ESSAYS IN LAW AND MICROECONOMIC THEORY

by

HOWARD F. CHANG

A.B., Government, Harvard College (1982);
M.P.A., Economics and Public Policy,
Princeton University (1985);
J.D., Harvard Law School (1987);
S.M., Economics, Massachusetts Institute of Technology (1983)

Submitted to the Department of Economics
in Partial Fulfillment of the Requirements of the Degree of
Doctor of Philosophy in Economics

at the

Massachusetts Institute of Technology

September 1992

Signature of Author: ______________________________
Department of Economics
June 18, 1992

Certified by: ______________________________
Oliver Hart
Professor of Economics
Thesis Supervisor

Accepted by: ______________________________
Richard Eckaus
Chairman, Department Graduate Committee
Department of Economics

ESSAYS IN LAW AND MICROECONOMIC THEORY

by

HOWARD F. CHANG

Submitted to the Department of Economics on June 18, 1992
in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy in Economics

ABSTRACT

This dissertation presents a series of three essays. Each chapter is an essay analyzing the effects of particular legal regimes upon the behavior of the private actors subject to these rules. Each essay develops a formal model and applies the tools of microeconomic theory to shed light on salient issues in a specific area of the law.

The first essay presents a model of cumulative innovation to investigate what factors should influence a court's decision when a patentee alleges that another inventor has infringed the patent with an improved version of the patented product. First, the model reveals how the optimal patent policy would extend broad protection not only to those inventions that are very valuable have (standing alone) relative to the improvements that others may subsequently invent, but also to those that very little value relative to possible improvements. Such inventions should receive broad patents in the sense that very few improvements should be permitted to avoid infringement. Second, the essay examines whether courts should allow the holder of a patent and competing inventors with improved versions of the patented product to enter collusive agreements. The model indicates that such a lax antitrust policy would create incentives for inefficient entry by imitators who "invent around" the original patent.

The second essay develops a sequential bargaining model of the negotiations in corpororate reorganizations under Chapter 11 of the Bankruptcy Code. The analysis identifies the expected outcome of the bargaining process and examines the effects of the legal rules that shape the bargaining. The analysis determines how much value equityholders and debtholders receive under the Chapter 11 process, and compares the value obtained by each class with the "contractual right" of that class. The essay identifies and analyzes three reasons that the equityholders can expect to obtain some value even when the debtholders are not paid in full. Finally, the essay shows how the features of the reorganization process and of the company filing under Chapter 11 affect the division of value, and in this way the model provides several testable predictions.

The third essay notes that neither the American rule of litigation cost allocation, under which each litigant bears its own expenses, nor the British rule, under which the losing litigant pays the attorneys' fees of the winning litigant, would induce plaintiffs to make optimal decisions to bring suit. The essay analyzes the effect of more general fee-shifting rules that are based not only upon the identity of the winning party but also on how strong the court perceives the case to be at the end of the trial -- that is, the "margin of victory." In particular, the analysis explores how one can design such a rule to induce plaintiffs to sue if and only if they believe their cases are sufficiently strong. The analysis suggests some considerations to guide the interpretation of Federal Rule of Civil Procedure 11.

Thesis Supervisor: Dr. Oliver Hart

Title: Professor of Economics

ACKNOWLEDGMENTS

This dissertation reflects not only my own efforts, but also the contributions and assistance of others. In particular, I owe special thanks to my thesis advisors Oliver Hart and Franklin Fisher for their many valuable suggestions throughout the writing process. Furthermore, I am most indebted to Lucian Bebchuk of Harvard Law School, with whom I wrote the second and third essays in this dissertation. I am grateful for the opportunity to work with him and for the inspiration and guidance he provided during our collaboration on both projects.

I wish to thank Louis Kaplow for stimulating my interest in the economic analysis of antitrust and patent law, for his careful reading of my first two essays, and for many useful discussions. I also appreciate the help of Steve Shavell, who (through the Harvard Law School Program in Law and Economics) invited me to present both papers at seminars at Harvard Law School and circulated earlier drafts of both essays as discussion papers. Kaplow, Shavell, and others at my seminars at Harvard, as well as participants at my seminars at M.I.T., provided helpful comments on both papers. I am also grateful to Jean Tirole, for teaching me the theory of industrial organization and for his perceptive comments on the first essay.

The first and second essays were much improved in response to comments from many others. Rebecca Henderson, Richard Nelson, Steven Salop, Andrea Shepard, and seminar participants at the University of California at Berkeley offered helpful reactions to the first essay. The comments of Marcel Kahan, Kevin Kaiser, Roberta Romano, Jeff Zwiebel, and two anonymous referees generated important insights in the second essay.

I also gratefully acknowledge financial support from the John M. Olin Foundation, which funded the fellowships I received during each of my four years of studies at M.I.T.

Finally, I am deeply indebted to Hilary Sigman, my colleague and companion, who provided encouragement and suggestions during the writing of this dissertation, and to my parents, who taught me the value of hard work, patiently provided financial support during my many years of education, and otherwise supported me in all my academic endeavors.

TABLE OF CONTENTS

CHAPTER I:

PATENT SCOPE, ANTITRUST POLICY, AND CUMULATIVE INNOVATION

Howard F. Chang

A. INTRODUCTION

The Patent Office and the courts are constantly called upon to decide the proper scope of patents. Courts frequently must decide whether a second invention infringes upon the patent of a preceding invention. These decisions have important effects on the pace of technological progress through the incentives to invent not only the first invention but also later inventions that build upon the first. These effects are critical in the design of the optimal patent policy, because the dynamic losses that flow from a slower pace of innovation can easily swamp the static deadweight losses from patent monopolies. Nevertheless, until recently scholars have devoted little attention to the scope of patent protection and the issue of cumulative innovation.

This essay presents a formal model of cumulative innovation to determine if (and how) the scope of patents should be limited so as to encourage the invention of valuable improvements on the patented product or process. Courts have declined to find infringement when the patentee complains of a competing product that features a significant improvement over the patented product. Accordingly, this essay investigates how such an improvement should influence a court's decision in an infringement case. The results suggest that, in light of the objective of promoting cumulative research, courts should curtail patent protection for an basic innovation so as to allow later innovators to develop improvements that are particularly valuable (among all possible improvements on that basic innovation) without violating the patent.

Most important, the formal model also suggests a rule for taking the value of these inventions into account that differs from that proposed by Merges and Nelson (1990, pp. 865-66), who suggest that courts decline to find infringement when the value of the original invention is small relative to the value of the improvement. Instead, courts should extend the broadest protection not only to the patent with very large stand-alone value relative to all possible subsequent improvements, but also to the patent with very little stand-alone value relative to the improvements that it may inspire. Thus, although the basic invention by itself may have almost no value, the courts should allow the original patentee to claim rights over nearly all the improvements built upon the patented technology. When the first invention has no economic value except as a necessary condition to subsequent innovation, then the only result that can permit the first inventor to appropriate any reward at all is a finding of infringement. Thus, patents for such basic inventions should be permitted wide scope in the sense that only the most extraordinary improvements should be permitted to avoid infringement.

This essay also examines a related question: whether the courts should interpret the antitrust laws so that the holder of a patent and competing inventors with improved versions of the patented product may enter collusive agreements and thereby avoid competition that would erode the profits that encourage innovation. This essay evaluates this issue in a model designed to provide a synthesis of some of the opposing considerations raised in the preceding literature. Whereas Scotchmer (1991) and Green and Scotchmer (1990) suggest that courts should permit collusive licensing between competing patentees, Kaplow (1984) has argued that such an antitrust policy rewards innovation only at an excessive social cost. The analysis seeks to reconcile the views of Kaplow and of Green and Scotchmer regarding the antitrust treatment of collusion between competing innovators.

The results provide only limited support for a permissive antitrust policy regarding collusion between competing patentees. Such a policy would create incentives for inefficient entry by imitators who "invent around" the original patent. The optimal policy in the context of cumulative innovation would permit collusion only between holders of basic patents with little value (standing alone) and imitators who contribute relatively valuable improvements, and even in these cases the analysis suggests that the benefits from such collusion may be small.

Section B of this essay reviews both the relevant economic theory developed in the preceding literature and the relevant patent and antitrust law. This background discussion sets forth the issues that the formal model will address. Section C sets forth a formal model of cumulative innovation and lays out the underlying assumptions. Section D presents the derivation of the optimal innovation policy. Section D.1 a derives an optimal policy that corresponds to a compulsory licensing scheme with courts setting patent royalties. This policy proves to be a useful benchmark for comparison to the policies studied next. In Section D.2, the courts can only influence the inventors' incentives through decisions on patent scope and antitrust law. The optimal policy bases the infringement and antitrust decisions on the values of the basic invention and of the improvement that is alleged to infringe the basic patent. Section E concludes with implications for policy from the results and suggests directions for further research and possible extensions of the model.

B. BACKGROUND

In an environment in which the regulatory authorities cannot observe research and development (R&D) costs, authorities seeking to encourage investment in R&D ex ante if and only if such efforts are socially desirable should set the inventor's private reward ex post equal to the social benefits of the invention. A system of patents cannot achieve this equality ex post unless firms can engage in perfect price discrimination and thereby appropriate all consumer surplus. Furthermore, in the absence of perfect price discrimination, patents impose the familiar ex post deadweight loss that results from monopoly pricing: such prices exclude some potential users of the invention that are willing to pay more than the marginal cost of production.

This essay focuses on one particular example of the ex post social costs of patents: the inhibition of R&D by other inventors with ideas for improvements in the patented product or process. Once another inventor has invested in developing an improvement, the original patentee may use its monopoly over its innovation to appropriate some of the value created by such a complementary innovation, even if the second innovator patents the improvement. If the second innovator can market its invention only with the consent of the original patentee, that patentee can increase its profits at the expense of the second innovator by bargaining to license the complementary technology at less than its full value. This "holdup" problem reduces R&D in complementary technologies by other inventors by reducing the expected return on their investment.[1]

As in the context of vertical integration and contracts discussed by Klein, Crawford, and Alchian (1978) and by Williamson (1979), opportunistic appropriation ex post (in this context, after R&D costs are sunk) discourages investments in assets (innovations) that have their highest value when used in conjunction with an asset (the original patent) held by another party. This underinvestment results whenever contracts are incomplete in that the parties cannot agree ex ante to specific levels of investment (because a court cannot observe levels ex post, for example) or to a particular division of the surplus generated by the investment. A complete contract ex ante could encourage the first-best level of investment by ensuring that the party investing receives an adequate return on its investment.

If the first innovation is a necessary condition for the second innovation, then the surplus from the second innovation is a joint product of both innovations. In this case, the holdup problem is bilateral: to the extent that the second innovator receives positive profits, then the first innovator will tend to underinvest in R&D relative to the first-best. As Scotchmer (1991, p. 34) and Green and Scotchmer (1990) observe, in markets with cumulative innovation, patent protection cannot offer both the first innovator and the second innovator the full surplus from the second innovation. As a result, some distortion of incentives is unavoidable under a patent system: at least one party will have too little incentive

[1]This particular social cost of patents merits special study because it is likely to have important implications for policy. The "holdup" problem is important, because the costs of a small dynamic effect can rapidly accumulate to dwarf the costs of a static effect such as the deadweight loss from monopoly pricing. See Scherer (1980, p. 407). The policy implications of the holdup problem will differ from those of the deadweight loss problem, because various policy instruments will have different effects on these two problems. For example, lax antitrust regulation of horizontal agreements may increase the static costs of patents, but as discussed below, it may decrease the dynamic costs of patents by encouraging subsequent innovation. On the other hand, whereas price discrimination ex post may reduce the static costs of patents, it cannot solve the "holdup" problem unless the original patentee can commit ex ante to reward a subsequent inventor ex post.

to invest in innovation.

Faced with such a problem, the parties to an incomplete contract ex ante can allocate residual rights (i.e., ownership) to one party or the other before any investments are made and, as Grossman and Hart (1986) show, thereby influence any renegotiation to divide the surplus ex post. By setting the "threat point" in such ex post bargaining, the contract can trade off distortions in one party's ex ante incentives against those in the ex ante incentives of the other. Similarly, courts can affect ex ante incentives by allocating intellectual property rights: they decide how different a second invention must be from a prior patented invention to avoid infringement ex post. Green and Scotchmer observe that by adjusting patent scope in this manner, courts can shift surplus between the inventors ex post (by affecting their bargaining over a licensing agreement, for example) and thereby seek to minimize the distortion of R&D investment decisions ex ante.[2]

At this point, it will prove useful to review the law on patent scope which raises the issues that we will address. A patent application consists of a specification of the invention and a set of claims as to the subject matter that the applicant regards as protected by the patent. See Patent Act, 35 U.S.C. § 112 (1988). Under the "doctrine of equivalents," a court will hold that a device that falls outside a patent's claims nevertheless infringes the patent if the device has avoided those claims only through minor variations in the invention.

Courts generally provide broad patent protection for particularly significant inventions. Patents for "pioneer" inventions, which cover "a function never before performed, a wholly novel device, or one of such novelty and importance as to mark a distinct step in the progress of the art" as defined in Westinghouse v. Boyden Power Brake Co., 170 U.S. 537, 561-62 (1898), are entitled to a broad

---

[2]Copyright protection presents similar problems, because it raises the costs of creating new works that borrow or build upon material from a prior work. Thus, Landes and Posner (1989) and Menell (1989, pp. 1079-88) analyze the scope of copyright protection as an instrument with which to balance these costs against the benefits flowing from enhanced incentives for the original work.

range of equivalents. Inventions that represent less marked advances over prior technology are entitled to a correspondingly more limited range of equivalents. See Lipscomb (1987, §§ 22:41-:45).

A court may refuse to find an infringement under the doctrine of equivalents if the allegedly infringing device features major improvements rather than unimportant or insubstantial changes. Indeed, on rare occasions, a court may even hold that a patent does not cover a subsequent improved version of the invention that falls within the patentee's claims. See Chisum (1990, Vol. 4, § 18.04). The Supreme Court applied this "reverse" doctrine of equivalents in Westinghouse v. Boyden Power Brake Co., 170 U.S. at 568-73. In that case, George Westinghouse had invented a train brake that used reservoirs of compressed air for stopping power, and George Boyden had invented an ingenious improvement that provided added stopping power. Noting that the Westinghouse patent "did not prove to be a success until certain additional members had been incorporated into it," id. at 572, the Court refused to find that Boyden infringed, because his new design featured substantial improvements.[3]

One who invents an improvement on the patented product of another can obtain a patent on the improved feature alone. See Patent Act, 35 U.S.C. § 101 (1988). In this case the original patent is called the "dominant" patent, and the improvement patent is called the "subservient" patent. The two inventors then hold complementary "blocking patents": the original inventor cannot practice that particular improvement without a license from the second inventor, but the second inventor cannot practice the improvement either without a license from the original inventor. See Chisum (Vol. 3, § 9.03[2][ii]; Vol. 4, § 16.02[1][a]). For this reason, the second inventor would prefer a patent free of anyone else's claims and

[3]The doctrine of equivalents is somewhat at odds with the general design of the patent system, which avoids official decisions by the courts on the relative technological, social, or economic value of inventions. See Chisum (Vol. 4, pp. 18-101 to -102). Usually, improvements of (or additions to) a patented invention will not avoid infringement. See Lipscomb (§§ 22:29-:30).

will not often voluntarily characterize its invention as subservient. The original inventor, on the other hand, would object to the second inventor's claim to a competing patent. In the course of litigation, a court may find that the second inventor's patent infringes the prior patent, even while upholding the infringer's patent as valid with respect to an improved feature. Such a holding in effect creates a complementary subservient patent, and courts frequently resolve infringement cases in this way. See Merges and Nelson (1990, pp. 861, 864-65)

As discussed above, however, a dominant patent poses a holdup problem for the holder of the subservient patent. For this reason, Merges and Nelson (pp. 857-68, 909-11) defend the doctrine of equivalents as applied in Westinghouse. Patent cases indicate that courts have broad discretion under this doctrine to decide the scope of patent protection, and Merges and Nelson propose that courts use this doctrine on a case-by-case basis to reduce the holdup problem posed by patents in the context of cumulative innovation.[4] In particular, Merges and Nelson (1990) assert that the holdup problem posed by a finding of infringement becomes most acute when the value of the improvement embodied in the infringing invention is large relative to the value of the original invention. Therefore, Merges and Nelson (pp. 865-66) suggest that courts should find infringement when the original invention contributes most of the value of the improved invention, but find no infringement when the original patent contributes very little value compared to the improvement. The model in this essay is designed to shed light on these recommendations regarding the proper criteria for such patent scope decisions.

The courts can affect not only the division of profits between innovators

---

[4]Merges and Nelson (pp. 866 n.118) also suggest that a system of compulsory licensing, under which the patentee would be obliged to license its invention in exchange for a "reasonable royalty," would probably be the most efficient solution to the holdup problem. Although compulsory licensing is well established as a remedy when courts find that the patentees have violated the antitrust laws, in general it remains disfavored in patent law. See Scherer (1980, pp. 456-57) and Areeda and Kaplow (1988, ¶¶ 190, 284(c)-(d)). Merges and Nelson (p. 911) conclude that as long as compulsory licensing remains anathema, courts should rely on the doctrine of equivalents to solve the holdup problem.

through decisions on patent scope, but also the division of surplus among consumers and innovators through antitrust policy. Furthermore, antitrust law, like patent law, leaves courts with considerable discretion to make policy. Kaplow (1984) addressed the inherent tension between antitrust and patent policies: antitrust regulation of the practices of patentees restrict their ability to exploit their market power. While such restrictions may reduce the static deadweight loss associated with patents, they also reduce the returns to the patentee's investment in R&D.

For example, horizontal agreements regarding patents pose an antitrust problem because they can be used by competing firms to collude. If a patent licensing agreement may contain restrictions on the prices to be charged in the product market, then any patent may form the basis for a price-fixing cartel.[5] Similarly, if multiple firms that possess competing patents may cross-license their innovations to one another with price restrictions, they can cartelize an industry. Kaplow (pp. 1855-60) concludes that courts should interpret the antitrust laws to prohibit price-restricted licenses.[6]

Even if antitrust law were to prohibit such collusive cross-licensing among competing patentees, however, these firms can achieve a similar result by selling

---

[5]One firm could license its innovation (even if the innovation is of little economic value) to all other firms in the industry and include a price floor in the licensing agreements. The patentee in such a case could realize profits far greater than the social value of the invention. Such large rewards for trivial inventions would misallocate R&D resources through perverse and excessive incentives.

[6]Although price restrictions in patent licenses are not illegal <u>per se</u>, licenses may not fix prices if the licenses cover substantial parts of the market and thus cartelize an industry through these price restrictions. See Areeda and Kaplow (1988, p. 449) and Sullivan (1977, pp. 545-46, 551-54). To protect its profits from erosion without any price restrictions, a patentee may instead raise the price charged by licensees through a royalty based on units of output by the licensee. Competition between the patentee and its licensee would drive the licensee's prices toward the sum of the licensee's marginal costs and its per-unit royalty. In this case, the reward obtained through such a royalty cannot exceed the value of the patented innovation, because it cannot exceed what the licensee is willing to pay for it (the increase in the licensee's profits attributable to the innovation).

their competing patents to one party.[7] Thus, Priest (1977, pp. 358-64) and Kaplow (pp. 1867-73) address the issue of horizontal combinations among competing patentees and conclude that such practices also should be suspect under the antitrust laws.[8] Scotchmer (1991, pp. 33-34) and Green and Scotchmer (1990) suggest, however, that if a second innovation builds and improves upon the first, and the innovating firms cannot agree ex ante (before investment in the second innovation) to conduct joint R&D, then the courts should allow competing innovators to use collusive licensing agreements ex post to avoid competition that would dissipate their profits.

Green and Scotchmer focus, however, on the possibility of agreements between firms ex ante and on what such research joint ventures imply for the optimal patent-antitrust policy. They do not do not explore the optimal tradeoff between the two innovators' incentives when such ex ante agreements are not possible. This essay will analyze the features of the optimal patent-antitrust policy under such circumstances, keeping in mind that the optimal antitrust policy regarding patents should be determined simultaneously with other aspects of the optimal patent policy.

This essay presents a formal model with two periods and two firms. The first firm can invent and patent a product in the first period; if and only if firm 1

---

[7]Suppose that a firm must possess a patent to sell in the relevant market and that each firm's patent is a perfect substitute for the others, so that no firm would be willing to license any patent from any other in the absence of price restrictions. Whereas product-market competition among the patentees would drive the profits of each toward zero, by selling their patents to one owner (who can then charge a monopoly price) the patentees can reap and divide the joint monopoly profits as a cartel would.

[8]The acquisition of many patents is not illegal <u>per se</u>, but the pooling of competing patents is analogous to a merger: a combination of patents may violate the antitrust laws if it creates excessive market power. See Areeda and Turner (1978, Vol. 3, ¶¶ 704b, 705b, 819c) and Sullivan (pp. 515-20). Courts will consider efficiency justifications for an otherwise suspect arrangement. Patent pools (which share royalties among the participants) often result from the settlement of patent disputes, for example, and the general policy in favor of settlements -- which save the parties and society the costs of litigation -- may be a factor that militates in favor of the legality of the arrangement. See, e.g., <u>Standard Oil Co. (Indiana) v. United States</u>, 283 U.S. 163 (1931).

invents, then the second firm can invent and patent an improved version of the product in the second period. The incentives for the firms to invent depend upon the patent-antitrust policy implemented by the courts. These courts must commit ex ante to a policy that will apply to a large set of potential R&D projects of various expected values and costs, but the courts must design their policy knowing only the joint probability distribution over the possible values and expected costs of these innovations.

In particular, I shall consider a court's decision when firm 1 claims that the firm 2 has infringed firm 1's patent. Green and Scotchmer assume that patent policy takes the form of a cutoff value that is a function of the value of the original product: if the value of the improvement is greater than the cutoff value, the improved product does not infringe; otherwise, it does infringe. In contrast, the assumptions in this essay do not restrict patent policy to take such a form.[9] Instead, I shall evaluate (i) whether the court's decision on infringement should turn on the values of the two innovations, and if so, (ii) precisely how the optimal policy will depend on these values. I shall also determine when licensing agreements ex post between competing patentees are desirable, as Green and Scotchmer suggest, and when such collusive agreements are inefficient, as Kaplow argues. The model will assume that courts may decide, for each pair of values for the two innovations, whether to allow collusive licensing agreements, both when the improved invention competes lawfully with the first invention and when it instead infringes the original patent (so that the improvement properly

[9]Green and Scotchmer do not show that such a cutoff function is the optimal patent policy. If a cutoff function dominates other possible policies, presumably, it is because when one must reduce incentives to the first innovator in order to offer greater incentives to the second innovator, it is best to sacrifice the first innovator's profits in those cases in which the expected value of the improvement is largest relative to the value of the original innovation. In Green and Scotchmer's model, however, the expected value of the improvement (as well as the R&D cost for the first innovation) does not vary. Thus, their model does not address the considerations that arise in these circumstances, and as this essay will show, these considerations have important implications for other aspects of the optimal patent-antitrust policy.

complements the first invention).[10] This essay will also consider a richer set of policy instruments, including compulsory licensing for patents as a hypothetical instrument.

