}(\sigma|\phi)], \\ S_j(\sigma|\phi) &= \pi^N - \mu\sigma\overline{\psi}(\sigma|\phi). \end{aligned}$$

Recalling that $S_i(\cdot|\phi^*) = L_i^0(\cdot)$, it follows that

$$S_0(\sigma|\phi) \geq L_0^0(\theta) \iff \sigma[1 + n\overline{\psi}(\sigma|\phi)] \geq \theta[1 + n\psi^*], \tag{17}$$

$$S_j(\sigma|\phi) \geq L_j^0(\theta) \iff \sigma\overline{\psi}(\sigma|\phi) \leq \theta\psi^*. \tag{18}$$

(i) ⇒ *(ii)*: Consider $\phi(\cdot)$ over-proportional. Fix $\theta \in (0,1)$. We know there exists some $\sigma > \theta$ such that $S_i(\sigma|\phi) \geq L_i^0(\theta)$ for all i. From identities (17) and (18) we have that, for this particular σ,

$$\sigma[1 + n\overline{\psi}(\sigma|\phi)] \geq \theta[1 + n\psi^*] \quad \text{and} \quad \sigma\overline{\psi}(\sigma|\phi) \leq \theta\psi^*.$$

Given that $S_j' < 0$, we have $\overline{\psi}(\sigma|\phi) \geq 0$. Hence, if $\overline{\psi}(\sigma|\phi) > \psi^*$, then $\sigma\overline{\psi}(\sigma|\phi) \geq \theta\psi^*$ which is a contradiction with the second inequality.

(ii) ⇒ *(i)*: Suppose that $\overline{\psi}(\sigma|\phi) < \psi^*$ for all σ. Then, evaluating at $\sigma = \theta$ we have $\theta\overline{\psi}(\theta|\phi) < \theta\psi^*$. By continuity, this inequality holds for $\sigma = \theta + \varepsilon$, for a small ε. We can increase σ to reach a magnitude $\hat{\sigma}$ such that $\hat{\sigma}\overline{\psi}(\hat{\sigma}|\phi) = \theta\psi^*$. For this magnitude $\hat{\sigma}$ then $\hat{\sigma}[1 + n\overline{\psi}(\hat{\sigma}|\phi)] \geq \theta[1 + n\psi^*] \Leftrightarrow \hat{\sigma} \geq \theta$. Therefore, $\phi(\cdot)$ is over-proportional.

(ii) ⇔ *(iii)*: Directly from the definitions.

(iv) ⇒ *(i)*: Fix $\theta \in (0,1)$. Notice that any $\sigma \geq \min\{\widetilde{\sigma}_0(\theta|\phi), \widetilde{\sigma}_j(\theta|\phi)\} > \theta$ satisfies the incentive compatibility and hence ϕ is over-proportional.

(i) ⇒ *(iv)*:

Fix $\theta \in (0,1)$. Because ϕ is over-proportional, the set

$$I(\theta) = \{\sigma \ : \ \sigma > \theta, \ S_i(\sigma|\phi) \geq L_i^0(\theta)\}$$

is not empty. [**Are we assuming** $S_j' < 0$?] If $S_j' < 0$, clearly, $\widetilde{\sigma}_0(\theta|\phi), \widetilde{\sigma}_j(\theta|\phi) \in I(\theta)$. Therefore, we only need to show that $\widetilde{\sigma}_0(\theta|\phi) \leq \widetilde{\sigma}_j(\theta|\phi)$, which is true because $S_0(\widetilde{\sigma}_0(\theta|\phi)|\phi) \leq S_0(\widetilde{\sigma}_j(\theta|\phi)|\phi)$ and S_0 is strictly increasing. $\square$