C. THE MODEL

Consider a simple two-firm two-period model with no discounting. At the start of period 1, an idea for a product occurs exogenously to firm 1. For simplicity, I assume all consumers are identical and each demands one unit of the product per period. If firm 1 invests an amount c_1 in R&D, then firm 1 develops product 1, and the consumers are willing to pay an amount v_1 per period for the product. If and only if firm 1 develops product 1, then at the start of period 2 an idea for product 2 -- an improved version of product 1 -- occurs exogenously to a firm 2. If firm 2 invests an amount c_2 in R&D, then firm 2 develops product 2, which exceeds product 1 in value by an amount v_2 per period. There exist no other substitutes for these products. Once either product is developed, it can be produced at zero marginal cost.

At the time each of the innovators makes its investment decision, each knows the value and the R&D cost of its own innovation with certainty. Firm 1, however, faces uncertainty in period 1 over both the value and the cost of the second innovation in period 2. For simplicity, I assume that firm 1 knows only that v_2 is distributed continuously in the interval $[0, V_2]$ with positive density $f(v_2)$, where $F(v_2)$ is the corresponding cumulative distribution function and $V_2>0$, that c_2 is uniformly distributed in the interval $[0, C_2]$, where $C_2>0$, and that v_2 and c_2 are independently distributed.

[10]Thus, the model is general enough to consider the suggestion that complementary patents should receive more permissive antitrust treatment than competing patents. Priest (pp. 357-58) asserts that price restrictions are essential when firms cross-license complementary patents, because the firms will otherwise compete away the patent rents. Green and Scotchmer do not address this question, because their model does not allow the antitrust policy to be contingent upon whether the improved invention infringes the original patent.

Each firm is risk-neutral and chooses to develop its product if and only if its expected revenue exceeds its R&D cost. The expected revenue for each firm, however, depends upon public policy, and I shall use this model to explore the effects that alternative policies have upon the incentives to invent. For example, suppose that in the absence of any property rights for firm 1 in its invention, other firms would enter the market and use firm 1's technology for free. Suppose also that Bertrand competition then drives the price of product 1 to 0. Anticipating this outcome, firm 1 would not invest in product 1, and firm 2 would never have the opportunity to develop product 2.

Suppose instead that each firm receives an exclusive right to market its invention, i.e., a patent. If firm 1 sells its product under a patent monopoly, with identical unit demands and unregulated prices, then firm 1 could appropriate all social surplus flowing from its sales. Specifically, if firm 1 sells product 1 free from competition from product 2, it will choose the monopoly price v_1 and leave no consumer surplus. If firm 2 sells product 2 in competition with product 1, however, I assume that Bertrand competition drives the price of product 1 to 0 and that of product 2 to v_2 (minus some infinitesimal amount). Therefore, only product 2 is sold in equilibrium, but its price is constrained by the potential competition from product 1. Due to this competition, those who buy the product would enjoy v_1 in consumer surplus.

Again public policy ex post affects firm 1's incentives to invent, in this case not only through the likelihood that firm 2 decides to develop product 2, but also through the consequences of such a decision for firm 1's expected revenues. If firm 2 develops product 2, for example, suppose that firm 1 brings a lawsuit seeking to enjoin the marketing of product 2, claiming that product 2 infringes on the patent for product 1. If the court decides the second product infringes the patent granted to firm 1, then that patent includes the right to block entry by firm 2. Although this holdup right increases firm 1's expected revenues and encourages innovation by firm 1, it also reduces firm 2's revenues and discourages innovation by firm 2.

If firms can reach agreement ex ante (before investment in the second innovation) to conduct joint R&D in product 2 or to transfer the patent for product 1 to firm 2, then they can avoid this holdup problem and also avoid competing with one another in the product market in period 2. I assume that such transactions are either illegal or too difficult to arrange. There may be too many innovators with ideas for possible improvements, or transaction costs may be too great, for ex ante contracts to be a complete solution. See Scotchmer (1991, p. 37) and Merges and Nelson (1990, p. 877 n.160). Firm 2 may find that it cannot induce firm 1 to agree to an R&D joint venture without disclosing its idea. Such disclosure, however, would undermine the bargaining power of firm 2. Similarly, such asymmetric information may lead to strategic behavior that would inhibit the ex ante licensing or sale of firm 1's patent to firm 2, as the two firms bargain over the division of surplus.[11]

Now suppose that the courts act as a principal with limited information that seeks to design a general innovation policy that will induce desirable behavior on the part of its agents (the innovators). I assume that the courts can observe only the different flow values of the innovations ex post, v_1 and v_2 (not their costs, c_1 and c_2). The courts also know the information available to firm 1 in period 1, including the distributions of c_2 and v_2. Finally, the courts know that the two products are perfect substitutes aside from their vertical differentiation, that the first innovation is necessary for the second, that c_1 is uniformly distributed in the interval $[0, C_1]$, where $C_1>0$, and that v_2, c_1, and c_2 are all distributed independently of v_1 and of each other.

The courts establish a reputation for adherence to a particular policy (by following precedent), so as to give their policy the desired effect. This policy is a function mapping each point in (v_1, v_2) space to a pair (π_1, π_2), which represents the second-period revenues received by each of the two firms. More precisely,

[11]See Caves, Crookell, and Killing (1983) for a discussion of market failures in the bargaining over technology licenses.

policy will allocate each pair of relative values, v_1/V_2 and v_2/V_2,[12] to a legal regime that determines both (1) the prices paid by consumers, and (2) how the resulting revenue is allocated between the two firms.

In Section D.1, I shall assume that the courts can pursue a hypothetical policy through which they can set the absolute value of these variables precisely for each (v_1, v_2) pair. To implement such a policy in practice, the courts would need to observe the absolute values of v_1 and v_2. In this model, this hypothetical policy is equivalent to a government agency that purchases each innovation from the corresponding patentee and sells the improved product to the public, but is required to balance its budget (i.e., set revenues equal to the total amount paid to the two innovators). This policy is also equivalent to the contract to which the parties would agree ex ante (before any opportunities for R&D occur), if the absolute values (but not the costs) of inventions are verifiable ex post in court. I refer to the optimal hypothetical policy as the "second best" insofar as it is the best that the court could implement given that it cannot observe R&D costs and given the "balanced budget" constraint (that is, the requirement that inventors be rewarded only with revenues from product sales).[13]

---

[12]Without loss of generality, we can normalize so that all variables and parameters express values relative to V_2. That is, the features of the optimal policy will depend only on values relative to V_2. Thus, the model is general enough to include cases with varying absolute values for V_2; one need only interpret the other variables and parameters accordingly.

[13]In reality, courts can observe the economic value of an invention only imperfectly. If the regulatory authorities could observe the absolute value of each invention perfectly, then one might think that an unconstrained system of rewards ex post for successful innovators (the "award" or "prize" system) would perform better than budget-constrained policies, including the patent system. Indeed a major rationale for the patent system is the fact that the regulatory authorities lack information that innovators possess, such as knowledge regarding the demand for inventions. For this reason, it may be better to rely on a patent system in which the monopoly profit extracted from the market is correlated (albeit imperfectly) with the social surplus created by the invention.

There may be other reasons, however, that one may prefer not to rely on an unconstrained government agency that transfers funds to innovators based on its appraisal of the values of their inventions. Laffont and Tirole (1990) model another problem that may arise if regulators are authorized to make transfers to firms: the agency may be

In Section D.2, I shall assume instead that the courts have only a cruder set of instruments: this "third best" policy will consist of a function mapping each point in (v_1, v_2) space to one of four possible patent-antitrust regimes. To implement this policy, however, the courts need not know the absolute values of either v_1 or v_2; it will suffice if the courts observe the relative values v_1/V_2 and v_2/V_2. For each (v_1, v_2) pair, the court would decide both (1) whether the second product infringes the first patent, and (2) whether the two firms should be allowed to collude through a licensing agreement (or a transfer of one of the patents). These two binary decisions create a matrix of four possible regimes, each of which implies not only particular second-period revenues (π_1, π_2) for firm 1 and firm 2, respectively, but also a particular level of consumer surplus.

The second-best policy would dominate this set of four possible outcomes, because the hypothetical policy in Section D.1 could replicate each of the four outcomes as a special case. Indeed, under that hypothetical policy, courts could divide the social surplus among the consumers and the two innovators in any way desired. As shown in Section D.1, courts in theory could implement the second-best policy through a compulsory licensing policy. Current law, however, restricts courts to the cruder third-best policy studied in Section D.2, perhaps because a general compulsory licensing policy would require courts to gather and process too much information. See Areeda and Kaplow (1988, ¶ 284(d), pp. 444-45). In any

---

"captured" by the industry. The award system also suffers from the problem that the authorities are tempted to "hold up" the inventor through conservative estimates of the invention's value.

In any event, the justification of the patent system itself is beyond the scope of this essay. Like much of the literature on the optimal patent policy, e.g., Green and Scotchmer (1990) and Klemperer (1990), this essay will instead take the constraints on policy as given and assume for the sake of argument that the regulatory authorities have full information regarding the demand for the innovations in question. The model makes this assumption simply to address the question: to the extent that the courts can acquire information on the value of inventions, how (if at all) should that affect their decisions? Even without any direct measure of value, courts can observe the degree of differentiation between the products, for example -- a characteristic that may be correlated with the economic value of the improvement.

event, analysis of the second-best policy will prove useful in understanding the features of the third-best policy.

Using either policy, the court can affect social welfare through the payoffs to each of the two firms. Let $\pi_2(v_1, v_2)$ denote the revenue (from the sales of product 2) received by firm 2 if it develops the product. This $\pi_2(v_1, v_2)$ will be a function of the courts' innovation policy. Thus, firm 2 would develop product 2 if and only if its expected net payoff is positive, that is, if and only if $\pi_2(v_1, v_2)$ exceeds c_2. Similarly, firm 1's R&D decision would depend on the innovation policy as specified in Section D.

Before proceeding to that discussion, it will prove useful to introduce some more notation. The social value flowing from sales of product 2 is v_2. Let $S(v_1)$ denote the expectation in period 1 of the net social value (conditional on firm 1 having developed product 1 with value v_1) created by the possibility that firm 2 will develop product 2 in period 2. Thus, $S(v_1)$ equals the expected value of v_2-c_2, conditional on the development and sale of product 2, multiplied by the probability of the development and sale of product 2, $Pr[\pi_2(v_1, v_2) > c_2]$. For any given v_1, I assume:

$$C_2 \geq \pi_2(v_1, v_2) \qquad (1)$$

for all possible v_2 and for all possible innovation policies in period 2.[14] Then we may express $S(v_1)$ as follows:

$$S(v_1) = \int_0^{V_2} \int_0^{\pi_2(v_1,v_2)} \frac{v_2 - c_2}{C_2} dc_2 \; dF(v_2), \qquad (2)$$

where we integrate with respect to v_2 to include all possible ideas for improvements

---

[14]This assumption ensures that greater rewards for firm 2 will always induce greater expected R&D expenditures. In this model, any policy that increases the expected payoff to firm 2 would encourage more expected innovation in period 2 only by inducing firm 2 to undertake projects that were otherwise too costly to be worthwhile. To create this effect, there must be some probability that the cost of developing product 2 exceeds the revenue to firm 2. Assumption (1) is necessary because c_2 can realize values only as high as C_2. In terms of equation (2), this simplifying assumption allows us to use $\pi_2(v_1, v_2)$ as an upper limit of integration rather than $\min[C_2, \pi_2(v_1, v_2)]$.

and with respect to c_2 to include all possible R&D costs for any given v_2. For any given v_1, S will depend on how the innovation policy would treat a second invention (i.e., what π_2 each possible value of v_2 would imply).

Let $\Delta\pi_1(v_1, v_2)$ represent the change in firm 1's second-period revenues induced by the development of product 2. This $\Delta\pi_1(v_1, v_2)$ will be a function of the courts' innovation policy. Let $P(v_1)$ denote the expected net private benefit for firm 1 (conditional on having developed product 1 with value v_1) created by the possibility that firm 2 will develop product 2. Thus, $P(v_1)$ equals the expected value of $\Delta\pi_1(v_1, v_2)$, conditional on the development and sale of product 2, multiplied by the probability of the development and sale of product 2, $Pr[\pi_2(v_1, v_2) > c_2]$. That is:

$$P(v_1)=\int_0^{V_2}\int_0^{\pi_2(v_1,v_2)}\frac{\Delta\pi_1(v_1,v_2)}{C_2}dc_2\ dF(v_2), \tag{3}$$

where we integrate with respect to v_2 to include all possible ideas for improvements and with respect to c_2 to include all possible R&D costs for any given v_2. For any given v_1, P will depend on the $\Delta\pi_1$ and π_2 implied by each v_2 under the existing innovation policy. We can now analyze the optimal innovation policy.

D. THE OPTIMAL INNOVATION POLICY

Suppose that the courts seek to maximize social welfare through their innovation policy. Given unit demands, as long as the price of a product does not exceed its value, that price merely determines how the social surplus from the product is divided between producer and consumer; there is no deadweight loss from prices above marginal cost in this model. To the extent that policies affect a product price within this range, then, they would affect social welfare only through the incentives for producers to invent and market products. The "first best" policy from this perspective would align the private incentives for producers with the social value derived from these products.

For example, to optimize the incentives to firm 1 in the absence of firm 2, courts would permit firm 1 to charge the monopoly price for product 1, v_1, which would allow firm 1 to appropriate as much social surplus as possible. To price any lower would lead firm 1 not to invest in R&D in cases in which the social benefit of the invention would exceed its cost. The lower the price below v_1, the greater the expected loss in social surplus, both because these cases become more numerous and because the average value (net of R&D costs) of the invention foregone increases. To price any higher than v_1, however, would eliminate any sales (and thus any benefits, private or social) of the product. Similarly, given the existence of product 1, to optimize incentives for firm 2 to invent product 2, a court would give firm 2 the full social value of product 2, v_2. That is, $S(v_1)$ is maximized by $\pi_2(v_1, v_2) = v_2$.

To maximize welfare, however, the courts must also consider the incentives for the first inventor, given the possibility of the second invention. Consider the optimal innovation policy in the relatively simple case in which the technology for the second invention is independent of the first invention, that is, in which the probability distribution for the cost and value of the second innovation is the same whether or not firm 1 develops product 1. In such a case, a finding of noninfringement and a prohibition on collusion between firm 1 and firm 2 would provide optimal incentives to both firms. In competition with product 1, firm 2 would receive the social benefit of its invention, which would be the amount by which product 2 exceeds product 1 in value, v_2. Moreover, given the possibility of the invention of product 2, firm 1 receives the social benefit of its invention, which is v_1 for the first period, plus v_1 for the second period if firm 2 does not invent product 2, and 0 if firm 2 does invent product 2.

In this model, however, as is often the case in reality, firm 2 would not have invented product 2 unless firm 1 had first invented product 1. The social benefit from the invention of product 1, then, would include the spillover to firm 2: product 1 provides firm 2 with the inspiration for product 2. To provide socially optimal incentives to firm 1, then, the court would increase firm 1's payoffs to include the expected value (in period 1) of this spillover (in period 2): if firm 2 receives socially optimal incentives, then this spillover is the expected value of $\max(0, v_2-c_2)$. The social value of firm 1's innovation, then, would be:

$$2v_1 + E[\max(0, v_2-c_2)]. \qquad (4)$$

In the second period, however, a patent system cannot offer both producers together more than the v_1+v_2 that consumers are willing to pay for the improved product. The problem for the courts, then, is to allocate the social surplus so as to create the optimal trade-off between the incentives to the two firms.

In this model, for any possible patent policy (optimal or otherwise) that the court could implement, the following lemma will hold:

Lemma 1: Any patent policy will offer firm 1 less than the expected social value of its innovation.

Proof: The expected social value of firm 1's innovation equals the sum of firm 1's expected revenues, firm 2's expected surplus (its expected revenues π_2 minus its expected R&D costs c_2), and expected consumer surplus. A patent system can offer firm 1 and firm 2 no more revenue than what consumers are willing to pay for product 1, v_1, plus (if firm 2 invents product 2) what they are willing to pay for the improvement embodied in product 2, v_2, because the patent system has no other source of revenues. Furthermore, if firm 2 is to invent product 2, then a patent system must offer firm 2 at least its R&D costs, c_2. Indeed, because c_2 may vary, and the patent authorities cannot observe its realized value, any given revenue π_2 offered to firm 2 would leave firm 2 positive expected surplus, because the expected value of c_2 (conditional on firm 2's invention of product 2) must be less than π_2. Thus, a patent system must offer consumers a nonnegative share of the social surplus, and firm 2 a positive share. Therefore, firm 1 cannot appropriate all the social surplus flowing from its innovation. ■

Lemma 1, in turn, implies the following lemma:

Lemma 2: The welfare-maximizing court would allow firm 1 to charge the monopoly price v_1 in the first period, and if firm 2 does not invent product 2, in the second period as well.

Proof: Any lower price would transfer value to consumers without increasing social surplus ex post, but would reduce firm 1's incentives to invent ex ante. Reducing firm 1's incentives must reduce social welfare, because in this model, as Lemma 1 states, the patent system already offers firm 1 less than the social value of its innovation. ■

Given the price v_1 in Lemma 2, we can state:

$$\pi_1(v_1, v_2) = v_1 + \Delta\pi_1(v_1, v_2). \tag{5}$$

The expected payoff to firm 1 for inventing product 1 equals $2v_1$, which is its revenues in the absence of product 2, plus $P(v_1)$, which is the net change in its expected payoff due to the possibility of the second innovation. For any given v_1, I assume:

$$C_1 \geq 2v_1 + P(v_1) \tag{6}$$

for all the possible innovation policies in period 2.[15] Then we may express the social welfare function (conditional on a particular v_1), denoted by $W(v_1)$, as follows:

$$\begin{aligned} W(v_1) &= \int_0^{2v_1+P(v_1)} \frac{2v_1+S(v_1)-c_1}{C_1} dc_1 \\ &= \{\frac{1}{C_1}[2v_1+P(v_1)]\}\{[2v_1+S(v_1)]-\tfrac{1}{2}[2v_1+P(v_1)]\} \end{aligned} \tag{7}$$

where we integrate with respect to c_1 to include all possible R&D costs for any given v_1. The second expression for $W(v_1)$ is the product of two terms. The first term is the probability that firm 1 will invent product 1. The second term is the net social benefit from that invention -- $2v_1$ plus $S(v_1)$ minus the expected R&D costs for the first invention.

Expected social welfare (unconditional) would be the expectation of $W(v_1)$ taken with respect to v_1. To maximize this expected welfare, however, the court

---

[15]Assumption (6), by reasoning similar to that in footnote 13, ensures that greater rewards for firm 1 would always induce greater expected R&D expenditures. In terms of equation (7), this simplifying assumption allows us to use $2v_1+P(v_1)$ as the upper limit of integration rather than $\min[C_1, 2v_1+P(v_1)]$.

need only maximize $W(v_1)$ for each v_1, because the court can observe v_1 and tailor its innovation policy accordingly. Furthermore, policy can also depend on v_2, and the optimal policy must be such that no small change in policy for any (v_1, v_2) pair can cause a welfare improvement.

To study the features of a welfare-maximizing policy, then, we shall consider the effect of changes in $\Delta\pi_1$ and π_2 for one (v_1, v_2) pair. In particular, consider the effect of such changes in innovation policy on welfare conditional on a given v_1. As we can see from (7), policy affects $W(v_1)$ only through $S(v_1)$ and $P(v_1)$. For small policy changes, such as changes pertaining only to one (v_1, v_2) pair, we can totally differentiate (7) to obtain:

$$dW(v_1) = \{[2v_1+P(v_1)]dS(v_1) + [S(v_1)-P(v_1)]dP(v_1)\}/C_1. \tag{8}$$

Note that both the coefficient on $dS(v_1)$ and that on $dP(v_1)$ must be nonnegative.[16]

These two coefficients represent respectively the relative importance of the two objectives of the optimal innovation policy. Ceteris paribus, the courts should optimize the incentives to firm 2 so as to increase $S(v_1)$, the expected social surplus from the invention of product 2 (conditional on the invention of product 1 with value v_1). The importance of this objective is proportional to the probability of the first invention, $[2v_1+P(v_1)]/C_1$. Ceteris paribus, the courts should also optimize the incentives to firm 1, which entails increasing $P(v_1)$ insofar as $P(v_1) < S(v_1)$. The importance of this objective is proportional to the degree to which firm 1 cannot appropriate the social benefits of product 2, because the marginal gain in welfare from reducing the difference $S(v_1)-P(v_1)$ is proportional to this difference.

In general, then, any small policy change that increases both $S(v_1)$ and $P(v_1)$ must be desirable. To be precise, we can state the following lemma:

Lemma 3: Welfare given any particular v_1, $W(v_1)$, is nondecreasing in $S(v_1)$

---

[16]First, as long as $v_1>0$, we know $2v_1+P(v_1) > 0$. To see this result, recall that $P(v_1)$ can be no less than $-v_1$, because even the policy least favorable to firm 1 can only reduce its second-period payoff to 0. (If $v_1=0$, then $2v_1+P(v_1) = 0$ is possible.) Second, as long as there is some positive probability that firm 2 will invent product 2, we know $S(v_1)>P(v_1)$, because Lemma 1 and Lemma 2 together imply that firm 1's private benefit from the invention of product 2 must fall short of the social benefits from that invention.

and strictly increasing in $P(v_1)$. Furthermore, if either $v_1>0$ or $P(v_1)>0$, then $W(v_1)$ must be strictly increasing in $S(v_1)$ also.

1. The Second-Best Policy

Suppose that the courts can pursue the second-best policy described above in Section III; that is, they can set both (1) the prices paid by consumers, and (2) how the resulting revenue is allocated between the two firms. For any (v_1, v_2) pair, the optimal price will be the monopoly price, v_1+v_2, because there is no deadweight loss from monopoly pricing in this model. Any lower price could be raised to increase the royalties for firm 1, for example, which by Lemma 1 must improve the incentives to invent product 1. The more complex issue is the allocation of these monopoly profits between firm 1 and firm 2.

Consider the cases in which $v_2>0$.[17] Then we can always express the payoff to firm 2 as a multiple of the value of product 2. That is, let $\pi_2(v_1, v_2) = \alpha(v_1, v_2)v_2$. The payoff to firm 1, then, will be the remaining revenue: $\pi_1(v_1, v_2) = v_1 + [1-\alpha(v_1, v_2)]v_2$, which together with (5), implies that $\Delta\pi_1(v_1, v_2) = [1-\alpha(v_1, v_2)]v_2$. The court's optimal policy would consist of some $\alpha>0$ for each (v_1, v_2) pair.[18] That is, the court would choose some $\alpha>0$ for each v_2 to maximize $W(v_1)$.

We can substitute for π_2 in (2) and express $S(v_1)$ as:

[17]The case of $v_2=0$ is trivial; then innovation by firm 2 is never desirable, and the optimal policy would always discourage such R&D by ensuring that firm 2 receives no revenues in such a case: $\pi_2(v_1, 0) = 0$.

[18]If α were nonpositive, then firm 2 would never develop product 2, even if the social value v_2 were to exceed the cost c_2. Therefore, increasing α slightly above 0 could impose only a welfare gain through the incentives for firm 1 and firm 2. One can see from (2) and (3) that any (v_1, v_2) pair with $\alpha=0$ (and thus $\pi_2=0$) would be making no contribution to either $S(v_1)$ or $P(v_1)$; with any α in the (0, 1) interval (and thus with $\pi_2>0$ and $\Delta\pi_1>0$), however, it would make a positive contribution to both $S(v_1)$ and $P(v_1)$. Given Lemma 3, these effects must improve welfare.

$$S(v_1)=\int_0^{v_2}\int_0^{\alpha(v_1,v_2)v_2}\frac{v_2-c_2}{C_2}dc_2\ dF(v_2)$$
$$=\frac{1}{C_2}\int_0^{v_2}[\alpha(v_1,v_2)-\tfrac{1}{2}\alpha^2(v_1,v_2)]v_2^2\ dF(v_2), \tag{9}$$

and substitute for $\Delta\pi_1$ and π_2 in (3) and express $P(v_1)$ as:

$$P(v_1)=\int_0^{v_2}\int_0^{\alpha(v_1,v_2)v_2}\frac{[1-\alpha(v_1,v_2)]v_2}{C_2}dc_2\ dF(v_2)dv_2$$
$$=\frac{1}{C_2}\int_0^{v_2}[\alpha(v_1,v_2)-\alpha^2(v_1,v_2)]v_2^2\ dF(v_2). \tag{10}$$

Taking the derivatives of (9) and (10) with respect to α, and using (8), we find that the first-order condition for this maximization problem implies that:

$$\{[2v_1+P(v_1)](1-\alpha) + [S(v_1)-P(v_1)](1-2\alpha)\}v_2^2f(v_2)/C_1C_2 = 0. \tag{11}$$

The left-hand side of (11) represents the net marginal benefit from increasing α.

Lemma 4: For each (v_1, v_2), the optimal α is strictly less than 1 and must satisfy:

$$\alpha = [2v_1+S(v_1)]/[2v_1+2S(v_1)-P(v_1)]. \tag{12}$$

Proof: Note that the left-hand side of (11) must be negative if $\alpha \geq 1$ and $v_2>0$. Therefore, welfare must improve as α is reduced from such a value to below 1, and any $\alpha \geq 1$ cannot be optimal. Thus, α always lies strictly within the (0, 1) interval. One can see from (9) and (10), therefore, that once we exclude cases in which $v_2=0$, both $S(v_1)>0$ and $P(v_1)>0$.

Solving the first-order condition (11) for α yields (12). Note that the right-hand side of (12) must be strictly greater than ½ for any v_1, because $P(v_1)>0$, but also strictly less than 1, because $S(v_1)>P(v_1)$. Any such α would also meet the second-order condition. Taking the second derivative of $W(v_1)$ with respect to α, one can show that it must be negative for any α in the interval (½, 1). Thus, any α meeting the first-order condition is a maximum, and the optimal α must satisfy (12). ■

Note that the optimal α in (12) depends on v_1 but is independent of v_2. The marginal costs and benefits of increasing α for any given (v_1, v_2), as (11) indicates, are both simply proportional to v_2^2. Therefore, the optimal $\alpha(v_1, v_2)$ will be a function of v_1 only. We can now derive an explicit solution for the unique optimal policy $\alpha(v_1)$, which would allow firm 2 to capture most of the social value of its innovation, but no more than the total social value of the innovation:

Proposition 1: The second-best policy would give firm 2 a fraction $\alpha(v_1)$ of the social value of its invention, v_2, and give firm 1 the remaining share, where:

$$\alpha(v_1)=\frac{E(v_2^2)-2v_1C_2+\sqrt{[E(v_2^2)]^2+8E(v_2^2)v_1C_2+4v_1^2C_2^2}}{3E(v_2^2)} \tag{13}$$

for all (v_1, v_2). This $\alpha(v_1)$ must lie in the interval [⅔, 1) and is strictly increasing in v_1. If $v_1>0$, then $\alpha(v_1)$ is strictly greater than ⅔, and is strictly increasing in C_2, but strictly decreasing in $E(v_2^2)$.

Remark: Note that if $v_1=0$, then $\alpha(v_1)$ equals ⅔. The explanation for this result is as follows. If $v_1=0$, then the possibility of the second innovation is the only incentive for firm 1 to invent. Therefore, the probability of invention by firm 1 is proportional to $P(v_1)$. At $\alpha(v_1)$ = ⅔, firm 1 and firm 2 evenly divide the expected social surplus (net of firm 2's expected R&D costs, which will consume half of firm 2's expected revenues) from the second innovation, so that $P(v_1)$ = ½$S(v_1)$. As one can see from (8), this equality implies that increasing $P(v_1)$ and increasing $S(v_1)$ are equally important imperatives at this point. Furthermore, once $\alpha(v_1)$ rises to ⅔, $P(v_1)$ is declining at the same rate as $S(v_1)$ is rising. Thus, the marginal cost of further increases in α has risen to equal the (diminishing) marginal benefit.

If $v_1>0$, then as one can see from (8), increasing $S(v_1)$ is more important than increasing $P(v_1)$ at $\alpha(v_1)$ = ⅔. The marginal benefit of further increasing α above ⅔ is now greater than the marginal cost, rather than just equal to marginal cost at ⅔ (as it would if $v_1=0$). Thus, as v_1 grows larger, the marginal benefit schedule rises, and so does the optimal α. Indeed, as v_1 rises toward infinity, $P(v_1)$ makes a proportionally less significant contribution to firm 1's incentives. Then increasing $S(v_1)$ becomes far more important than increasing $P(v_1)$. As the implications for the incentives for firm 2 become relatively more important, the optimal α grows arbitrarily close to 1.

Increasing C_2 or decreasing $E(v_2^2)$ similarly increases the relative importance of v_1 in firm 1's incentives. A larger C_2 makes the second innovation less likely; a smaller $E(v_2^2)$ makes the second innovation less important. Thus, if $v_1>0$, then as $C_2/E(v_2^2)$ ranges toward infinity, the optimal α again grows arbitrarily close to 1.

Proof: See Appendix A.

Finally, note that courts can implement this second-best policy through an idealized compulsory licensing policy. Suppose the court finds that the second product infringes the first patent, but allows the second innovator to retain a subservient (and therefore complementary) patent. The court can then require each innovator to license its innovation to anyone who desires to do so at what the court determines to be a "reasonable royalty." Thus, firm 2 (or its licensee) could license the basic technology from firm 1 and market product 2. Following Tandon (1982), suppose that the court has complete discretion in setting the per-unit royalty rate for use of each innovation and thereby controls the flow of revenue received by each firm. Therefore, the court can set firm 1's royalty at $v_1 + [1-\alpha(v_1)]v_2$ and firm 2's royalty at $\alpha(v_1)v_2$ minus some infinitesimal amount. If firm 2 invents product 2, then competing firms will seek to license the complementary innovations from firm 1 and firm 2, driving down the price paid by consumers until that price just covers the cost of the royalties. Consumers would choose to buy the improved product at a price equal to v_1+v_2 minus some infinitesimal amount, and given that firm 1 would make v_1 on each such sale, firm 1 would not have any incentive to undercut this price by selling product 1 for less than v_1.

2. <u>The Third-Best Policy: Antitrust Regulation and Patent Scope</u>

I now restrict the court's control over the innovators' payoffs. Under current law, courts in general cannot resort to compulsory licensing. Suppose the court can only implement the third-best policy described above in Section III; that is, the court can affect the firms' behavior only by its decisions on infringement and on collusive licensing. The court observes v_1 and v_2, then assigns that case to one of four patent-antitrust regimes. The firms may then bargain over a licensing agreement. Let the firms evenly divide the joint surplus from any licensing agreement, collusive or otherwise; that is, the parties reach the Nash bargaining solution. The court's decisions would set the threat point for such bargaining and thereby determine both $\Delta\pi_1$ and π_2 for that (v_1, v_2) pair. Thus, the patent-antitrust policy implies a pair of functions, $\Delta\pi_1(v_1, v_2)$ and $\pi_2(v_1, v_2)$, which in turn determine the welfare function, $W(v_1)$ in (7), through $S(v_1)$ in (2) and $P(v_1)$ in (3). I shall next describe each of these four regime and the payoffs that each would imply in the second period:

<u>a. Infringement</u>: If the court finds that product 2 does infringe (let I denote this finding), then it declares firm 2's patent to be subservient to firm 1's dominant patent: firm 2 cannot sell product 2 unless firm 1 gives its consent, that is, unless firm 1 licenses its technology to firm 2. In this case, the second innovator may compensate the first for its consent through the licensing agreement. Firm 2 may also license its innovation to firm 1, which cannot use the patented improvement without the consent of firm 2. The courts also may allow the two firms to avoid competition in the product market through collusive licensing or a transfer of one of the patents (let $I \cap C$ denote this regime), or they may prohibit such collusive agreements (let $I \cap NC$ denote this regime). Under either I regime, if the firms do not agree on a licensing agreement (collusive or otherwise), firm 1 would deny its consent to firm 2 (and obtain v_1) in the second period, rather than compete with firm 2 (and obtain 0). These payoffs constitute the threat point for the bargaining over any surplus created by any licensing agreeement.

<u>(i) Infringement with No Collusion</u>: Suppose the courts prohibit both price

restrictions in patent licensing agreements and transfers that enable one party to own both patents. Nevertheless, the two firms can enjoy the maximum possible joint producer surplus: v_1+v_2. For example, as Kaplow (pp. 1860-62) notes, owners of complementary patents can cross-license and charge per-unit royalties to one another that sum to the monopoly price, v_1+v_2. This arrangement leads to the monopoly price as the equilibrium price, because only at that price will each party have no incentive to change price and thereby shift sales between the two firms: each would receive the same revenue from each unit sold regardless of which party sells the unit. Furthermore, even if the courts prohibit per-unit royalties, the two firms can reap monopoly profits. In particular, firm 2 can license its technology to firm 1, and as long as firm 1 does not also license its technology to firm 2, firm 1 can sell the improved product as a monopolist. Firm 2 cannot compete with firm 1, even if firm 1 cannot tie sales of product 1 and of product 2, because firm 2's complementary patent is subservient. In any event, the firms would choose some licensing arrangement that produces monopoly profits, then divide the surplus between them. Specifically, they divide the surplus v_2 that licensing creates beyond the v_1 that firm 1 could obtain on its own. Thus, under the I∩NC regime, firm 1 receives $v_1+\frac{1}{2}v_2$, and firm 2 receives $\frac{1}{2}v_2$.

<u>(ii) Infringement with Collusion</u>: Under the I∩C regime, the firms can agree to sell the improved product only at the monopoly price, v_1+v_2. Again, firm 1 receives $v_1+\frac{1}{2}v_2$, and firm 2 receives $\frac{1}{2}v_2$. Thus, each of the two I regimes produce the same payoffs: the first innovator can appropriate half of the value created by the improvement. The court's antitrust policy does not matter in cases of infringement, so we can refer to the I regime without specifying an antitrust policy. Under this regime, neither firm would receive the full social value generated by its innovation: firm 1 receives less under any regime, as we know from Lemma 1, and firm 2 receives less than v_2. In terms of the policy described above in Section D.1, this regime is equivalent to choosing $\alpha=\frac{1}{2}$. Compared with the second-best policy, then, firm 2 receives too small a share of v_2, and firm 1 receives one too large, and as Proposition 1 indicates, the magnitude of these deficiencies increase with v_1.

b. No Infringement: If the second product does not infringe (let NI denote this finding), it may be sold in (Bertrand) competition with the first. In this case, the court may either allow the holders of competing patents to collude through a licensing agreement or by a transfer of one of the patents (let $NI \cap C$ denote this regime) or it may prohibit such collusion (let $NI \cap NC$ denote this regime).

(i) No Infringement with No Collusion: Under the $NI \cap NC$ regime without licensing, competition yields 0 for firm 1 and v_2 for firm 2. With competition ex post, producer surplus can be no more than v_2; therefore, there is no incentive for firm 2 to agree to a licensing agreement.[19] Under this regime, then, firm 2 appropriates the full social value of its innovation, v_2, while firm 1 receives nothing. For firm 2, this regime dominates the I regime. Firm 1, however, receives the lowest possible payoff under this regime; thus, for firm 1, all other regimes dominate this regime. Furthermore, firm 1 receives less profit, and firm 2 receives more, than it would under the second-best policy.

(ii) No Infringement with Collusion: Under the $NI \cap C$ regime, the firms would agree to sell only the improved product, fix its price at the monopoly level, and divide the surplus from such a collusive agreement. Therefore, firm 1 receives $\frac{1}{2}v_1$, and firm 2 receives $\frac{1}{2}v_1+v_2$. Thus, under the $NI \cap C$ regime, the second innovator can expropriate half of the value of the first product. Firm 1 fares better under this regime than under the $NI \cap NC$ regime, but not as well as it does under the I regime. For firm 2, this regime dominates all others: firm 2 receives the highest possible payoff. Indeed, firm 2 receives more than the full social value of its innovation; under this regime, firm 2 has an excessive incentive to innovate. In

[19] Even cross-licensing with per-unit royalties cannot by itself sustain prices any higher. To suppress competition, a more elaborate agreement would be necessary. For example, an agreement can maintain monopoly prices if the royalties are based on all of a firm's sales regardless of which patent the firm uses for the particular units sold. The royalties might then be pooled and the surplus divided between the firms. Alternatively, each firm could agree to sell only its rival's product, a per-unit royalty of v_1 for product 1 could maintain monopoly prices, and the royalty for product 2 could provide firm 2 with the desired share of the monopoly profits. I assume that courts would recognize such collusive arrangements and prohibit them whenever they would prohibit explicit price-fixing.

terms of the policy described above in Section D.1, this regime is equivalent to choosing $\alpha(v_1, v_2) = 1 + v_1/2v_2$. Compared with the second-best policy, then, firm 1 receives too little profit, and firm 2 receives too much profit. For any given v_1, however, the magnitude of these deficiencies decrease with v_2.

To summarize the preceding results, the table below presents the second-period revenues for firm 1 and firm 2, respectively, as (π_1, π_2) pairs in the following payoff matrix:

	No Infringement (NI)	Infringment (I)
No License (threat point)	$(0, v_2)$	$(v_1, 0)$
No Collusion (NC)	$(0, v_2)$	$(v_1 + \frac{1}{2}v_2, \frac{1}{2}v_2)$
Collusion (C)	$(\frac{1}{2}v_1, \frac{1}{2}v_1 + v_2)$	$(v_1 + \frac{1}{2}v_2, \frac{1}{2}v_2)$

An optimal policy would assign all (v_1, v_2) pairs to one of these three regimes: I, NI∩C, or NI∩NC. Each pair would be assigned such that it could not raise $W(v_1)$ to switch it to another regime. The resulting configuration implies a particular $S(v_1)$ and a particular $P(v_1)$, with $P(v_1)$ strictly less than $S(v_1)$.

Consistent with assumption (1), let $C_2 \geq \frac{1}{2}v_1 + v_2$ for all possible v_2 (i.e., let $C_2 \geq \frac{1}{2}v_1 + V_2$). Then we may express $S(v_1)$ as follows:

$$S(v_1)=\int_I \int_0^{\frac{1}{2}v_2} \frac{v_2-c_2}{C_2} dc_2 dF(v_2) + \int_{NI\cap C} \int_0^{\frac{1}{2}v_1+v_2} \frac{v_2-c_2}{C_2} dc_2 dF(v_2) + \int_{NI\cap NC} \int_0^{v_2} \frac{v_2-c_2}{C_2} dc_2 dF(v_2) \tag{14}$$
$$=\frac{1}{C_2}\left[\int_I (\tfrac{3}{8}v_2^2) dF(v_2) + \int_{NI\cap C} (\tfrac{1}{2}v_2^2-\tfrac{1}{8}v_1^2) dF(v_2) + \int_{NI\cap NC} (\tfrac{1}{2}v_2^2) dF(v_2)\right]$$

where the probability of the second invention would depend on the regime to which it would be assigned. We must integrate with respect to c_2 over a different interval for each regime, and integrate with respect to v_2 within each regime separately. Note that the NI∩NC regime induces the optimal behavior for firm 2. Thus, the patent-antitrust policy that assigns all v_2 (from 0 to V_2) to that regime would maximize $S(v_1)$.

We may also express $P(v_1)$ as follows:

$$P(v_1)=\int_I \int_0^{\frac{1}{2}v_2} \frac{\frac{1}{2}v_2}{C_2} dc_2 dF(v_2) + \int_{NI\cap C} \int_0^{\frac{1}{2}v_1+v_2} \frac{-\frac{1}{2}v_1}{C_2} dc_2 dF(v_2) + \int_{NI\cap NC} \int_0^{v_2} \frac{-v_1}{C_2} dc_2 dF(v_2) \tag{15}$$
$$=\frac{1}{C_2}\left[\int_I (\tfrac{1}{4}v_2^2) dF(v_2) + \int_{NI\cap C} (-\tfrac{1}{4}v_1^2-\tfrac{1}{2}v_1 v_2) dF(v_2) + \int_{NI\cap NC} (-v_1 v_2) dF(v_2)\right]$$

where the probability of the second invention would depend on the regime to which it would be assigned. Again, we must integrate within each regime separately. Note that only under the I regime can firm 1 appropriate some value from the improvement; in any NI regime, the entry of firm 2 in period 2 will reduce firm 1's payoff. Thus, the patent-antitrust policy that assigns all v_2 (from 0 to V_2) to the I regime would maximize $P(v_1)$; nevertheless, in that case, $P(v_1) < V_2^2/4C_2$. Consistent with assumption (6), then, let $C_1 \geq 2V_1 + V_2^2/4C_2$. We can then derive the following propositions:

<u>Proposition 2</u>: For any $v_1 > 0$, the optimal patent-antitrust policy is infringement for v_2 below a cutoff value, which lies strictly between 0 and V_2, but noninfringement (either with collusion or without collusion) for v_2 above this cutoff value.

Remark: In general, courts should find that improvements infringe on the original patent, so as to preserve the incentives to the first inventor. Of the three regimes, this one most often provides the payoffs closest to the second-best described above in Section D.1. In terms of the second-best policy, this regime chooses the α that maximizes the contribution of the (v_1, v_2) pair to $P(v_1)$, but at the expense of making its contribution to $S(v_1)$ too low.

For a sufficiently large gain in terms of encouraging the second innovator, it will be worthwhile to curtail the first innovator's patent protection at the margin. In particular, consider the noninfringement regime without collusion as an alternative for some (v_1, v_2) pairs. This regime maximizes the contribution a pair makes to $S(v_1)$, but minimizes its contribution to $P(v_1)$. Thus, switching some (but not all) pairs to this alternative, then, would serve to offset the infringement regime's deficiencies relative to the second-best policy.

Placing especially valuable improvements outside the scope of the original patent provides a relatively large gain compared to the sacrifice in the first inventor's expected profits. The benefits (in terms of $S(v_1)$) of switching to this alternative would derive from increasing π_2 from $\frac{1}{2}v_2$ to v_2. These benefits, then, would increase with v_2 and would be independent of v_1. The costs (in terms of $P(v_1)$) would derive from reducing $\Delta\pi_1$ from $\frac{1}{2}v_2$ to $-v_1$. For any $v_1>0$, then, these costs would also increase with v_2, but at a proportionally slower rate than the benefits would increase, because the sacrifice in terms of firm 1's profits would depend on v_1 as well as v_2. For any given $v_1>0$, then, a switch to this alternative yields the greatest net benefit (i.e., at a relatively small cost) for the largest v_2.[20] Thus, it will prove optimal to allow a subsequent inventor to market a particularly valuable improvement -- one with a high v_2/V_2 ratio -- without sharing its value

[20]Conversely, for any v_1, the infringement regime will be better than this alternative for sufficiently small v_2. The same will be true of the other alternative, noninfringement with collusive licensing. See the Remark accompanying Proposition 4.

with the first inventor.[21]

Proof: See Appendix B.

Proposition 3: As v_1 approaches 0, the cutoff value for noninfringement (either with or without collusion) rises toward V_2. For $v_1=0$, the optimal patent-antitrust policy is infringement for any v_2.

Remark: Recall that under the second-best policy, the optimal share α for firm 2 increases in v_1, because the need to increase $S(v_1)$ becomes more important relative to the need to increase $P(v_1)$. Therefore, the infringement regime comes closer to the second-best policy as v_1 approaches 0. As v_1 grows smaller, there is less scope for a welfare improvement by switching any (v_1, v_2) pairs to another regime. Indeed, if $v_1=0$, then the probability of the first invention is so small that the need to improve firm 1's incentives (i.e., increase $P(v_1)$) outweighs the countervailing objective of rewarding subsequent inventors (i.e., to increase $S(v_1)$) for any v_2. Thus, even if the economic value of the second invention is much greater than the original invention, a finding of noninfringement would never improve welfare.

The most important implication of this proposition is a criterion for the court's finding of noninfringement that is more complex than that described by Merges and Nelson (1990, pp. 865-66). The decision is not simply a function of the v_2/v_1 ratio. Indeed, the cutoff value v_2 for the second product not to infringe is not even monotonically increasing in v_1. Instead, this proposition reveals that this cutoff value is decreasing in v_1, at least when v_1 is small relative to the value

[21]Thus, for any given $v_1>0$, only the most valuable improvements should avoid infringement. Note that this result emerges not because subservient improvement patents, which permit holders of dominant patents to "holdup" subsequent inventors, are a less attractive option for the courts as v_2 rises to such high levels, as Merges and Nelson (p. 866 & n.117) suggest. Proposition 1 reveals that ideally the original inventor would be entitled to the same share of v_2, whether that improvement is very valuable or not. Rather, the cutoff value for infringement emerges because the alternative (noninfringement) becomes relatively more attractive as v_2 grows larger: its social benefits grow faster than its social costs.

of possible improvements. Thus, the v_2/v_1 ratio is not sufficient to distinguish cases in which the courts should find infringement from those in which the courts should not. Instead, the optimal policy depends on both v_1/V_2 and v_2/V_2. Specifically, in cases of basic inventions with low v_1/V_2 ratios, courts should establish a very high v_2/V_2 hurdle for subsequent inventors to avoid infringement.

This result contradicts the reasoning of the Supreme Court in the Westinghouse case, in which the Court noted that although the basic patent in that case was a "pioneer" insofar as it broke new ground in the technology for train brakes, the design was not a success until later improvements perfected it. The Court cited the patent's shortcomings as a reason for a narrower range of equivalents. This proposition, however, indicates that the argument for noninfringement is weakest in just such a case: because the basic patent by itself lacked great commercial value, the original inventor should be entitled to receive a large share of the value produced by the improvements inspired by his own pioneering invention. Only the most extraordinary improvement among the universe of possible improvements -- the improvement that contributes such value that an invention of its importance would be deemed very unlikely ex ante -- should be permitted to escape infringement of such a pioneer invention.

Proof: See Appendix B.

Proposition 4: For sufficiently small v_1, collusive licensing agreements will be optimal for all noninfringement cases. For sufficiently large v_1, including any $v_1 > v_2$, however, it will be optimal to prohibit collusive licensing agreements in all noninfringement cases.

Remark: Switching a case from infringement to noninfringement with collusion can increase the contribution a (v_1, v_2) pair makes to $S(v_1)$, but must reduce its contribution to $P(v_1)$. For any v_1, this alternative is most attractive when v_2 is large. Thus, switching some (but not all) of the largest v_2 values to this alternative, just as switching some to noninfringement without collusion, can serve to offset the infringement regime's deficiencies relative to the second-best policy.

Noninfringement with collusion, however, provides firm 2 with excessive rewards that may encourage inefficient invention, simply to threaten entry and share in firm 1's revenues.[22] This excess incentive grows as v_1 grows larger. Indeed, if $v_2 < v_1$, then this alternative would be unambiguously inferior to the infringement regime: both $S(v_1)$ and $P(v_1)$ would be lower. That is, this alternative would not only reduce firm 1's profits and incentives to invent, but also distort firm 2's incentives even more than the infringement regime would. If the improvement is less valuable than the original invention, then the overinvestment by firm 2 under that alternative regime would be larger than the underinvestment by firm 2 under the infringement regime.

Nevertheless, for small v_1, noninfringement with collusion can pose less of a threat to the profits for firm 1 than noninfringement with competition does. Although competition gives the second inventor the full surplus from product 2, and so maximizes the contribution of a (v_1, v_2) pair to $S(v_1)$, it also erodes the incentives for the first invention more (than noninfringement with collusion does) in the event of the second invention. If v_1 is large relative to v_2, on the other hand, then the excess incentive for firm 2 under the collusion regime would be significant: it could increase the probability of entry by firm 2 by such a large

[22]In the Green and Scotchmer (1990) model, the courts can always avoid this problem. In their model, the second innovator is uncertain of the value of its invention, and the distribution of this variable is the <u>same</u> for all inventors of improvements who could be subject to the patent policy. Therefore, even with collusive licensing, the courts can adjust the patent scope so that the expected payoff to the second innovator is always exactly equal to the expected value of the second innovation. Their cutoff function can optimize incentives to the second innovator simply by randomizing between infringement (and partial appropriation by the first innovator) and noninfringement (which allows the second innovator to expropriate some value created by the first innovation through a collusive agreement with the first innovator).

In contrast, the model in this essay incorporates the possibility that second innovators face <u>different</u> prospects and thus explores the implications of inefficient entry. The Kaplow (1984) model addresses the problem of wasteful R&D by imitators who "invent around" the original patent, but does not introduce the possibility of "imitators" who contribute valuable improvements. Nor does he include patent scope as a policy instrument, which allows courts to decide whether imitators have competing patents or merely complementary patents that infringe on the original patent.

factor that the regime with collusion erodes firm 1's expected profits more than the regime with competition.

The relative merits of the two noninfringement regimes, then, will vary significantly with v_1. If the first invention is not so valuable, then the risk of inefficient R&D by firm 2 is relatively insignificant, and the need to maintain the firm 1's incentives (i.e., increase $P(v_1)$) is great relative to the need to optimize firm 2's incentives (i.e., increase $S(v_1)$). Therefore, collusion is optimal in such noninfringement cases. If the first invention is valuable relative to improvements, however, then the risk of inefficient entry is significant, and the first inventor's incentives are less important. Therefore, competition is optimal in such noninfringement cases.

Proof: See Appendix B.

Proposition 5: As v_1 grows very large relative to V_2, the cutoff value for noninfringement (without collusion) rises toward V_2.

Remark: For any given policy, as v_1 grows larger, the need to increase $S(v_1)$ tends to grow relative to the need to increase $P(v_1)$. This effect tends to make the noninfringement regime (without collusion) attractive relative to the infringement regime, because it gives firm 2 better incentives. A second effect, however, opposes the first: an increase in v_1 implies that such a noninfringement finding entails a greater sacrifice in firm 1's incentives, and thus $P(v_1)$ would decline (whereas $S(v_1)$ would be unaffected) insofar as patent policy includes the noninfringement regime. This increase in the distortion of firm 1's incentives tends to make the need to increase $P(v_1)$ more important relative to the need to increase $S(v_1)$, which in turn makes the infringement regime (which better protects firm 1's profits) more attractive. Furthermore, if v_1 is large, then so is the cost of a noninfringement finding in terms of $P(v_1)$, relative to the corresponding gain in terms of $S(v_1)$. Therefore, if v_1 is large, then the second effect dominates the first effect: the noninfringement regime becomes less attractive as v_1 grows still larger.

In the limit, the value of all possible improvements becomes trivial in comparison with the original invention. In this case, the social costs of diminishing firm 1's incentives outweigh the social benefits of encouraging improvements, even those with high v_2/V_2 ratios. If v_1/V_2 is large, then the optimal policy would confront firm 1 with only a small ex ante probability of noninfringement by setting a high v_2/V_2 hurdle for firm 2.

Proof: See Appendix B.

E. CONCLUDING REMARKS

Proposition 2 supports the suggestion that courts respond to the holdup problem in cumulative innovation by placing particularly valuable improvements outside the scope of the original patent, at least as long as the original product has some positive stand-alone value. Thus, the model offers some basis in economic theory for the courts' reliance on the value of the inventions in infringement cases. Courts, however, should not find that any improvement avoids infringement. Given the option of subservient improvement patents, courts should be circumspect in allowing imitators to "invent around" a patent. Thus, the optimal patent policy may entail broader or narrower patent protection than courts currently provide.

Although the model supports the general practice of declaring improved products not "equivalents" of patented predecessors, the model also suggests a refinement in how courts take the value of these inventions into account. In particular, patents should provide the broadest protection to two distinct classes of basic inventions: not only those that are very valuable relative to possible improvements (Proposition 5), but also those that have very little value relative to the improvements that one would expect to follow (Proposition 3). Courts should grant a broad range of equivalents for inventions falling toward either extreme; if instead a basic invention would be expected to generate improvements with values on the same order of magnitude as the value of the basic invention, then it should receive relatively narrow patent protection.

That is, contrary to the Supreme Court's suggestion in the Westinghouse case, courts should recognize an invention with little stand-alone value as a pioneer if it provides "stepping stones" for others to render the invention much more valuable. Thus, the definition of a pioneer invention should not turn simply on the value it has standing alone. Although a basic invention may have trivial value by itself, it may also be a technological breakthrough in that it generates great spillovers in the form of improvements likely to be far more valuable than the basic invention itself. Like basic research, such pioneer inventions present a strong case for some form of subsidy because a private inventor is unlikely to undertake such R&D at levels commensurate with their social value. The doctrine of equivalents, then, can help such pioneer inventors appropriate the external benefits of their research.

The foregoing discussion presumes that the original invention is a necessary condition for the allegedly infringing innovation that follows. As the discussion in Section D indicated, if the patented innovation does nothing to facilitate the subsequent innovation, then a finding of noninfringement (with collusion prohibited) would be appropriate. This reasoning lends support both for the general policy against collusion among holders of competing patents and for the critique by Merges and Nelson (pp. 903-04, 914-15) of cases like Scripps Clinic and Research Foundation v. Genentech, Inc., 666 F. Supp. 1379 (N.D. Cal. 1987). In that case, Genentech had invented a recombinant DNA method for producing a human blood-clotting protein with major advantages over an earlier technique of purifying the protein from natural blood. Scripps, however, held a product patent based on the older technology, and the court held that Genentech had infringed the earlier patent. Although the patentees in such cases have in reality invented processes for purification of substances that occur in nature, they seek and often obtain product patents. See Chisum (Vol. 1, § 1.02[9]). These patents pose a holdup problem even for those who invent entirely different (and better) processes for producing the same purified substance.[23] Such product patents inhibit R&D

[23]The appropriate policy for such innovations, which are independent of the preceding substitute innovations, would appear to be the awarding process patents rather than product patents. Merges and Nelson support application of the doctrine of equivalents

in new biotechnology like recombinant DNA research, because the new processes will be held to infringe the prior patents. Furthermore, because such new techniques do not build on the older patented techniques, there is no reason that holders of patents on those older technologies should profit from such unrelated subsequent research. The holdup problem in such cases is unilateral, not bilateral: it could be eliminated entirely by assigning control over the new technology to the second innovator alone. Patents should protect against imitators who would otherwise free-ride on the work of patentees, not independent innovators who could not possibly gain any benefit from the patentee's work.

Finally, the model provides very little support for lax antitrust scrutiny of collusion between patentees, even though the model excludes considerations of deadweight loss. Such collusion would be not only undesirable outside of the context of cumulative innovation, as in the Genentech situation, but also unnecessary between holders of dominant and subservient patents. Collusion between holders of competing patents, even in the context of cumulative innovations which Green and Scotchmer discuss, would be desirable only in limited circumstances. Once we recognize the potential for inefficient entry, a policy favoring collusion proves to be an unappealing alternative. Proposition 4 indicates that such a policy would only be superior to the alternatives only if v_2 is large, both relative to v_1 (so that the risk of inefficient entry is relatively small) and relative to most other possible improvements (so as to justify a finding of noninfringement, which should be rare when v_1 is small).

It is difficult to imagine an antitrust defense tailored to these circumstances. If a court would find these cases at all difficult to identify, the best policy would be a uniform rule against collusion. After all, in the circumstances described by Proposition 4, noninfringement without collusion is likely to be an adequate substitute for noninfringement with collusion: if v_1 is small relative to v_2, then the improvements in the two inventors' payoffs when they collude will be

after the fact to achieve the same reduction in patent scope.

correspondingly small. Collusive licensing, then, seems to be a poor response to the problem of cumulative innovation.

These conclusions, however, should be qualified: the model presented here include many simplifying assumptions to render the analysis tractable. Further research is necessary to determine how robust the results in this essay are when these assumptions are relaxed. Given the many complexities excluded from the model, one must be cautious in drawing implications for public policy. Even given this uncertainty, however, the preliminary analysis undertaken here does contribute some basic insights that one would expect to survive further study. My conjecture is that most plausible extensions of the model would preserve the two main results of this essay: (1) if an invention has very little stand-alone value relative to most possible improvements that others may subsequently invent, this fact militates in favor of broad patent protection;[24] and (2) collusive agreements between holders of competing patents create incentives for inefficient entry by imitators who "invent around" the original patent.[25]

For example, the model assumes that the parties litigate and do not settle

---

[24]Of course, specific information about the distribution of R&D costs would imply special results. If a class of patents were known to generate sufficient revenues standing alone to cover their R&D costs nearly always, then extra royalties derived from anticipated improvements would not induce much additional valuable R&D, and narrower patent protection for that class of inventions may be in order. It would be difficult, however, to make this inference from historical data on the cost of R&D investments. Because the success of any given R&D project is uncertain, inventors must be rewarded by a risk premium to compensate for the possibility of failure. Thus, one would need information on probability of success and on risk aversion, as well as actual costs, to calculate the revenue necessary to provide an adequate reward. Moreover, because R&D investments are endogenous, one would never expect to observe R&D costs regularly in excess of the stand-alone value v_1 unless the existing patent system already ensured broad protection for the resulting patent.

[25]I am less confident about some of the other specific results of the model. For example, Proposition 2 may be sensitive to the assumption that c_1 and c_2 are distributed uniformly. It appears possible that the model would not produce a cutoff function as the optimal patent policy if c_2 is less likely to take on its highest values: then increasing the reward for the second inventor may be particularly unlikely to induce additional innovation in cases of high v_2, and findings of infringement may prove optimal instead in those cases.

out of court. Insofar as litigation costs are symmetric, however, settlements should reflect the expected outcome of trials, and the ex ante incentives should be the same. Similarly, the model assumes that courts have full information and make no errors. Even with judicial error, however, courts could create the same incentives ex ante as long as they offer the firms the correct payoffs on average. Indeed, to the extent there is a stochastic element in infringement decisions, courts can attempt to do better than the third-best policy described in Section D.2: they can pursue a randomized "mixed strategy" that better approximates the expected payoffs of the ideal compulsory licensing policy described in Section D.1.[26]

Settlements may not reflect the expected outcome of a trial, however, to the extent that settlements allow collusion that would not be permitted otherwise. Antitrust scrutiny of such settlements, then, should ensure that patentees with substitute (rather than complementary) innovations do not fix prices or otherwise collude. To the extent that rival patents actually derive from complementary innovations, particularly those that arise in the context of cumulative innovation, the patentees should be allowed to cross-license with per-unit royalties. This arrangement may lead to monopoly pricing, but as Section D.1 revealed, the optimal policy would allow such innovators to maximize their joint producer surplus. To the extent that courts find complementary and competing patents difficult to distinguish, however, antitrust policy should not be so permissive as to allow explicit price-fixing. See Kaplow (1984, pp. 1860-62) and Priest (1977, p. 358).

This model also makes the strong assumption that competing duopolists would engage in Bertrand competition. One could relax this assumption, so that

[26]This task is easier if the first innovation is necessary for the second innovation, as the model in this essay assumes. In reality, courts may be uncertain whether the second innovation required the first. One could extend the model to consider other possible relationships between the innovations: for example, the second innovation may have occurred earlier than it would otherwise because of the first, or it may have cost less because of the first. To the extent a finding of infringement leads to excessive rewards for the first innovator in these circumstances, a finding of noninfringement may be preferable.

competing patentees could more closely approximate a collusive outcome without any formal agreement. In that case, findings of noninfringement (even with collusive agreements prohibited) would become less attractive because they would entail the problems of inefficient entry posed by collusion, which would be greatest in the case of a patent with a relatively large v_1. Thus, introducing more imperfect competition into the model would have the primary effect of bolstering the case for broad patent protection for inventions with relatively large stand-alone values.

A complete analysis of the effects of monopoly pricing would include the deadweight loss created by pricing above marginal cost. One could relax the assumption of unit demands,[27] address the case of downward-sloping demand curves, and thereby introduce the deadweight loss from patent monopolies. It is ambiguous, however, whether this change in the model would make collusive pricing more or less attractive. Of course, the deadweight loss militates in favor of competition and lower prices. Downward-sloping demand curves, however, would also imply consumer surplus even at monopoly prices, which in turn would imply that innovators always reap too little profit to encourage all socially desirable R&D. This consideration militates in favor of monopoly pricing.[28]

Finally, this paper includes assumptions in the model that ensure that patents tend to offer inventors too little incentive to invest in R&D, in the belief that this problem is most important. For example, whereas the model assumed that

---

[27]This simplifying assumption allowed us to focus on the trade-off between the incentives to the two innovators. This assumption may be reasonable for a process innovation that reduces the cost of producing a given product. Although this essay describes the model as one of product patents, the analysis applies to process patents as well. The assumption of unit demands will also approximate reality whenever consumer surplus and deadweight loss are small, which may be the case when producers can price discriminate among consumers or when the market demand curve is concave.

[28]The optimal balance between these considerations would depend on the particular shape of the demand curve, the nature of the improvement, and other factors that would be difficult for a court to observe. Further research may prove useful. The simple model in this essay, however, may be a better guide to policy than a more complex model that implies a different result only in special cases that would be difficult to distinguish from the usual case.

there is no rivalry in R&D, the literature using "patent race" models indicate that rivalry in R&D can lead to excessive R&D expenditures. See Reinganum (1989). To include this effect of patent races, one might modify the model in this essay so that both firms participated in both stages of innovation simultaneously rather than inventing in sequence.[29] Rivalry in R&D would tend to offset the problem of underinvestment in cumulative innovation, and would thereby bolster the case against collusion among patentees.

Furthermore, I would expect that such a model of rivalrous R&D would introduce other considerations in support of broad patent protection for "stepping stone" innovations. An inventor engaged in a patent race would be reluctant to patent such an innovation in the absence of broad protection, because a rival inventor could improve upon the imperfect technology disclosed by the patent and claim the reward generated by the ultimate success of this line of research. If others can "invent around" patents too easily, an inventor would prefer to keep a design secret in its intermediate stages to increase the chances of perfecting the technology first.[30] This secrecy would promote duplicative research by others, who must waste resources to "reinvent the wheel." Whereas narrow patent protection would inhibit the disclosure of discoveries useful to others, broad patent protection would encourage the dissemination of this information, thereby opening the field for others to invent improvements. Thus, broad patent protection would serve the disclosure objectives of the patent system, encourage licensing, and thereby mitigate the "winner take all" aspect of patent races that can induce excessive levels of R&D investment. At the same time, rapid dissemination of new

---

[29]Gillespie (1990) employs such a model to study the impact of patent scope upon R&D, but in his model the second innovation is simply an imitation of the first and features no valuable improvements.

[30]Levin, Klevorick, Nelson, and Winter (1987, pp. 794-95) surveyed R&D managers and found that the respondents considered secrecy to be more effective than process patents in protecting the returns to such innovation. The most important weakness of patents, according to the respondents, was the ease with which rivals could invent around a patent. See Levin, et al. (pp. 802-03).

technology increases the probability of quick and successful perfection of the technology. Thus, broad patent protection for inventions with little stand-alone value (but with great potential for valuable improvements) could promote more efficient R&D investments without reducing the pace of innovation.[31]

[31]Further research could also explore whether the results in this model generalize to a sequence of innovations in an n-stage model. I would expect the two-stage model in this essay to capture the considerations that would prove relevant between each pair of stages in a longer sequence of innovations. Other extensions that could qualify the results in this essay include horizontal product differentiation, see Klemperer (1990) and Waterson (1990), ex ante agreements to form research joint ventures, see Green and Scotchmer (1990), and patent length as another policy instrument, see Gilbert and Shapiro (1990), Klemperer (1990), and Tandon (1982).

APPENDIX A

Proof of Proposition 1: The expressions involving α in (9) and (10) may be brought outside the integrals, because α is independent of v_2. In each equation, what is left of the integral is simply $E(v_2^2)$. Substituting these revised expressions for $S(v_1)$ and $P(v_1)$ in (12), and rearranging terms, yields the following quadratic equation:

$$3E(v_2^2)\alpha^2 + [4v_1C_2 - 2E(v_2^2)]\alpha - 4v_1C_2 = 0. \quad \text{(A.1)}$$

Solving for α using the quadratic formula, and excluding any solutions for α that are negative, we obtain (13) as the unique solution for $\alpha(v_1)$.

Total differentiation of (A.1) yields:

$$[(6\alpha-2)E(v_2^2) + 4v_1C_2]\, d\alpha + [4C_2(\alpha-1)]\, dv_1$$
$$+ [4v_1(\alpha-1)]\, dC_2 + [\alpha(3\alpha-2)]\, dE(v_2^2) = 0, \quad \text{(A.2)}$$

which implies our comparative statics results. The partial derivative of α with respect to v_1 is strictly positive, because we know from the proof of Lemma 4 that $\frac{1}{2}<\alpha<1$. Similarly, the partial derivative of α with respect to C_2 is nonnegative, and strictly positive if $v_1>0$. For $v_1>0$, α must be greater than $\frac{2}{3}$, because $\alpha'(v_1) > 0$ and (13) implies that if $v_1=0$, then $\alpha=\frac{2}{3}$. Therefore, the partial derivative of α with respect to $E(v_2^2)$ is nonpositive, because $\alpha \geq \frac{2}{3}$, and is strictly negative if $v_1>0$, because then $\alpha>\frac{2}{3}$. ■

APPENDIX B

In this appendix, I prove the propositions presented in Section D.2. To characterize the optimal policy, it is useful to describe first the regions in (v_1, v_2) space where each regime dominates the others. I compare the three regimes in a series of three pairs.

First, given the optimal policy, consider the hypothetical change in welfare resulting from switching a point in (v_1, v_2) space from I to NI∩NC. Switching one such point in (14) would entail a $dS(v_1)$ proportional to $\frac{1}{8}v_2^2f(v_2)/C_2$, and in (15), a $dP(v_1)$ proportional to $-(\frac{1}{4}v_2^2+v_1v_2)f(v_2)/C_2$. Therefore, we know from (8) that $dW(v_1)$ would be positive or negative as:

$$\tfrac{1}{8}[2v_1+P(v_1)]v_2^2 - [S(v_1)-P(v_1)](\tfrac{1}{4}v_2^2+v_1v_2)$$

is positive or negative. Excluding points where $v_2=0$ (and thus where $dW(v_1)=0$), then, $dW(v_1)$ would be positive or negative as:

$$2v_1v_2 - [2S(v_1)-3P(v_1)]v_2 - 8[S(v_1)-P(v_1)]v_1 \qquad (B.1)$$

is positive or negative. We can see from (14) and (15) that $0 \leq 2S(v_1)-3P(v_1)$. Thus, if $2v_1 \leq 2S(v_1)-3P(v_1)$, then (B.1) implies that $dW(v_1)<0$ for any $v_1>0$, and I will strictly dominate NI∩NC for any v_2. If $2v_1 > 2S(v_1)-3P(v_1)$ instead, then (B.1) implies that NI∩NC will dominate I for any:

$$v_2 > 8[S(v_1)-P(v_1)]v_1/\{2v_1-[2S(v_1)-3P(v_1)]\}, \qquad (B.2)$$

and I will dominate NI∩NC if the inequality is reversed.

Second, consider the hypothetical change in welfare resulting from switching a point in (v_1, v_2) space from I to NI∩C. Switching one such point in (14) would entail a $dS(v_1)$ proportional to $\frac{1}{8}(v_2^2-v_1^2)f(v_2)/C_2$, and in (15), a $dP(v_1)$ proportional to $-(\frac{1}{4}v_1^2+\frac{1}{4}v_2^2+\frac{1}{2}v_1v_2)f(v_2)/C_2$. Therefore, we know from (8) that $dW(v_1)$ would be positive or negative as:

$$\tfrac{1}{8}[2v_1+P(v_1)](v_2^2-v_1^2) - [S(v_1)-P(v_1)](\tfrac{1}{4}v_1^2+\tfrac{1}{4}v_2^2+\tfrac{1}{2}v_1v_2)$$

is positive or negative. Excluding the origin, where $v_1+v_2 = 0$ (and thus where $dW(v_1)=0$), then, $dW(v_1)$ would be positive or negative as:

$$2v_1v_2 - 2v_1^2 - [2S(v_1)-P(v_1)]v_1 - [2S(v_1)-3P(v_1)]v_2 \qquad (B.3)$$

is positive or negative. We can see from (14) and (15) that $0 < 2S(v_1)-P(v_1)$.

Thus, if $2v_1 \leq 2S(v_1)-3P(v_1)$, then (B.3) implies that $dW(v_1)<0$ for any $v_1>0$, and I will strictly dominate NI∩C for any v_2. If $2v_1 > 2S(v_1)-3P(v_1)$ instead, then (B.3) implies that NI∩C will dominate I for any:

$$v_2 > \{2v_1^2+[2S(v_1)-P(v_1)]v_1\}/\{2v_1-[2S(v_1)-3P(v_1)]\}, \qquad \text{(B.4)}$$

and I will dominate NI∩C if the inequality is reversed.

Finally, consider the hypothetical change in welfare resulting from switching a point in (v_1, v_2) space from NI∩NC to NI∩C. Switching one such point in (14) would entail a $dS(v_1)$ proportional to $-\frac{1}{8}v_1^2f(v_2)/C_2$, and in (15), a $dP(v_1)$ proportional to $(\frac{1}{2}v_1v_2-\frac{1}{4}v_1^2)f(v_2)/C_2$. Therefore, we know from (8) that $dW(v_1)$ would be positive or negative as:

$$-\tfrac{1}{8}[2v_1+P(v_1)]v_1^2 + [S(v_1)-P(v_1)](\tfrac{1}{2}v_1v_2-\tfrac{1}{4}v_1^2)$$

is positive or negative. Excluding $v_1=0$ (and thus where $dW(v_1)=0$), then, $dW(v_1)$ would be positive or negative as:

$$4[S(v_1)-P(v_1)]v_2 - 2v_1^2 - [2S(v_1)-P(v_1)]v_1 \qquad \text{(B.5)}$$

is positive or negative. Thus, NI∩C will dominate NI∩NC for any:

$$v_2 > \{2v_1^2+[2S(v_1)-P(v_1)]v_1\}/4[S(v_1)-P(v_1)], \qquad \text{(B.6)}$$

and NI∩NC will dominate NI∩C if the inequality is reversed.

Given this information, we can now derive the following propositions:

<u>Proof of Proposition 2</u>: For any $v_1>0$, we know the optimal policy must be the I regime for some positive v_2 sufficiently close to 0. Consider the limit of (B.1) or (B.3) for any v_1 as v_2 grows arbitrarily small. These limits are negative: therefore, a switch from either NI∩NC or NI∩C to I would improve welfare. That is, the I regime dominates either alternative for sufficiently small v_2.

Nevertheless, for any $v_1>0$, the optimal policy cannot rely on the I regime exclusively. Otherwise, we know from (14) and (15) that $S(v_1)-P(v_1) < V_2^2/8C_2$, and $2S(v_1)-3P(v_1) = 0$. These relationships, together with (B.1), would imply that for any $v_1>0$ a shift from I to NI∩NC for a v_2 sufficiently close to V_2 will yield a welfare improvement, which would contradict the supposed optimum. (Recall that $C_2>V_2$.) Therefore, for any $v_1>0$, the optimal policy must instead feature at least one cutoff function, either the critical v_2 in (B.2) or that in (B.4), strictly between $v_2=0$ and $v_2=V_2$ to separate the infringement from the noninfringement regime■

<u>Proof of Proposition 3</u>: We know that $2v_1 > 2S(v_1)-3P(v_1)$ for all $v_1>0$. If we were to suppose otherwise, then both (B.1) and (B.3) would be negative for all $v_1>0$, as a comparison of (14) and (15) confirms. If both (B.1) and (B.3) were negative, however, then a shift for <u>any</u> $v_2>0$ from either NI∩NC or NI∩C to I would yield a welfare improvement, which would contradict Proposition 2.

We also know from Proposition 2, (14), and (15) that $2S(v_1)-3P(v_1) > 0$ for any $v_1>0$. As v_1 becomes arbitrarily close to 0, then, $2S(v_1)-3P(v_1)$ must stay between $2v_1$ and 0, which can occur only if fewer and fewer v_2 values are assigned to either NI∩NC or NI∩C. Therefore, the lowest cutoff value for noninfringement, either the critical v_2 in (B.2) or that in (B.4), must rise toward V_2 as v_1 approaches 0.

If $v_1=0$, then no $v_2>0$ can be assigned to either NI∩NC or NI∩C. If we were to suppose otherwise, then we know from (14) and (15) that $0 < 2S(0)-3P(0)$. This inequality, together with (B.1) and (B.3) evaluated at $v_1=0$, imply that a shift for any $v_2>0$ from either NI∩NC or NI∩C to I would yield a welfare improvement, which would contradict the supposed optimum. (Note that at the origin in (v_1, v_2) space, policy is irrelevant, because any one of the three regimes would yield the same welfare ex post: zero.) ■

<u>Proof of Proposition 4</u>: Consider the critical value for v_2 in (B.6), which denotes indifference between NI∩C and NI∩NC. This critical value for v_2 falls to 0 as v_1 grows arbitrarily close to 0. Furthermore, Proposition 3 implies that the lowest cutoff value for noninfringement -- either that in (B.2) or that in (B.4) -- grows close to V_2 (and is therefore bounded away from $v_2=0$) as v_1 approaches 0. Therefore, for sufficiently small v_1, all noninfringement cases fall in the region where NI∩C will dominate NI∩NC.

Nevertheless, NI∩NC must dominate NI∩C for larger values of v_1. In particular, consider (B.3) where $v_2=v_1$. Note that (B.3) must be negative along the $v_2=v_1$ line. Therefore, the cutoff function for NI∩C in (B.4) must always lie strictly above this line. For any $v_1 \geq v_2$, then, collusive licensing cannot be optimal in noninfringement cases. For any $v_1>V_2$, for example, courts should prohibit collusive licensing agreements in all cases of noninfringement. ■

Proof of Proposition 5: Proposition 4 implies that for large v_1, we can exclude the NI∩C regime from (14) and (15). In these cases, the critical v_2 in (B.2) is the relevant cutoff value for noninfringement. Let $v_2^*(v_1)$ denote this cutoff function. Thus, in our expressions for $S(v_1)$ in (14) and for $P(v_1)$ in (15), we would integrate with respect to v_2 from 0 to $v_2^*(v_1)$ for the I regime and from $v_2^*(v_1)$ to V_2 for the NI∩NC regime. Although we cannot solve explicitly for $v_2^*(v_1)$, we can substitute our new expressions for $S(v_1)$ and $P(v_1)$ into (B.2). Multiplying both numerator and denominator by C_2/v_1, we derive the following condition:

$$v_2^*(v_1)=\frac{\int_0^{v_2^*(v_1)} v_2^2 dF(v_2)+\int_{v_2^*(v_1)}^{V_2}(4v_2^2+8v_1v_2)dF(v_2)}{2C_2-\int_{v_2^*(v_1)}^{V_2}[(v_2^2/v_1)+3v_2]dF(v_2)}. \qquad \text{(B.7)}$$

Consider the right-hand side of (B.7) and suppose that $v_2^*(v_1)$ is simply a constant. As v_1 grows larger, both the numerator and the denominator tend to increase. For a sufficiently large v_1, however, the numerator will increase by a greater proportion than the denominator as v_1 increases; that is, the right-hand side of (B.7) is increasing in v_1. Then the equality (B.7) cannot hold if $v_2^*(v_1)$ is constant as v_1 increases. Nor can (B.7) hold if $v_2^*(v_1)$ is decreasing in v_1, for this would only cause the right-hand side of (B.7) to rise faster: $v_2^{*\prime}(v_1)<0$ would add another positive effect upon the numerator and a negative effect upon the denominator. Instead, for sufficiently large v_1, $v_2^*(v_1)$ must increase in v_1, which will dampen the increase in the right-hand side of (B.7). Proposition 2 implies that $v_2^*(v_1)$ must remain below V_2, and to keep the right-hand side of (B.7) bounded as v_1 increases without bound, $v_2^*(v_1)$ must approach V_2 asymptotically. ■

REFERENCES

Areeda, Phillip and Louis Kaplow (1988), Antitrust Analysis, 4th ed. (Boston: Little, Brown & Co.).

Areeda, Phillip and Donald F. Turner (1978), Antitrust Law, Vol. 3 (Boston: Little, Brown & Co.).

Caves, Richard E., Harold Crookell, and J. Peter Killing (1983), "The Imperfect Market for Technology Licenses," Oxford Bulletin of Economics and Statistics, Vol. 45, 249-267.

Chisum, Donald S. (1990), Patents (New York: Matthew Bender & Co., Inc.).

Gilbert, Richard and Carl Shapiro (1990), "Optimal Patent Length and Breadth," RAND Journal of Economics, Vol. 21, 106-112.

Gillespie, William H. (1990), "The Influence of the Patent System on Research and Development," Yale University, Ph.D. dissertation.

Green, Jerry R. and Suzanne Scotchmer (1990), "Antitrust Policy, the Breadth of Patent Protection and the Incentive to Develop New Products," University of California at Berkeley, John M. Olin Working Paper in Law and Economics No. 90-12.

Grossman, Sanford J. and Oliver D. Hart (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," Journal of Political Economy, Vol. 94, 691-719.

Kaplow, Louis (1984), "The Patent-Antitrust Intersection: A Reappraisal," Harvard Law Review, Vol. 97, 1813-1892.

Klein, Benjamin, Robert G. Crawford, and Armen A. Alchian (1978), "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," Journal of Law and Economics, Vol. 21, 297-326.

Klemperer, Paul (1990), "How Broad Should the Scope of Patent Protection Be?", RAND Journal of Economics, Vol. 21, 113-130.

Laffont, Jean-Jacques and Jean Tirole (1990), "The Politics of Government Decision Making: Regulatory Institutions," Journal of Law, Economics, and Organization, Vol. 6, 1-31.

Landes, William M. and Richard A. Posner (1989), "An Economic Analysis of

Copyright Law," Journal of Legal Studies, Vol. 18, 325-363.

Levin, Richard C., Alvin K. Klevorick, Richard R. Nelson, and Sidney G. Winter (1987), "Appropriating the Returns from Industrial Research and Development," Brookings Papers in Economic Activity, No. 3:1987, 783-820.

Lipscomb, Ernest B. (1987), Walker on Patents, 3d ed., Vol. 6 (Rochester, N.Y.: Lawyers Cooperative Publishing Co.).

Menell, Peter S. (1989), "An Analysis of the Scope of Copyright Protection for Application Programs," Stanford Law Review, Vol. 41, 1045-1104.

Merges, Robert P. and Richard R. Nelson (1990), "On the Complex Economics of Patent Scope," Columbia Law Review, Vol. 90, 839-916.

Priest, George L. (1977), "Cartels and Patent License Arrangements," Journal of Law and Economics, Vol. 20, 309-377.

Reinganum, Jennifer F. (1989), "The Timing of Innovation: Research, Development, and Diffusion," Ch. 14 of Handbook of Industrial Organization, Vol. 1, edited by Schmalensee, Richard and Robert D. Willig (New York: North-Holland).

Scherer, F.M. (1980), Industrial Market Structure and Economic Performance, 2d ed. (Boston: Houghton Mifflin Co.).

Scotchmer, Suzanne (1991), "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law," Journal of Economic Perspectives, Vol. 5, 29-41.

Sullivan, Lawrence A. (1977), Antitrust (St. Paul, Minn.: West Publishing Co.).

Tandon, Pankaj (1982), "Optimal Patents with Compulsory Licensing," Journal of Political Economy, Vol. 90, 470-486.

Waterson, Michael (1990), "The Economics of Product Patents," American Economic Review, Vol. 80, 860-869.

Williamson, Oliver E. (1979), "Transaction-Cost Economics: The Governance of Contractual Relations," Journal of Law and Economics, Vol. 22, 233-261.

CHAPTER II:

BARGAINING AND THE DIVISION OF VALUE IN CORPORATE REORGANIZATION

Lucian Arye Bebchuk and Howard F. Chang

A. INTRODUCTION

Reorganization is one of the two alternatives open to an insolvent corporation under the Bankruptcy Code. A bankrupt corporation may file either for liquidation under Chapter 7 or for reorganization under Chapter 11 of the Bankruptcy Code.[1] Upon a filing for liquidation, a court immediately appoints a trustee to sell the firm's assets, either piecemeal or as a going concern, to outside buyers. The proceeds from this sale are divided among those who have rights against the corporation, with the division made according to the ranking of these rights by legal priority.

The firm may instead go into Chapter 11, under which the firm can be "reorganized." In reorganization there is no sale to third parties. Rather, there is a "hypothetical sale" of the firm to the existing "participants" -- all those who hold claims or rights against the insolvent company. These participants surrender their claims and rights in exchange for claims and rights against the new corporation. For example, a bankrupt company may emerge from reorganization with all its debt cancelled and with the former debtholders holding some or all of the equity of the reorganized company.

If the court supervising a reorganization could observe the value of the reorganized firm, it would allocate that value among the participants according to the legal priority of their claims. For example, consider a company that has equityholders and that owes debtholders $200, and suppose that the court observes that the value of the reorganized company will be $150; in this case, the court would order that all of the reorganized company's securities be given to the debtholders, and none be given to the equityholders. Below we will use the term "contractual right" to mean that which a class would receive if the bankruptcy court could observe the firm's value and distributed it among the classes according to the

[1]A firm may commence a voluntary bankruptcy case by filing a petition with a bankruptcy court under section 301 of the Bankruptcy Code. For a general discussion of both bankruptcy proceedings, see White (1984, 1989). For a review of the law on business reorganizations, see Trost (1979).

initial contracts, i.e., strictly in the order of the legal priority of their claims.

Because a court cannot determine accurately an objective figure for this value, however, the law leaves the division of the reorganized company's value to a process of bargaining among the classes of participants. Under section 1129(a)(8) of the Bankruptcy Code, each class of equityholders and debtholders whose interests are impaired must vote to approve a reorganization plan, which would include a division of value. The approval by a class requires (as specified by section 1126 of the Bankruptcy Code) a certain majority of the class members to vote in favor of the plan. As is generally believed by participants in reorganization and as will be shown in this essay, the outcome of this bargaining process often diverges from the contractual rights of the classes.

Section B of this essay develops a sequential bargaining model of the negotiations in corporate reorganizations under Chapter 11. We identify the expected outcome of the bargaining process and examine the effects of the legal rules that shape the bargaining. We determine how much value each class will receive and what factors give each class an advantage or a disadvantage in the bargaining process. We also compare the share of the reorganized firm's value that each class obtains under the existing legal regime with the contractual right of that class.[2]

---

[2]In spite of a growing interest among economists in corporate reorganization, little work has been done thus far to model bargaining under Chapter 11. Brown (1989) analyzed this bargaining, but his model does not include certain important features of the process that we seek to incorporate. First, whereas time plays no role in Brown's model, we analyze the implications of the critical fact that upon filing for Chapter 11, a company may remain under Chapter 11, protected from creditors, for some time. As our model shows, the bargaining is shaped by the possibility of such a delay, because the value available for division may well change over time as a result of the realization of uncertainty and the company's incurring financial distress costs. This delay plays no role in Brown's model, because he does not allow for uncertainty or financial distress costs. Moreover, whereas in Brown's model no class can make many offers, we develop a sequential bargaining model in which each class can make many during the Chapter 11 period.

In surveying past work, note should be made of the related literature about debt renegotiation outside of Chapter 11. See, e.g., Bergman and Callen (1991), Giammarino (1989), Hart and Moore (1989), and Webb (1987). (Of these papers, that of Bergman and

The critical features of the situation that we model are as follows. Once an insolvent company files for reorganization under Chapter 11, an "automatic stay" prevents debtholders from seizing the company's assets as long as the company is in Chapter 11. Of course, the company will not remain in Chapter 11 indefinitely; if there is no agreement on a reorganization plan, eventually the supervising court would convert the bankruptcy proceedings to a Chapter 7 sale. For the debtholders to obtain any value before such a conversion, however, the two classes, equityholders and debtholders, must agree on a division of value (i.e., on a reorganization plan).

If the parties do not agree immediately, and the company remains in Chapter 11 for some time, then the value available for division may well change, for two reasons. First, the company would incur some "financial distress costs," to be discussed later, which would reduce its value. Second, uncertainty will be realized: random shocks may increase or decrease the firm's value. Furthermore, if the two classes ultimately do not reach agreement, and the assets are sold under Chapter 7, such a sale might sometimes involve a loss of value. These potential consequences of the parties' failure to reach agreement are important elements of our model, because they provide the background against which the parties would decide which offers to make or accept.

Callen is the closest to ours in modelling approach; they also use a sequential bargaining model, though not one in which the firm value may fluctuate during the bargaining period due to the realization of uncertainty.) Debt renegotiation prior to formal bankruptcy proceedings, however, differs from that within such proceedings. Indeed, pre-bankruptcy bargaining is always shaped by the parties expectations about what will happen should the company file for bankruptcy, and the above literature therefore makes various assumptions about the outcome of such a filing -- assumptions that can be justified ultimately only by constructing a model of the bargaining within bankruptcy proceedings. Our focus, of course, is on the bargaining that follows filing for Chapter 11 and on how the legal rules of Chapter 11 affect this process.

Finally, after writing this paper, we learned about two independent attempts to develop sequential bargaining models of Chapter 11 by Baird and Picker (1991) and Kaiser (1991). While they share our approach, their models differ substantially from ours; in particular, they assume that the company's value does not fluctuate during during the Chapter 11 period.

Empirical studies of corporate reorganization, e.g., Eberhart, Moore, and Roenfeldt (1990), Weiss (1990), Franks and Torous (1989), confirm that Chapter 11 often enables equityholders to obtain a share of the value of the reorganized company even when that value is less than sufficient to cover debtholders' claims. The debtholders often agree to plans giving the equityholders a share of this value even when the debtholders are not paid in full. Our model identifies and analyzes three sources of the equityholders' ability to obtain value and the parameters that determine how much value they will obtain.

The fact that the equityholders' consent is necessary for a division of value enables them to obtain value, even if the company is insolvent, for three reasons. The first two reasons arise from the ability of the equityholders to delay the adoption of a reorganization plan from the beginning of the Chapter 11 period to a later date. First, if equityholders delay agreement, there may be a favorable resolution of uncertainty that would cause the value of the firm to exceed the value of its debt. Thus, the equityholders have an "option value," and to forego it they must be compensated by the agreed distribution of value. Second, if the equityholders withhold their consent and thereby delay agreement, the company will be expected to incur "financial distress costs" that will erode the value that debtholders can expect to receive. Thus, the equityholders' consent to a division of value can save the firm "financial distress costs," and therefore they can obtain a share of these savings. The fraction of these savings that they obtain depends upon the rules that govern the Chapter 11 bargaining. In particular, the equityholders in effect have an exclusive power to propose reorganization plans for a period of time, which increases the fraction of the savings of "financial distress costs" that the equityholders can capture.

The third reason that the equityholders can obtain value (even when the value of the reorganized firm is insufficient to pay the debt in full) arises when the parties expect a Chapter 7 sale to entail a loss of value. In this case, the equityholders' consent to a division of value is necessary to avoid this loss. Therefore, they can obtain a fraction of the value gained by avoiding a Chapter 7 sale.

Section C.1 of this essay reveals how much each of the elements described above contributes to the value that equityholders will receive. Sections C.2 and C.3 present some comparative statics analysis. In particular, the amount that equityholders will receive tends to increase in (1) the volatility of the value of the company's assets, (2) the extent to which delay in reorganization imposes "financial distress costs," (3) the length of the reorganization period, (4) the length of the period during which the equityholders have the exclusive right to make offers, (5) the extent to which liquidation imposes a loss in value, and (6) the extent to which the value of the company's assets cover the company's debts. As will be discussed, these results provide several testable implications of the model. Section D presents some informal discussion of various possible extensions of our formal model.

Finally, Section E offers concluding remarks on some issues for further research. While this essay focuses on the positive analysis of the bargaining process under Chapter 11, this analysis is a prerequisite for a normative analysis of the efficiency costs of Chapter 11. Accordingly, Section E discusses the implications of the model for the welfare effects of Chapter 11.

B. THE MODEL

We consider a company with one class of equity and one class of debt that files for bankruptcy under Chapter 11. Let V be the value of the company's assets at the time of filing. Specifically, V is the present discounted value of the future stream of earnings that the assets would produce if the the company had a capital structure such that the firm did not incur the "financial distress costs" to be described below.[3] We assume that all equityholders and debtholders share a common discount rate. Upon filing for bankruptcy, however, the company owes an amount D to debtholders, where $D>V$. Thus, the company is insolvent.

[3]For simplicity, we can assume this capital structure is the optimal capital structure (that which maximizes the value of the company).

(Section D.2 considers an extension of the model to the case in which $D \leq V$.)

Faced with this insolvency, the company files for Chapter 11 at time $t=0$. The firm is to emerge from the reorganization with an all-equity structure or some other capital structure that avoids the "financial distress costs" to be described below.[4] Under Chapter 11, the corporate reorganization process ends when the various classes of investors accept a plan including a division of value. Reorganization is not open-ended, however: if the parties fail to agree on a reorganization plan, the supervising court would eventually convert the bankruptcy proceedings to a Chapter 7 liquidation (usually following a petition by creditors). For concreteness, we assume that if the parties fail to reach an agreement within a period T, the company will be liquidated. The firm remains in Chapter 11 reorganization until time $t=T$ unless the parties reach agreement before then.[5]

For an agreement on the distribution of value, one class must propose a reorganization plan, and then both classes must accept it. The rules governing the

---

[4]Once the firm is reorganized and its shares are distributed, its new shareholders would have the power to select the company's management. We do not analyze the possibility that either the equityholders or the debtholders have any special management skills necessary for running the company. Ignoring this possibility appears quite appropriate for publicly traded firms, which are often run by professional managers. Baird and Picker (1991), who focus on the reorganization of closely held firms with a small group of shareholder-managers, consider the interesting implications of the presence of shareholder-managers with some special management skills.

[5]Our assumption that the firm cannot emerge from Chapter 11 before $t=T$ without an agreement is a simplification of the legal rules. In reality, the equityholders can impose a plan on the debtholders so long as the debtholders are paid in full. The equityholders would choose to exit Chapter 11 by paying D in full if during the bargaining the value of the firm were to rise sufficiently above D, for much the same reasons (to be discussed below in Section D.2) that they would rather avoid Chapter 11 if V were sufficiently greater than D.

We adopt our simplifying assumption for convenience; a more complex model could include this option for the equityholders to exit Chapter 11, but would not yield results significantly different from ours. In particular, this extension of our model would not change the direction of our comparative statics results. The addition of such an "exit option" would only increase the share of the firm obtained by the equityholders. This option can only strengthen the equityholders' bargaining power, because they would exercise this option only if it would improve their payoffs.

solicitation of acceptances, however, require some minimum time to pass between offer and response. In particular, section 1125(b) of the Bankruptcy Code requires that after the filing of a Chapter 11 case, acceptances may not be solicited unless the plan is transmitted to the accepting class with a written disclosure statement, which the supervising court must approve as containing adequate information after notice and a hearing. Furthermore, Bankruptcy Rule 3017(a) requires the court to hold this hearing on at least 25-days notice to the parties in interest. These notice and hearing requirements introduce a delay -- no offer can be accepted immediately. In this respect, bargaining in the reorganization context differs from that in other contexts, in which each round of bargaining -- an offer and the response to it -- can take a very short period of time.

Let Δt be the length of one round of bargaining, i.e., the time required until a proposed reorganization plan can be accepted or rejected under Chapter 11. Let $n=T/\Delta t$, assumed for simplicity to be an integer greater than 1, be the number of bargaining rounds during the reorganization period. Let V_i denote the value of the company (discounted to time $t=0$ by the common discount rate) if the parties adopt a reorganization plan at the end of round i, where $i = 1, \ldots, n$.[6]

The firm's value V_i evolves over time. That is, if the parties approve a reorganization plan in any round after round 1, the value divided might differ from V_1. Specifically, if delay were to occur, the value may change for the following two reasons. One reason is that the firm will bear "financial distress costs" -- efficiency costs that the firm must incur while it is in Chapter 11 and that it would not incur

[6]We assume that each proposed plan includes the optimal capital structure. That is, the capital structure itself is not the subject of bargaining. We rule out disagreements among the parties regarding capital structure in order to focus on the division of value between debtholders and equityholders.

If management had an interest in a suboptimal capital structure, as assumed in Hart and Moore (1990), then its preferences would diverge from those of the equityholders as well as from those of the debtholders. To model such a three-player bargaining game, it would be necessary to specify the preferences of management and to modify the bargaining protocol. We do not believe that abstracting from principal-agent problems, however, significantly affects the basic thrust of our results regarding the division of value.

if it had a new capital structure. There are several sources of financial distress costs. First, Chapter 11 bankruptcy involves significant administrative costs. Indeed, the fees paid to lawyers, accountants, various professional consultants, and expert witnesses in a Chapter 11 reorganization of a publicly traded company are often in the order of tens of millions of dollars. In one recent reorganization of a major corporation, the bankruptcy expenses of the company and of the creditor committees came to $3.5 million per month (see Cutler and Summers (1988, p. 167)). Second, potential business partners may be reluctant to deal with the company or demand especially favorable terms, for those doing business with a firm in financial distress may incur greater information costs, monitoring costs, enforcement costs, and collection costs. Third, financial distress might lead to inefficient management decisions with respect to the choice of projects and investments. Since Jensen and Meckling (1976), the economic literature has analyzed how a firm's indebtedness can lead management to pursue suboptimal investments strategies.[7] We assume that as a result of these financial distress costs, the company loses an amount of its value (discounted to time $t=0$) continuously over time at the rate of α per period Δt, i.e., per round of bargaining, as long as the company is in Chapter 11.[8]

---

[7]During the reorganization period, the incumbent management retains authority to manage the company as "debtor in possession." That management, as discussed below, is commonly assumed to represent the equityholders, whose interests may be served by decisions that do not maximize the value of the company. Jensen and Meckling (1976) and Myers (1977) showed that the value of a firm with debt will tend to deteriorate then, because management will make suboptimal decisions. For example, the management will have an incentive to invest inefficiently in risky projects, because equityholders may have little to lose from a downward turn in the company's fortunes and more to gain from an upward turn. Hart and Moore (1990) criticize the common assumption that management acts on behalf of shareholders. As they show, however, even if one assumes that management pursues only its own "empire-building" objectives, a large level of debt may cause management to pass up profitable projects.

[8]Note that financial distress costs in our model may result merely from the firm's presence under Chapter 11 and would be eliminated only if the parties reach a reorganization agreement (creating a sound financial structure). Bergman and Callen (1991) present an interesting model in which the shareholders threaten to cause some

The second reason for V_i to change over time is that uncertainty will be realized. Random shocks may increase or decrease the value of the company's assets. We assume that during each round of bargaining, unpredictable events would cause this value to be either higher or lower than expected by the amount θ, and that either state of the world occurs with equal probability. The parameter θ, then, represents the volatility of the firm's value. For concreteness, we assume that at the beginning of each round, information is revealed that causes the value of the firm (discounted to time $t=0$) either to rise or to fall instantaneously by the amount θ.

Let the value realized after revelation of information at the start of round 1 be V and V_i be the value at the end of round i. Thus, $V_1 = V-\alpha$, and $V_i = V_{i-1}+\delta_i$ for $i = 2, ..., n$, where $\delta_i = -\alpha-\theta$ with probability ½, and $\delta_i = -\alpha+\theta$ with probability ½. The value of δ_i is realized at the start of round i, before the ith offer is made. We assume $0<\alpha<\theta$, so that the company's value will increase with probability ½ despite the financial distress costs. We also assume that $(n-1)(\alpha+\theta) < V_1$, so that the company's value cannot disappear completely during the reorganization period.

We also assume the following timing of events. In each round, once the uncertainty is realized, one party proposes a reorganization plan, i.e., a division of the value that will exist at the end of the round. At the end of the round, after the delay created by the rules governing the solicitation of acceptances, the value has deteriorated by the amount α. At this point, the two classes decide whether to accept or reject the plan. The class proposing the plan, naturally, votes to accept the proposal. If the other class also accepts the proposal, the plan is confirmed, and the firm immediately emerges from Chapter 11 with its value divided according

deterioration in the situation of a solvent firm as a bargaining device in debt renegotiations. In their model, the erosion in the firm's value is a strategy chosen by the equityholders deliberately to force concessions from debtholders, and a critical question with respect to such threats is that of credibility. The credibility issue does not arise with respect to the financial distress costs in our model, however, because these costs would be incurred as long as the firm is in Chapter 11, and only reorganization can avoid these costs.

to the plan. If the other class instead rejects the proposal, then the parties go to the next round of bargaining.[9]

As in any sequential bargaining model, it is necessary to specify a procedure that determines which party makes the offer in each round. In the case of Chapter 11 bargaining, the Bankruptcy Code determines some of this bargaining protocol. Section 1121(b) of the Bankruptcy Code establishes a period during which only the debtor (i.e., the incumbent management) has the right to propose reorganization plans. The incumbent management generally will have at least six months in which to produce a reorganization plan and obtain the necessary acceptances, and courts may grant extensions of this period (see Trost (1979, p. 1325)). During the first several months of reorganization, then, usually the parties can adopt only a plan proposed by management.

Although the insolvent corporation and its management have fiduciary duties to creditors as well as shareholders, most observers expect management to favor the interests of shareholders. As Normandin (1989, pp. 56-58) notes, management not only is elected by the shareholders but also often holds a substantial interest in the corporation's stock. Consistent with this observation, this essay assumes that with respect to proposals to divide the value of the reorganized firm between

---

[9]The above description of the sequence of events assumes implicitly that each class of claimants acts as a single agent. We assume that the set of equityholders and the set of debtholders are mutually exclusive. Given our assumptions, there is no conflict of interest among the class members, and they will all agree on whether a plan should be accepted or rejected.

Given the homogeneity of each class in the model, there is no need to introduce the possibility that exists under section 1129(a)(7) of the Bankruptcy Code for a veto by an individual member. Under that provision, dissenting members can veto a plan that fails the "best interests of the creditors" test -- that is, a plan that provides them with less than the liquidation value of their claims -- even if the required majority of the class votes to accept the plan (see Baird and Jackson (1990, pp. 59, 950)). If there were differences in preferences or in information among class members, then individual members may wish sometimes to exercise such a veto. If instead there are no such differences among class members, as we assume, then no member would wish to veto a plan that other members of the class voted to approve.

equityholders and debtholders, management acts on behalf of the equityholders.[10]

To represent the period during which the management has the exclusive right to propose plans, we assume that the equityholders make at least the first offer and possibly the first several offers. That is, they have the exclusive right to make offers for the first e rounds, where e is an integer such that $0<e<n$.[11] Over the remaining n-e rounds, both classes may make offers. (This possibility is indeed contemplated by Chapter 11, and firms often remain in Chapter 11 even after the exclusive period runs out.) To capture the possibility of either party making the offer in any given round, we assume that at the beginning of each of these n-e rounds, the identity of the class making the offer is determined randomly, with the probability of each equal to ½.[12]

If the parties were to fail to reach agreement in round n, we assume that the court would convert the proceedings to a Chapter 7 liquidation. In that event, the assets of the company would be sold. Let the random variable L denote the value obtained through such a sale (discounted to time $t=0$ by the common discount

---

[10]It is not our assumption that management acts in the equityholders' interests in all respects. The interests of the equityholders and of management may well diverge with respect to certain investment decisions. With respect to management's proposals to divide the value of the reorganized firm between equityholders and debtholders, however, we believe it reasonable to assume that management will prefer the interests of equityholders.

The assumptions that the managers initially have exclusive control of the bargaining agenda, and that they use this control to the advantage of the equityholders, do not matter much for our results. As will be noted, one can easily modify our model to accommodate the case in which there is no initial agenda control, or the case in which the control is not exercised in the interest of equityholders, and these modifications do not have important effects on our conclusions. See Section D.3 and footnote 18, respectively.

[11]We can easily extend our analysis to the case in which $e=n$. See footnotes 15, 20, and 34.

[12]Having each party make the offer with probability ½ in each round is one of the two conventional alternatives in modelling a situation in which two parties can make offers. The second alternative is to assume that the two parties alternate in making offers. In a model with a finite number of rounds (such as ours), however, the first alternative ensures that the division of value is not sensitive to whether the number of rounds, n, is even or odd. We prefer the first method because it renders it unnecessary to make such arbitrary assumptions regarding n.

rate), which would then be divided according to the legal priority of the parties' claims. That is, the debtholders would receive, to the extent that there is enough to give them, what they are owed. We assume that this amount in "real" terms is D; that is, we the interest rate on the debt equals the common discount rate.[13] Thus, equityholders get an amount VE_L = max(0, L-D) in present disounted value, and debtholders get the complementary share, VD_L = L - max(0, L-D) = min(L, D), also in present discounted value. At this point, we will assume that $L=V_n$; that is, we assume that the sale procedure occurs immediately and does not involve any loss of value. In Section C.3, however, we will drop the assumption that $L=V_n$ and extend the analysis to cover the case in which $L<V_n$.

We also assume that all parties are risk neutral. Thus, each class seeks to maximize the expected value of its share. Finally, we assume that the structure of the bargaining game described above, including the values of the parameters, V_1, D, L, α, θ, T, e, and n, is common knowledge to the participants.

C. THE DIVISION OF VALUE

Under the assumption of common knowledge, the parties will reach an agreement at the end of round 1. As in the sequential bargaining game analyzed by Rubinstein (1982), to obtain agreement, the party making the proposal in round i would offer the other at least what the other could obtain in expected value in round i+1, where i = 1, ..., n-1. Each party would offer just this amount to the other in order to maximize its own share. It cannot expect to do better by asking for any larger amount, which would delay agreement and allow the other party to make the next offer.

The party whose turn it would be in round i, denoted by X, offers to take the amount VX_i (in present disounted value) of the total value V_i. By the argument

[13]To the extent that the interest rate is smaller or that the court will not allow the accumulation of interest during the reorganization period, then there is another reason that equityholders can obtain value beyond the three reasons identified below. In such a case, delay would erode the value of the debt in real terms.

above:

$$VX_i = V_i - E_i[V_{i+1} - VX_{i+1}], \tag{1}$$

where VX_{i+1} denotes the present discounted value of the amount X would obtain under the plan proposed in the next round, and E_i denotes the expected value conditional on information available in round i, i.e., conditional on the realized value of V_i. By our assumptions about δ, we also know that the expected "efficiency gain" from an agreement in round i rather than round i+1 is:

$$V_i - E_i[V_{i+1}] = -E_i[\delta_{i+1}] = \alpha. \tag{2}$$

Using (2) to substitute into (1):

$$VX_i = \alpha + E_i[VX_{i+1}]. \tag{3}$$

Thus, when the party making the offer, X, holds the other to its expected payoff next round, X thereby takes for itself the expected "efficiency gain," α, plus its own expected payoff next round.

Under liquidation at the end of round n, equityholders would receive VE_L, and debtholders would receive VD_L, where $VE_L + VD_L = L = V_n$. In round n, the party making the offer cannot gain by proposing any other division, because the other party will reject any proposal offering it a smaller share. Therefore, $VE_n = VE_L$ and $VD_n = VD_L$. The party making the nth offer cannot capture any "efficiency gain" from avoiding financial distress costs, because we have assumed here that liquidation occurs costlessly and immediately upon rejection of the offer.

We solve for the unique subgame perfect equilibrium by backward induction. In round n-1, the party making the offer would hold the other to its expected round n payoff and take for itself the expected "efficiency gain" (from agreement in round n-1 rather than round n), α, plus its own expected round n payoff. Similarly, in round n-2, the offering party would hold the other party to its expected round n-1 payoff (which is ½α plus its expected round n payoff, if it would make the offer in round n-1 with probability ½) and take the balance. That balance amounts to the expected "efficiency gain" (from agreement in round n-2 rather than round n-1), α, plus its own expected round n-1 payoff. Its expected round n-1 payoff, in turn, is also ½α plus its expected round n payoff, if it would make the offer in round n-1 with probability ½.[14]

In sum, in each round, each party receives its expected round n payoff plus α times the number of offers it would expect to make from the current round to round n-1. In round 1, therefore, the equityholders receive their expected round n payoff plus α times the number of offers they would expect to make before round n. Because each party would expect to make ½(n-1-e) offers from round e+1 to round n-1, and the equityholders in addition would make all e offers from round 1 to round e, the equityholders receive:

$$VE_1 = \alpha e + \tfrac{1}{2}\alpha(n-1-e) + E_1[VE_n], \tag{4}$$

and the debtholders receive:

$$VD_1 = \tfrac{1}{2}\alpha(n-1-e) + E_1[VD_n], \tag{5}$$

at the end of round 1. Moreover, because we assume liquidation is costless, (4) equals:

$$VE_1 = \tfrac{1}{2}\alpha(e+n-1) + E_1[\max(0, V_n-D)], \tag{6}$$

and (5) equals:

$$VD_1 = \tfrac{1}{2}\alpha(n-1-e) + E_1[\min(D, V_n)]. \tag{7}$$

The second term in (6) is the expected value of what equityholders would receive if there is no agreement through round n and there is instead liquidation. In the alternative, we may express this expected value as:

$$E_1[\max(0, V_n-D)] = \Pr(V_n>D)E_1[V_n-D \mid V_n>D], \tag{8}$$

where $\Pr(V_n>D)$ is the probability of $V_n>D$. We can also express this expected value as a function of our basic parameters, making use of the particular probability distribution of V_n:

<u>Lemma 1</u>: The expected values of what the parties would receive from a costless liquidation if they failed to reach agreement by time t=T are as follows:

---

[14]All expectations are taken conditional on information available at the time of the current offer, i.e., in round n-2. Recall that by the law of iterated expectations, the expectation of an expectation conditional on more information is simply the unconditional expectation. In the example above:

$$E_{n-2}[E_{n-1}[VX_n]] = E_{n-2}[VX_n].$$

$$E_1[\max(0,V_n-D)]=\sum_{k=0}^{n-1}(½)^{n-1}\binom{n-1}{k}\max[0,V_1-D-(\alpha+\theta)(n-1)+2\theta k] \qquad (9)$$

for the equityholders, and:

$$E_1[\min(D,V_n)]=\sum_{k=0}^{n-1}(½)^{n-1}\binom{n-1}{k}\min[D,V_1-(\alpha+\theta)(n-1)+2\theta k] \qquad (10)$$

for the debtholders.

Proof: Each round is a Bernoulli trial in which δ may take on one of two values with equal probability, so n-1 such trials from round 2 to round n leads to 2^{n-1} equally likely sequences of δ_i. Let k represent the number of times δ takes on the high value, θ-α, from the end of round 1 to the end of round n, so that n-1-k is the number of times it takes on the low value, -θ-α. The number of distinct sequences with the same k is given by the binomial coefficient. Thus, k is a random variable, taking values from 0 to n-1, that follows the binomial distribution:

$$b(k,n-1)=(½)^{n-1}\binom{n-1}{k}=(½)^{n-1}\frac{(n-1)!}{k!(n-1-k)!}\ . \qquad (11)$$

Given any set of values for the parameters known in round 1, including V_1, there are n possible values for V_n, because there are n possible values for k:

$$V_n(k) = V_1-\alpha(n-1)+\theta k-\theta(n-1-k) = V_1-(\alpha+\theta)(n-1)+2\theta k. \qquad (12)$$

Therefore, the probability of each $V_n(k)$ is given by the binomial formula in (11). Together (11) and (12) yield the above lemma. ■

Using (9) and (10) to substitute into (6) and (7), respectively, we can conclude with the following proposition:[15]

Proposition 1: The two classes of claimants will adopt a reorganization plan in round 1, with the equityholders obtaining:

[15]The above discussion assumes that e<n. As long as we assume that liquidation is costless, the division of value when e=n is simply the same as that implied by this proposition when e=n-1. That is, if a Chapter 7 sale involves no loss of value, then it does not matter which class makes the offer in round n.

$$VE_1=\tfrac{1}{2}\alpha(e+n-1)+\sum_{k=0}^{n-1}(\tfrac{1}{2})^{n-1}\binom{n-1}{k}\max[0,V_1-D-(\alpha+\theta)(n-1)+2\theta k], \qquad (13)$$

and the debtholders obtaining the complementary share:

$$VD_1=\tfrac{1}{2}\alpha(n-1-e)+\sum_{k=0}^{n-1}(\tfrac{1}{2})^{n-1}\binom{n-1}{k}\min[D,V_1-(\alpha+\theta)(n-1)+2\theta k]. \qquad (14)$$

1. <u>The Sources of the Equityholders' Power</u>

Although debtholders' claims exceed the value of the firm in round 1, equityholders receive a round 1 share greater than 0. We can isolate two distinct sources of the equityholders' power to obtain more than its contractual right. The equityholders' share, VE_1 as expressed in (13), is composed of two terms. Each term corresponds to a different reason for equityholders to get a positive share even though $D>V_1$. (In Section C.3, when we introduce the possibility that conversion to Chapter 7 liquidation may cause the firm to lose value, there will be a third source of the equityholders' power to obtain value.)

<u>a. The Financial Distress Costs Created by Delay</u>: The total "efficiency gains" from reaching agreement in round 1 rather than leaving the firm in Chapter 11 for n rounds is $\alpha(n-1)$, because such agreement thereby avoids the deterioration in firm value that would occur during the next n-1 rounds. Because the consent of the equityholders is necessary to avoid incurring these financial distress costs, the equityholders can obtain part of these savings in exchange for that consent. The first term in (13) is the equityholders' share of these savings and represents the gain to equityholders from the presence of financial distress costs. If we let α go to 0 in the limit, then the savings from avoiding delay also go to 0, and the equityholders would receive only the second term, $E_1[VE_n]$.

The first term in (13), $\tfrac{1}{2}\alpha(e+n-1)$, may be expressed as the sum of two elements, as in (4). First, during the first e rounds, when the equityholders have the exclusive power to make offers, they expect to capture all the savings from agreement in round 1 rather than in round e+1. This surplus amounts to αe. Second, during the last n-e rounds, when both the equityholders and the debtholders can make offers, each class expects to capture half of the savings from agreement in round e+1 rather than in round n. Therefore, the equityholders can obtain half of this surplus, i.e., $\frac{1}{2}\alpha(n-1-e)$, in round 1.

<u>b. The "Option Value" Created by Delay</u>: The second term in (13) represents the gain to equityholders from the possibility of V_n greater than D. The equityholders could deny their consent to any plan and delay a division of value until round n. Given the volatility of V_i, there is some chance that at the end of round n, V_n will exceed D. Although the expected value of V_n is less than V_1, the actual value of V_n might exceed V_1 and even D. The Chapter 11 process, by giving the equityholders the ability to insist on delay, gives equityholders the option to receive the difference V_n-D. To give up this option, equityholders must get compensation in return.

Specifically, equityholders receive the expected value of VE_n, i.e., the full value of the equityholders' "option." This "option value" is positive if and only if $Pr(V_n>D) > 0$. Furthermore, $Pr(V_n>D) > 0$ if and only if $V_n>D$ when all uncertainty is resolved favorably, i.e., $\delta = \theta-\alpha$ in each round, so that $k = n-1$ in (12). Therefore, the following two conditions are equivalent:

$$Pr(V_n>D) > 0 \leftrightarrow (n-1)(\theta-\alpha) > D-V_1. \quad (15)$$

2. Comparative Statics

The equityholders' share of V_1 consists of two terms that depend on the following parameters: the initial undercoverage of debt, $D-V_1$; the volatility of the firm's value from period to period, θ; the financial distress costs per unit of time Δt, α; the length of the reorganization period measured in rounds of bargaining, n; and the length of the period during which the equityholders control the agenda, also measured in rounds of bargaining, e. We will now examine how each of these parameters affects the equityholders' share, VE_1.

a. The Initial Undercoverage of Debt: By the initial undercoverage of debt, we mean the extent to which the debt exceeds the value of the firm's assets at the end of round 1, $D-V_1$.

Proposition 2: The share of V_1 obtained by equityholders, VE_1, is nonincreasing in $D-V_1$ (and thus in D). Furthermore, if there is a positive probability that V_n will exceed D, then VE_1 is decreasing in $D-V_1$ (and thus in D).[16]

Remark: The intuition underlying Proposition 2 is as follows. A rise in $D-V_1$ affects only the second term in (13). In particular, it shifts down the entire probability distribution of V_n-D in the expression for the equityholders' "option value" in (13). That "option value" arises from the possibility of V_i rising from V_1 to above D due to positive shocks. Because a higher $D-V_1$ makes this event less likely, it reduces this "option value." Thus, as shown in the proof, a rise in $D-V_1$ cannot make equityholders better off, and if $D-V_1$ is small enough to imply $Pr(V_n>D) > 0$, then such a rise must cause their "option value" (and their share of V_1) to decrease.

Proof: See the Appendix.

b. The Volatility of Firm Value: The volatility of the firm's value over time, represented by θ, depends upon the nature of the firm's business and thus may vary greatly from firm to firm.

Proposition 3: The share of V_1 obtained by equityholders, VE_1, is nondecreasing in θ. Furthermore, if there is a positive probability that V_n will

---

[16]Specifically, if $D-V_1 < (n-1)(\theta-\alpha)$, then $Pr(V_n>D) > 0$, and increasing $D-V_1$ will cause VE_1 to decrease. If $D-V_1 \geq (n-1)(\theta-\alpha)$, then $Pr(V_n>D) = 0$, and further increasing $D-V_1$ will have no effect on VE_1.

exceed or equal D, then VE_1 is increasing in θ.[17]

Remark: The intuition underlying Proposition 3 is as follows. The volatility of the firm's value affects only the second term in (13). As shown in the proof, the greater that volatility, θ, the greater the likelihood that V_n will exceed D, and the greater will be the equityholders' "option value." Thus, a rise in θ must leave equityholders at least as well off as before, and if θ is large enough to imply $Pr(V_n \geq D) > 0$, then their "option value" (and their share of V_1) must increase in θ.

Proof: See the Appendix.

c. The Financial Distress Costs: The financial distress costs per unit of time that the firm remains in Chapter 11, α, is likely to vary greatly among firms. For example, a company that owns a number of buildings leased under long-term leases would incur small financial distress costs (because few transactions would have to take place during the Chapter 11 period) relative to a company that deals frequently with many business partners and makes frequent investment decisions.

Proposition 4: The share of V_1 obtained by equityholders, VE_1, is increasing in α.

Remark: The intuition underlying Proposition 4 is as follows. The size of the financial distress costs, α, will affect both terms in (13). Thus, a rise in α has two effects. On the one hand, it increases the financial distress costs that can be saved if the equityholders consent to a plan, thus improving their bargaining position. On the other hand, it decreases their "option value" because it makes it less likely that V_i can climb beyond D. If α increases to the point that $Pr(V_n > D) = 0$, then the equityholders' share is affected by α only through the first effect and therefore a rise in α makes equityholders unambiguously better off. Furthermore, the positive effect is always greater than the negative effect, because the financial distress costs are more likely to fall on the debtholders than on the equityholders.

[17]Specifically, if $\theta < \alpha + (D-V_1)/(n-1)$, then $Pr(V_n \geq D) = 0$, and increasing θ will have no effect on VE_1. If $\theta \geq \alpha + (D-V_1)/(n-1)$, then $Pr(V_n \geq D) > 0$, and increasing θ will cause VE_1 to increase.

Therefore, the equityholders' share increases in α.

Proof: See the Appendix.

d. The Length of the Reorganization Period: The length of the reorganization period, represented in our model by T (or equivalently by the number of rounds n), depends on how long the bankruptcy court will wait, in the face of failure to work out a reorganization plan, before converting from Chapter 11 to Chapter 7. Let us consider increases in T by multiples of Δt, which would cause n to rise by an integer, so as to maintain the structure of the model which assumes that n is an integer.

Proposition 5: The share of V_1 obtained by equityholders, VE_1, is increasing in n.

Remark: The intuition underlying Proposition 5 is as follows. A rise in T affects both terms in (13) through n. First, it raises the total financial distress costs that the firm would incur if equityholders never gave their consent to a plan. This effect strengthens their bargaining power. Second, once $Pr(V_n > D) > 0$, an increase in n also would have an effect on their "option value." As shown in the proof, some increases will cause an upward change in the "option value" (because high values for V_n become more likely) whereas others will cause a downward change (because the expected value of V_n falls). We show in the proof, however, that even in the case of a downward change, the overall effect on the equityholders' share is positive. In those cases in which a rise in n causes the equityholders' "option value" to fall, the positive effect upon the equityholders' share of the "efficiency gains" would be greater than the absolute value of the negative effect, because the costs of delay would be more likely to fall on the debtholders than the equityholders.

Proof: See the Appendix.

e. The Initial Agenda Control: As already noted, the equityholders have control over the agenda for the first six months of bargaining, and the courts often extend this period. Our model captures this feature of Chapter 11 by assuming that all offers in the first e rounds, 0<e<n, must be made by the equityholders.[18] The size of e depends on the courts' willingness to extend the initial six-month period.

Proposition 6: The share of V_1 obtained by the equityholders, VE_1, is increasing in e. Specifically, any unit increase in e causes VE_1 to increase by $\frac{1}{2}\alpha$.

Remark: The intuition underlying Proposition 6 is as follows. Control over the reorganization agenda is valuable. During each round in which the equityholders have this control, they would capture all of the financial distress costs, α. In any round in which both sides can make offers, however, the equityholders expect to capture only $\frac{1}{2}\alpha$. The incremental gain to the equityholders from any unit increase in e is therefore $\frac{1}{2}\alpha$. The greater are the financial distress costs, the greater is the value of this unit increase (the derivative of the equityholders' gain with respect to α is ½). Furthermore, a higher e implies that a rise in α benefits equityholders even more than it would otherwise, because only they gain bargaining power from the deterioration of firm value from round 1 to round e+1. The derivative of the "efficiency gains" component in (13) with respect to α is ½(e+n-1), and each unit increase in e raises this derivative by ½.

Proof: The proof is clear from the above remark.

3. Loss from Chapter 7 Liquidation

Until now we have assumed that a sale of the firm's assets under Chapter 7 would not impose any efficiency loss. That is, we assumed that after n rounds a Chapter 7 sale would produce a value of $L=V_n$. This assumption would be reasonable if one adopted the view of Baird (1986) and Jackson (1986, ch. 9).

[18]In reality, the incumbent management controls the agenda during this initial period, but we have assumed that managers act on behalf of the equityholders. One could easily adjust the model to suit alternative assumptions. For example, consider a case in which the debt is held by a bank which the managers expect to hold most of the stock of the company after reorganization. In this case, one might assume that the management, in making offers during the period in which it controls the agenda, would serve the interests of the debtholders. As a result, the debtholders rather than the equityholders would capture all the surplus $e\alpha$ that is created by avoiding the financial distress costs that would be incurred during the initial period.

According to that view, there is no reason to expect that a sale under Chapter 7, properly administered, would produce a value significantly below the firm's going-concern value. If the firm has greater value as a going concern, then the assets would be sold as a going concern.

This view of Chapter 7, however, is by no means universally accepted. Many scholars and players in corporate insolvency believe, correctly or incorrectly, that Chapter 7 would involve some efficiency loss -- that is, it might not bring in the full value V_n. For this reason, we now introduce the possibility that a Chapter 7 sale at t=T would produce a value $L<V_n$.[19] Specifically, let $L = V_n-\lambda$, where $0 \leq \lambda < V_1-(n-1)(\theta+\alpha)$, so that $0<L\leq V_n$ in all states of the world, i.e., for any k. All other assumptions are as before. Using the same methods as before, we find that the loss in value from liquidation introduces a third source of the equityholders' bargaining power:

Proposition 7: If the two classes of claimants expect a Chapter 7 sale to involve a loss of value λ, then they will adopt a reorganization plan in round 1, with the equityholders obtaining VE_1 =

$$\tfrac{1}{2}\alpha(e+n-1)+\tfrac{1}{2}\lambda+\sum_{k=0}^{n-1}(\tfrac{1}{2})^{n-1}\binom{n-1}{k}\max[0,V_1-D-\lambda-(\alpha+\theta)(n-1)+2\theta k]. \quad (16)$$

Proof: See the Appendix.

Let us now examine the change that is introduced by the loss λ from Chapter 7. As the following proposition indicates, the equityholders' share of V_1 tends to be larger as this loss is larger.

Proposition 8: The share of V_1 obtained by equityholders, VE_1, is nondecreasing in λ. Thus, the equityholders are at least as well off when $\lambda>0$ as they are when $\lambda=0$. Furthermore, if the probability that L will exceed D is less

[19]Furthermore, to the extent that liquidation does not occur immediately upon rejection of the last offer in round n, the firm would incur some financial distress costs between the end of round n and liquidation. In the alternative, then, the possibility of $L<V_n$ could represent the deterioration in the firm's value between the end of round n and liquidation.

than ½, as is likely, then VE_1 is increasing in λ.[20]

Remark: The intuition underlying Proposition 8 is as follows. The expectation of the liquidation loss λ would strengthen the equityholders' bargaining power and thereby increase their share VE_1. Without their consent to a reorganization plan, the firm would not only suffer financial distress costs for n rounds, but would also suffer the losses from a Chapter 7 sale for less than the firm's full value. Thus, the equityholders' consent can save the firm the liquidation loss, λ, and therefore they can expect to capture some part of these savings in exchange for their consent. At the same time, in our model, such a liquidation loss would reduce the equityholders' "option value," $E_1[VE_L]$. As shown in the proof, however, the positive effect is at least as great as the negative effect on this "option value." Furthermore, under conditions that ensure $Pr(L>D) < ½$, which is likely to hold for plausible parameter values, the positive effect must strictly dominate the negative effect, because then the loss would be more likely to fall on the debtholders than on the equityholders.

Proof: See the Appendix.

D. EXTENSIONS

In developing our formal model, we have made numerous assumptions to simplify the analysis. In this section, we discuss the implications of relaxing some of these assumptions. An informal examination of these extensions suggests that our simple model is likely to capture the most important features of the division of value in actual corporate reorganizations.

1. Renegotiation Prior to Chapter 11 Filing

Our model has assumed that the bargaining game begins upon the firm's filing under Chapter 11. One might consider, however, whether the parties would

[20]In particular, if $(D-V_1) + \alpha(n-1) + \lambda \geq \theta$, as is likely, then $Pr(L>D) < ½$, and VE_1 is increasing in λ. Furthermore, if e=n rather than e<n, then VE_1 is always increasing in λ.

renegotiate their claims prior to filing under Chapter 11 (see, e.g., Berkovitch and Israel (1991), Bergman and Callen (1991), and Giammarino (1989)). To the extent that some financial distress costs are incurred prior to the filing, as is likely as the company approaches insolvency, the parties would have much to gain from such an earlier renegotiation. By this reasoning, if the parties know that once there is filing for Chapter 11 the firm's value will be divided in a certain way, then there will be no such filing because they will agree to such a division beforehand. Brown (1989), for example, assumes in his model that such renegotiation will always take place.

Various factors, however, suggest that such renegotiation may not succeed outside of Chapter 11, because there may be a "free rider" or "holdout" problem (see Gertner and Scharfstein (1991) and Gilson, John, and Lang (1990)). Consequently, renegotiation would instead take place, as in our model, only after the filing, and in accordance with the rules governing Chapter 11 reorganization. Outside of Chapter 11, each debtholders' claim cannot be waived unless the debtholder individually consents. Thus, each debtholder may hold out, preferring that other debtholders make the necessary concessions. In contrast, under Chapter 11 a provision for majority rule solves this collective action problem: the specified majority in any given class can vote to make concessions on behalf of each of the members of the class, without the unanimous consent of its members, and can thereby impose the plan on its members who dissent.[21]

Thus, although the parties may renegotiate their claims before filing for Chapter 11 to avoid excessive delay within Chapter 11, they will face the problem of holdouts. To solve this problem, the company may file for Chapter 11 so that a majority vote on the proposed plan would bind all members within each class.

[21]Similarly, contractual mechanisms could solve this collective action problem outside of Chapter 11. The Trust Indenture Act, however, prohibits contractual provisions that would enable the majority of bondholders to concede any part of the principal on the bonds. As Roe (1987) notes, the unanimous approval requirements of that law may generate holdout problems within a class and thereby inhibit a troubled firm's ability to avert bankruptcy proceedings.

Through this mechanism, the negotiating parties can settle on a reorganization plan before the filing, then have the pre-packaged plan confirmed within Chapter 11. See Broude (1990, § 11.06). Nevertheless, as the parties formulate the pre-packaged plan, they would anticipate the outcome of Chapter 11 bargaining. Thus, under such pre-packaged plans, the equityholders would not receive less than they would expect from Chapter 11 bargaining.

2. Chapter 11 Filing by "Solvent" Companies

Thus far we have assumed that upon filing for reorganization, the value of the firm's assets is less than its debt: $V<D$. A company, however, need not be insolvent to file under Chapter 11, and observers generally believe that some firms file with $V \geq D$. We now sketch briefly what our model suggests with respect to the outcome of such a filing (and the ex ante incentive on the part of equityholders to have such a filing).

Suppose that a firm files for Chapter 11 with $V \geq D$, but with V still close enough to D to cause financial distress costs. Our model can be extended to such a case. As before, the parties would understand that delay in adopting a reorganization plan would have two consequences: the firm would incur financial distress costs, and uncertainty would be resolved. Therefore, Propositions 1 and 7 would still describe the shares of value received by the two classes. The interesting question, however, is whether this expected outcome of the bargaining game would lead equityholders to favor a filing for Chapter 11.

Our model suggests that if V exceeds D by a sufficiently small amount, the equityholders can still get more than their contractual right, V-D. Uncertainty would work in favor of the equityholders through their "option value."[22]

[22]This "option value" still increases in V-D as Proposition 2 would indicate. As long as the risk of insolvency at the time of liquidation persists, however, the "option value" would not increase as rapidly as the equityholders' contractual right, V-D. Once that risk falls to zero, the equityholders' "option value" will fall short of V-D.

Proposition 3 would require only minor modification. The "option value" would increase in θ if both insolvency and solvency occur with positive probability at the time of

Moreover, the financial distress costs that would be incurred during a delay in reorganization for a period T would come partly at the expense of the equityholders, but if V does not exceed D by too much, then most of these costs would come at the expense of the debtholders. Consequently, the debtholders would make concessions -- agree to be paid less than their contractual right -- and the equityholders would thus obtain more than their contractual right. Given this expected outcome, it would be in the equityholders' interest in such a case for the company's management to file for Chapter 11.

Our model also suggests, however, that Chapter 11 might lead to a partition of value that would give the equityholders less than their contractual right. Specifically, if V exceeds D by an amount small enough to cause financial distress costs but also large enough to place most of the burden of these costs on the equityholders, then Chapter 11 would favor the debtholders.[23] Such Chapter 11 cases, however, might not actually arise. In such cases, the interest of equityholders would not be served by the company filing for Chapter 11 -- they would prefer to have the company pay the debt in full, raising the necessary funds by selling assets or issuing extra equity. Thus, to the extent that management seeks to advance the interests of the equityholders, one would not expect the company to file for Chapter 11 in such cases.[24]

3. Informal Offers and Agenda Control

One feature of Chapter 11 that we sought to represent in our model is the fact that during a certain period only the company could propose reorganization

---

liquidation. Otherwise, θ would have no effect on the outcome.

[23]Thus, if $V \geq D$, then the costs of delay and liquidation may fall mainly on equityholders, and the comparative statics results in Propositions 4, 5, and 8 become ambiguous. The equityholders would still gain from an increase in e, however, as Proposition 6 would indicate.

[24]Managers may file for Chapter 11 in such a case, however, for some self-serving reasons.

plans. We assumed that the equityholders would make all offers during this period, because we assumed that the company's management acts in the interests of the equityholders. We assumed that thereafter either one class or the other would have control of the bargaining agenda during each bargaining round, and that the procedural rules governing formal proposals determined the duration of these rounds.

One might argue, however, that the model should not feature such exclusive agenda control, in light of the possibility of informal offers. For example, although the debtholders cannot formally propose plans in the initial period during which the company has an exclusive right to do so, they can tell the company informally what plans they would be willing to accept. Therefore, in spite of the exclusive control of the formal agenda under Chapter 11, one might model the bargaining as a symmetric game in which both sides may make offers. According to this view, the legal monopoly over the formal agenda is practically meaningless.

Such a change in assumptions would require only minor modifications in our model, because exclusive agenda control is not a critical element of the model. If both classes may make offers throughout the bargaining period, then each would simply obtain half of the savings created by avoiding the financial distress costs during this period. Nevertheless, because the provision for exclusive agenda control during the initial period often arises in policy discussions, we should explain why in our view, in spite of the possibility of informal offers, control over the formal agenda is important in practice and may give a significant bargaining advantage to the party with this control.

Recall that in any sequential bargaining model, it is advantageous to be the party that makes the offer in any given round. When the other side decides whether to accept the offer on the table, acceptance is the only alternative to waiting until the next round, and the standard premise in these models is that waiting (however brief the delay) imposes some cost. The party who makes the offer can capture almost all the surplus created by avoiding the delay, because the other party would rather accept any small share of the surplus than reject the offer and lose the surplus.

Similarly, when one class proposes a formal plan under Chapter 11, the other class must decide whether to accept, after the delay required by law. Each knows that rejection would result in the extra delay required for another proposal and another vote. The other class would find it rational to accept any plan that gives its members some fraction (however small) of the surplus created by avoiding that extra delay.

Of course, informal offers may precede a formal proposal, but what would be the implications of such an offer? Unlike a formal plan, an informal offer cannot be accepted by a binding class vote. The "acceptance" of an informal offer thus lacks the legal implications of the acceptance of a formal offer: an informal acceptance cannot partition the company's value nor bring the company out of Chapter 11. An informal proposal is merely a suggestion for the other side to make a particular formal proposal. Regardless of such suggestions, the party making the formal proposal can be expected to propose the plan most favorable to the proposing party among those that the other party would rationally accept.[25] Informal proposals cannot affect this plan if (as is the case in our setting) there is no mechanism by which the party making an informal proposal can make a credible commitment to reject any less favorable proposal.[26] Moreover, as long as the

[25]Consider a company in which the value of the assets is $100. Suppose that in one month the court will convert the Chapter 11 proceeding to Chapter 7, and that $30 would be lost in liquidation, leaving only $70 for the debtholders. Suppose also that one month must pass between the formal proposal by the equityholders and a vote on that proposal. The equityholders would propose a plan that would give them almost $30 and leave slightly more than $70 for the debtholders. Even if the debtholders had suggested informally a plan that would give them more, say $75, the equityholders would have no reason to modify their plan, because they know that at the end of the month the debtholders would rationally vote to accept any amount greater than $70.

[26]Informal offers would serve a purpose, however, if the debtholders can commit not to give up more than a certain fraction of the surplus. For example, a bank may be a repeat player in Chapter 11 reorganizations and may have a reputation that enables it to make such a credible commitment. In such a case, the bank can insist on a particular fraction of the surplus and communicate informally its position. To succeed, however, the bank must have not only the ability to make informal offers but also the ability to make

debtholders have no private information (as is the case in our setting), an informal offer by them cannot convey any information.

In sum, the possibility of informal offers does not render exclusive control over the formal agenda meaningless. This formal control may well give an important bargaining advantage to the party exercising this control.[27] Therefore, we have sought to capture this effect in our model.

E. CONCLUDING REMARKS

This essay has sought to develop a sequential bargaining model of the Chapter 11 negotiation process, to show the effects of the legal rules that govern this process, and to determine the shares of the firm's value that equityholders and debtholders may expect to receive under a reorganization plan. The model has identified and analyzed three possible sources of the equityholders' power to obtain value under Chapter 11 even if the value of the firm's assets is less than its debt: if they were to delay or prevent agreement, (i) the firm would incur financial distress costs, (ii) the volatility of its assets' value may create some probability that the firm becomes solvent, and (iii) a Chapter 7 sale may entail a loss in value.

Our analysis has identified how various features of the company shape the Chapter 11 division of value. Our results regarding the effects of the company's features provide testable implications of the model. In particular, one could test the hypotheses that equityholders tend to capture a larger fraction of the value of the reorganized company when (i) the value of the company's assets is volatile (as measured, say, by the past volatility of the company's total stock and bond value), (ii) the nature of the company's business is such that financial distress costs are likely to be relatively high, or (iii) the total value of the reorganized company is a

---

such a commitment.

[27]Even pre-packaged plans formulated before filing for Chapter 11 would reflect the bargaining advantage that the equityholders would enjoy through their exclusive control over the agenda during the initial period under Chapter 11. See Section D.1.

relatively large fraction of the outstanding debt.[28]

We have also analyzed how the various features of the legal regime contribute to the equityholders' ability to extract value, and our analysis reveals how changes in this regime would affect the equityholders' share. The length of the reorganization period -- the time that the parties expect the court to allow them to remain in Chapter 11, without adopting a plan, before converting to Chapter 7 -- proves to be a critical feature. We find that any change in judicial attitudes that would shorten this period would decrease the equityholders' share. In addition, the equityholders benefit from their control over the agenda during the initial period; thus, any reduction in the length of this period would tend to shift the division of value in favor of the debtholders.

This essay has focused on a positive analysis of the formal reorganization process under the bankruptcy laws. This description of the distribution of value, however, has not addressed an important issue: the efficiency effects of Chapter 11. One alleged benefit of Chapter 11 that it avoids Chapter 7 liquidation, which would waste value. This possibility, which we introduced in Section C.3, is the subject of debate as noted. See Baird (1986) and Jackson (1986, ch. 9). Whatever the size of this benefit of Chapter 11, however, a full assessment of Chapter 11 requires an understanding of its costs. It is worth discussing briefly these efficiency costs and in particular how our positive model does some important preparatory

---

[28] Recent empirical studies of Chapter 11 have confirmed that equityholders obtain more than their contractual right, but have not tested any of these three hypotheses. Our preliminary analysis of the data presented in Table II of Eberhart, Moore, and Roenfeldt (1990, p. 1463) offers some support for hypothesis (iii). A simple linear regression on their 30 observations reveals that the extent to which senior claims (those of creditors and of preferred shareholders) exceed total value (expressed as a percent of total value) has a statisically significant negative effect upon the percent of total value paid to common shareholders: we find the relevant coefficient to be -.04727, with a standard error of .02428. If we restrict the sample to the 18 cases of insolvent companies, the regression again reveals a negative effect, but it is no longer statistically significant: we find the coefficient to be -.00729, with a standard error of .01267. A proper hypothesis test, however, would require data on all the explanatory variables that the model indicates we should include in the regression.

work for a complete understanding of these costs. Chapter 11 entails both financial distess costs after filing and lost value due to suboptimal management before filing. Both ex post and ex ante costs are important in evaluating proposed reforms in the law of corporate reorganization.

1. Ex Post Welfare Costs

In the model, because it was assumed that there was no asymmetric information, the resolution of the bargaining occurred in round 1. Note that because the rounds are not of negligible duration, the company incurred some financial distress costs: the firm lost α in value during round 1. Thus, even in this model with perfect information, the parties bear some limited efficiency costs.

Suppose one were to introduce asymmetric information about the firm's value V. For example, suppose that equityholders (or the management of the debtor who might be acting in their interests) and debtholders may have different estimates of the values of the parameters of the model, and that one's beliefs about these values are private information not known by others. In this case, the sequential bargaining literature suggests that reaching an agreement may well take some time (as indeed often happens in reality[29]) and that the parties will incur significant efficiency costs. See, e.g., Grossman and Perry (1986), Fudenberg and Tirole (1983), and Crampton (1984).[30]

The structure imposed upon the bargaining by Chapter 11 is crucial to this result. Gul and Sonnenschein (1988) point out that asymmetric information

---

[29]White (1984, pp. 35, 37) found the average length of the reorganization period, from bankruptcy filing to confirmation of a plan, to be 17 months. See also Franks and Torous (1989, p. 753).

[30]Since writing this essay, we learned of the work of Kaiser (1991) in the direction outlined above. He introduces asymmetric information in a bargaining model of reorganization and confirms that significant delay can result. Kaiser focuses on the possibility of delay and the factors that determine the length of this delay, not on the factors that determine the division of value between equityholders and debtholders. Thus, his paper complements this essay.

between bargaining parties cannot by itself explain delay in reaching agreement, because as the length of each round is allowed to go to 0, so does the total delay. This objection does not apply to the reorganization context, however, because the legal rules governing Chapter 11 bargaining imply that, unlike the case in other contexts, there is a limit to how short the bargaining rounds can be.

2. Ex Ante Welfare Costs

As the analysis has shown, Chapter 11 enables equityholders to obtain some value even if the firm is insolvent, that is, if the value of the firm is not enough to pay the debtholders' claims. Because the parties anticipate this outcome, Chapter 11 has ex ante effects. One ex ante effect is an initial interest rate on the debt that is higher than it would be otherwise. Chapter 11, then, does not actually create a net expected transfer from debtholders to equityholders, because the possibility of an ex post transfer is reflected in the interest rate chosen ex ante. Furthermore, these effects do not imply any efficiency consequences.

Chapter 11 has another ex ante effect, however. The equityholders anticipate obtaining some value in the event of insolvency, and this expectation may affect the way in which the management runs the firm ex ante. As Bebchuk (1991) shows, this expectation leads to inefficient ex ante management decisions, which should be regarded as a significant efficiency cost.[31]

[31]The above discussion suggests that Chapter 11 produces significant efficiency costs. For a proposal for changing Chapter 11 to eliminate these costs, see Bebchuk (1988). This proposal suggests a method for distributing the value of the reorganized firm. The distribution would be fully consistent with all the participants' contractual rights and would occur without delay. As a result, it would eliminate both the ex post and the ex ante efficiency costs of Chapter 11.

APPENDIX

In this appendix we derive several propositions, including our comparative statics results. To analyze the effect of the parameters on VE_1, it will prove useful to make some preliminary observations. Consider the general model in Section C.3, in which $\lambda \geq 0$ and $L \leq V_n$. (To apply the following reasoning to the model in Section C.2, which addresses a special case of the more general model, let $\lambda = 0$, $L = V_n$, and $VE_L = VE_n$.)

First, we derive a lemma pertaining to the "option value" component of VE_1. Note that VE_L is a continuous function of L-D. In particular, as the parameters change, VE_L will equal (and therefore increase and decrease one-for-one with) L-D if L>D, will remain constant at 0 if L<D, and will increase one-for-one (but not decrease) with L-D if L=D. Therefore, we can state the following lemma:

<u>Lemma A1</u>: For any parameter x that does not affect the probability distribution of the random variable k, i.e., for $D-V_1$, θ, and α, the following holds for increases in x:[32]

$$\frac{\partial E_1[VE_L]}{\partial x} = E_1\left[\frac{\partial VE_L}{\partial x}\right] \tag{A.1}$$
$$= \Pr(L{=}D)\max\left[0, \frac{\partial(L-D)}{\partial x}\Big|_{L=D}\right] + \Pr(L>D)E_1\left[\frac{\partial(L-D)}{\partial x}\Big| \; L>D\right].$$

Second, we will find it useful to place an upper limit on Pr(L>D). Given the symmetric probability distribution of k, if n is odd, then the median value for k is $E_1[k] = \frac{1}{2}(n-1)$. In that case, the median value for L is similarly $E_1[L] = V_1 - \alpha(n-1) - \lambda$, which must be less than D. Therefore, if n is odd, then $\Pr(L>D) < \frac{1}{2}$.

---

[32]The notation for the partial derivative of VE_L (or its expected value) in this lemma and in the proofs below will represent the limit of $\Delta VE_L/\Delta x$ as we approach x from the right side, in order to describe the effects of an increase in x. To describe decreases in x, we would approach x from the left side and would replace "max" in (A.1) with "min." These two limits are not equal (i.e., although $VE_L(x)$ is continuous, it is not differentiable) at values of x where Pr(L>D) changes, because this probability changes only in discrete amounts.

If n is instead even, then there is no unique median value for k. In this case, $\Pr(L>D) < \frac{1}{2}$ if and only if $L \leq D$ for $k = \frac{1}{2}n$. As one can see from (12), if $k = \frac{1}{2}n$, then $V_n = V_1 - \alpha(n-1) + \theta$. Therefore, if $(D-V_1) + \alpha(n-1) + \lambda \geq \theta$, then $\Pr(L>D) < \frac{1}{2}$. Furthermore, $\Pr(L>D) > \frac{1}{2}$ would require that $L>D$ for $k = \frac{1}{2}(n-2)$. As one can see from (12), however, if $k = \frac{1}{2}(n-2)$, then $V_n = V_1 - \alpha(n-1) - \theta$, which must be less than D. Therefore, $\Pr(L>D) > \frac{1}{2}$ is impossible. We summarize these results in the following lemma:

<u>Lemma A2</u>: $\Pr(L>D) \leq \frac{1}{2}$, and $\Pr(L>D) < \frac{1}{2}$ if and only if either $(D-V_1) + \alpha(n-1) + \lambda \geq \theta$ or n is odd.

Using these two lemmas, we can now provide proofs of the propositions stating our comparative statics results:

<u>Proof of Proposition 2:</u> Note that the derivative of V_n-D with respect to D-V_1 is -1 in any state of the world, i.e., for any k. Therefore, recalling that $VE_L = VE_n$ and $L = V_n$, and using Lemma A1:

$$\frac{\partial E_1[VE_n]}{\partial(D-V_1)} = -\Pr(V_n > D). \tag{A.2}$$

Thus, increases in D-V_1 will decrease the equityholders' "option value" until D-$V_1 \geq (n-1)(\theta-\alpha)$. At that point, by (15), the probability of $V_n > D$ (and the "option value") reaches 0, and by (A.2), increases in D-V_1 then will have no further effect. ■

<u>Proof of Proposition 3:</u> First, note that the partial derivative of V_n in (12) with respect to θ is 2k-(n-1), which is greater or less than 0 as k is greater or less than $\frac{1}{2}(n-1)$. That is, V_n increases in θ in those states of the world in which k is larger than $E_1[k] = \frac{1}{2}(n-1)$, which by (12) are those states in which V_n is larger than $E_1[V_n] = V_1 - \alpha(n-1)$. Because V_n can affect $E_1[VE_n]$ only in those states in which V_n exceeds or equals D, which in turn exceeds $V_1 - \alpha(n-1)$ (indeed, exceeds V_1), θ can affect $E_1[VE_n]$ only in those states in which V_n is increasing in θ. In the <u>relevant states</u>, therefore, 2k-(n-1) > 0, and a rise in θ can only increase V_n-D in the

expression for the equityholders' "option value" in (13). Recalling that $VE_L = VE_n$ and $L = V_n$, and using Lemma A1, then:

$$\frac{\partial E_1[VE_n]}{\partial \theta} = \Pr(V_n \geq D) E_1[2k-(n-1) \mid V_n \geq D]. \qquad (A.3)$$

Furthermore, by reasoning analogous to that used to derive (15), the following two conditions are equivalent:

$$\Pr(V_n \geq D) > 0 \leftrightarrow (n-1)(\theta-\alpha) \geq D-V_1. \qquad (A.4)$$

A rise in θ, by (A.3) and (A.4), will have no effect on the expected value of VE_n until $\theta \geq \alpha + (D-V_1)/(n-1)$; further increases will raise equityholders' "option value." ■

<u>Proof of Proposition 4:</u> First, note that the equityholders' "option value" is nonincreasing in α. A rise in α shifts the entire probability distribution of V_n downward and so reduces V_n-D in the expression for the equityholders' "option value" in (13). In all states of the world, i.e., for any k, the derivative of V_n-D with respect to α will be -(n-1). Recalling that $VE_L = VE_n$ and $L = V_n$, and using Lemma A1, then:

$$\frac{\partial E_1[VE_n]}{\partial \alpha} = -(n-1)\Pr(V_n > D). \qquad (A.5)$$

Thus, α will have no effect if $\Pr(V_n > D) = 0$ already. That is, by (15), an increase in α will reduce the "option value" until $\alpha \geq \theta + (V_1-D)/(n-1)$. Then the "option value" will be 0, and further increases in α will have no effect on it.

It will be useful to place an upper limit on the absolute value of the negative effect of α on the "option value." By Lemma A2, $\Pr(V_n > D)$ can be at most ½. Therefore, (A.5) implies that the derivative of $E_1[VE_n]$ with respect to α can range from 0 to -½(n-1).

The full effect of α on the equityholders' share, unlike the effect of θ or of D-V_1, will also depend on an effect on the "efficiency gains" component, i.e., the first term in (13). An increase in α has an unambiguously positive effect on this component of the equityholders' share: the derivative of $\frac{1}{2}\alpha(e+n-1)$ with respect to α is $\frac{1}{2}(e+n-1)$. This effect will always be greater than the absolute value of the negative effect on the "option value" component in (13), because $e>0$. ■

Proof of Proposition 5: By (15), the "option value" component of the equityholders' share will remain at 0 until $n > 1 + (D-V_1)/(\theta-\alpha)$. Once n rises above this level, their expected round n payoff rises above 0. Thus, at that point the "option value" increases in n.

The "option value," however, will not be monotonically nondecreasing in n. Instead, it may decrease in particular cases. Note that $E_1[V_n]=V_1-\alpha(n-1)$, so that each unit increase in n causes the expected value of V_n to fall by α. Indeed, an increase from $n=N$ to $n=N+1$ will reduce the expected value of V_n, conditional on any particular V_N, by the amount: $V_N - E_N[V_{N+1}] = \alpha$. Therefore, it is easy to construct cases in which the expected value of V_n, conditional on $V_n>D$, will fall with an increase from $n=N$ to $n=N+1$. The equityholders' expected round n payoff will then fall, provided that $Pr(V_n>D)$ does not rise.[33]

It will prove useful to place a ceiling on the absolute value of the negative effect that n may have on the equityholders' "option value." A unit rise in n cannot cause the expected value of V_n, conditional on $V_n>D$, to fall by more than α. Thus, a rise in n from N to N+1 can reduce $E_1[VE_n]$ by at most $\alpha Pr(V_N>D)$, because it can only reduce VE_n in those states of the world in which $V_n>D$. Recall also that

[33]As n increases, new probability mass will rise above D at regular intervals: first, the V_n associated with $k=n-1$ will exceed D, then that with $k=n-2$, and then that with $k=n-3$, and so forth. As one can see in (12), $V_n(k)$ falls by 2θ for each drop in k, and the V_n corresponding to each n-1-k will rise by $\theta-\alpha$ for each unit rise in n. Therefore, n must rise by $2\theta/(\theta-\alpha)$ to bring the next possible V_n above D. During these intervals, probability mass will be falling below D because each unit increase in n reduces the expected value of the V_n just above D; in particular, half of the probability mass at this marginal V_n falls below D. Once the marginal V_n (which by definition can be no more than $D+2\theta$) is at least $D+\theta-\alpha$ (so that the half of the probability mass that falls will at least offset the half that rises), but still no more than $D+\theta+\alpha$ (so that no new probability mass will rise above D), $E_1[VE_n]$ falls unambiguously with a unit increase in n.

Lemma A2, with $\lambda=0$ and $L=V_n$, implies that $Pr(V_n>D)$ can never exceed ½ and can equal ½ only if n is even. Therefore, if $Pr(V_n>D)$ equals ½, then a rise in n from N to N+1 results in an odd n, so the median V_{N+1} is strictly less than D, and $E_1[VE_n]$ must fall by strictly less than $\alpha Pr(V_N>D)$. Thus, a unit rise in n will cause the equityholders' "option value" to fall by an absolute value strictly less than ½α.

To consider the effect of n upon the "efficiency gains" component, note that the derivative of $\frac{1}{2}\alpha(e+n-1)$ with respect to n is ½α, which is unambiguously positive. This positive effect upon the "efficiency gains" component of the equityholders' share must be greater than the absolute value of the negative effect upon the "option value" component. Thus, each unit increase in n must cause VE_1 to increase. ■

<u>Proof of Proposition 7:</u> By the same reasoning used earlier in Section C -- backward induction from round n to round 1 -- the equityholders would obtain (4), i.e., $VE_1 = \frac{1}{2}\alpha(e+n-1) + E_1[VE_n]$. If $L<V_n$, however, then VE_n will not equal VE_L. Consider the reasoning used to derive (3) with respect to the financial distress costs, but substitute λ for α, round n for round i, and liquidation for round i+1. One can show that the party making the offer in round n would hold its opponent to its payoff from liquidation and thereby obtain the surplus from avoiding liquidation, λ, plus its own payoff from liquidation. In liquidation, the equityholders would receive $VE_L = \max(0, V_n-\lambda-D)$, and the debtholders would receive the complementary share $VD_L = L-VE_L = V_n-\lambda-\max(0, V_n-\lambda-D) = \min(V_n-\lambda, D)$.

Each party would make the offer in round n (and thus capture the surplus λ) with probability ½. In this case, $E_1[VE_n] = \frac{1}{2}\lambda + E_1[VE_L]$, and substituting this expression in (4) yields:

$$VE_1 = \tfrac{1}{2}\alpha(e+n-1) + \tfrac{1}{2}\lambda + E_1[VE_L]. \qquad \text{(A.6)}$$

Consider the reasoning used to prove Lemma 1, but applied to L rather than to V_n.

An analogous lemma implies that (A.6) equals (16).[34] ■

Proof of Proposition 8: The loss from liquidation, λ, affects VE_1 through both the second and the third terms in (16). Note that the derivative of L-D with respect to λ is -1 in any state of the world, i.e., for any k. Therefore, using Lemma A1:

$$\frac{\partial E_1[VE_L]}{\partial \lambda} = -Pr(L>D). \tag{A.7}$$

Bv (A.7), the derivative of the third term, $E_1[\max(0, L-D)]$, with respect to λ, is -Pr(L>D). The derivative of the second term with respect to λ, however, equals ½ if e<n (and equals 1 if e=n). The positive effect on the second term would be greater than or equal to the absolute value of the negative effect on the third term, because by Lemma A2, $Pr(L>D) \leq ½$. Therefore, VE_1 is nondecreasing in λ. Moreover, if either Pr(L>D) < ½ or e=n, then VE_1 is increasing in λ. Lemma A2, which states that if either $(D-V_1) + \alpha(n-1) + \lambda \geq \theta$, then Pr(L>D) < ½, completes the proof. ■

---

[34]If e=n rather than e<n, then only the equityholders could make the offer in round n and capture λ. In this case, $E_1[VE_n] = \lambda + E_1[VE_L]$. This value plus the $\alpha(n-1)$ they would capture from round 1 to round n-1 yields:

$$VE_1 = \alpha(n-1) + \lambda + E_1[VE_L].$$

In either case, the equityholders would have an additional source of bargaining power represented by the new second term.

REFERENCES

Baird, Douglas G. (1986), "The Uneasy Case for Corporate Reorganizations," Journal of Legal Studies, Vol. 15, 127-147.

Baird, Douglas G. and Thomas H. Jackson (1990), Cases, Problems, and Materials on Bankruptcy Law, 2d ed. (Boston: Little, Brown & Co.).

Baird, Douglas G. and Randal C. Picker (1991), "A Simple Noncooperative Bargaining Model of Corporate Reorganizations," Journal of Legal Studies (forthcoming).

Bebchuk, Lucian A. (1988), "A New Approach to Corporate Reorganizations," Harvard Law Review, Vol. 101, 775-804.

Bebchuk, Lucian A. (1991), "The Effects of Chapter 11 and Debt Renegotiation on Ex Ante Corporate Decisions," Harvard Law School, mimeo.

Bergman, Yaacov Z. and Jeffrey L. Callen (1991), "Opportunistic Behavior in Debt Renegotiations and an Interior Optimal Capital Structure of the Firm Without Deadweight Costs," Journal of Financial Economics (forthcoming).

Berkovitch, Elazar and Ronen Israel (1991), "The Bankruptcy Decision and Debt Contract Renegotiation," University of Michigan, School of Business Administration, mimeo.

Broude, Richard F. (1990), Reorganizations Under Chapter 11 of the Bankruptcy Code (New York: Law Journal Seminars-Press).

Brown, David T. (1989), "Claimholder Incentive Conflicts in Reorganization: The Role of Bankruptcy Law," Review of Financial Studies, Vol. 2, 109-123.

Crampton, Peter C. (1984), "Bargaining with Incomplete Information: An Infinite-Horizon Model with Two-Sided Uncertainty," Review of Economic Studies, Vol. 51, 579-593.

Cutler, David M. and Lawrence H. Summers (1988), "The Costs of Conflict Resolution and Financial Distress: Evidence from the Texaco-Pennzoil Litigation," Rand Journal of Economics, Vol. 19, 157-172.

Eberhart, Allan C., William T. Moore, and Rodney L. Roenfeldt (1990), "Security Pricing and Deviations from the Absolute Priority Rule in Bankruptcy Proceedings," Journal of Finance, Vol. 45, 1457-1469.

Franks, Julian R. and Walter N. Torous (1989), "An Empirical Investigation of U.S. Firms in Reorganization," Journal of Finance, Vol. 44, 747-769.

Fudenberg, Drew and Jean Tirole (1983), "Sequential Bargaining with Incomplete Information," Review of Economic Studies, Vol. 50, 221-247.

Gertner, Robert and David Scharfstein (1991), "A Theory of Workouts and the Effects of Reorganization Law," University of Chicago, Graduate School of Business and Law School, and Massachusetts Institute of Technology, Sloan School of Management, mimeo.

Giammarino, Ronald M. (1989), "The Resolution of Financial Distress," Review of Financial Studies, Vol. 2, 25-47.

Gilson, Stuart C., Kose John, and Larry H.P. Lang (1990), "Troubled Debt Restructurings: An Empirical Study of Private Reorganization of Firms in Default," Journal of Financial Economics, Vol. 27, 315-353.

Grossman, Sanford J. and Motty Perry (1986), "Sequential Bargaining under Asymmetric Information," Journal of Economic Theory, Vol. 39, 120-154.

Gul, Faruk and Hugo Sonnenschein (1988), "On Delay in Bargaining with One-Sided Uncertainty," Econometrica, Vol. 56, 601-611.

Hart, Oliver and John Moore (1989), "Default and Renegotiation: A Dynamic Model of Debt," Massachusetts Institute of Technology, Department of Economics, Working Paper No. 520.

Hart, Oliver and John Moore (1990), "A Theory of Corporate Financial Structure Based on the Seniority of Claims," Massachusetts Institute of Technology, Department of Economics, Working Paper No. 560.

Jackson, Thomas H. (1986), The Logic and Limits of Bankruptcy Law (Cambridge, Mass.: Harvard University Press).

Jensen, Michael C. and William H. Meckling (1976), "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," Journal of Financial Economics, Vol. 3, 305-360.

Kaiser, Kevin M.J. (1991), "A Bargaining Model of Chapter 11 Reorganization," Northwestern University, J.L. Kellogg Graduate School of Management, mimeo.

Myers, Stewart C. (1977), "Determinants of Corporate Borrowing," Journal of

Financial Economics, Vol. 5, 147-175.

Normandin, Charles P. (1989), "The Changing Nature of Debt and Equity: A Legal Perspective," in Are the Distinctions Between Debt and Equity Disappearing?, edited by Kopcke, Richard W. and Eric S. Rosengren (Boston: Federal Reserve Bank of Boston), 49-66.

Roe, Mark J. (1987), "The Voting Prohibition in Bond Workouts," Yale Law Journal, Vol. 97, 232-279.

Rubinstein, Ariel (1982), "Perfect Equilibrium in a Bargaining Model," Econometrica, Vol. 50, 97-109.

Trost, J. Ronald (1979), "Business Reorganizations Under Chapter 11 of the New Bankruptcy Code," The Business Lawyer, Vol. 34, 1309-1346.

Webb, David C. (1987), "The Importance of Incomplete Information in Explaining the Existence of Costly Bankruptcy," Economica, Vol. 54, 279-288.

Weiss, Lawrence A. (1990), "Bankruptcy Resolution: Direct Costs and Violation of Priority of Claims," Journal of Financial Economics, Vol. 27, 285-314.

White, Michelle J. (1984), "Bankruptcy Liquidation and Reorganization," Ch. 35 of Handbook of Modern Finance, edited by Logue, Dennis E. (Boston: Warren, Gorham, and Lanort).

CHAPTER III:

AN ECONOMIC ANALYSIS OF RULE 11 AND FEE-SHIFTING BASED ON THE MARGIN OF VICTORY

Lucian Arye Bebchuk and Howard F. Chang

A. INTRODUCTION

Under fee-shifting rules a court can require the losing litigant to pay the attorneys' fees of the winning litigant. This essay analyzes the effect of fee-shifting rules that are based not only upon the identity of the winning party but also on how strong the court perceives the case to be at the end of the trial -- that is, the "margin of victory." In particular, the analysis examines the effect upon a plaintiff's incentive to bring suit and reveals how one can design such a rule to induce plaintiffs to sue only if they believe their cases are sufficiently strong.

This analysis of fee-shifting rules based on the margin of victory is not only of theoretical interest but also of practical significance: courts have interpreted Federal Rule of Civil Procedure 11 as an example of such a rule. Rule 11 requires an attorney to sign pleadings, motions, and other papers before filing them in court, thereby certifying that to the best of the attorney's knowledge "after reasonable inquiry" the paper:

> is well grounded in fact and is warranted by existing law or a good faith argument for the extension, modification, or reversal of existing law, and that it is not interposed for any improper purpose, such as to harass or to cause unnecessary delay or needless increase in the cost of litigation.... If a pleading motion, or other paper is signed in violation of this rule, the court, upon motion or upon its own initiative, shall impose upon the person who signed it, a represented party, or both, an appropriate sanction, which may include an order to pay to the other party or parties the amount of the reasonable expenses incurred because of the filing of the pleading, motion, or other paper, including a reasonable attorney's fee.

Fed. R. Civ. P. 11. For example, the rule allows the court to shift the burden of the defendant's fees to the plaintiff if the plaintiff's case is so frivolous that the court finds that the plaintiff's attorney should have known the suit was without merit when filed. In enforcing Rule 11, courts often focus on the merits of claims and defenses in this way.[1] Furthermore, in the overwhelming majority of cases

---

[1]Schwarzer (1988), however, argues that courts should shift their scrutiny from the merits to the adequacy of the attorney's prefiling inquiry. Because it is difficult to predict whether a court will find a particular paper "frivolous," see Levinson (1986), Schwarzer argues that such a shift from merits to attorney conduct would promote more predictable

imposing sanctions under Rule 11, courts have punished the party filing the "frivolous" paper by awarding costs and fees to the opposing party. See Nelken (1986, p. 1333).[2]

Suppose that society seeks to induce plaintiffs to sue if and only if they believe they are entitled to prevail at trial. (The analysis of this essay is more general, that is, also allows for other objectives.) This objective implies two goals: that plaintiffs bring "good" suits -- those in which the plaintiff expects to prevail -- and that plaintiffs not bring "bad" suits -- those in which the plaintiff does not expect to prevail. How can fee-shifting rules accomplish these goals?

One important consideration is the fact that the outcome of a trial is unlikely to be certain to the plaintiff when it decides whether to sue. The court can lack information available to the plaintiff or can err in its judgment given the information that it can observe.[3] The plaintiff can also be wrong; for example, it might lack information that a trial would later reveal to the court.

Consider first the standard American rule, under which each litigant bears its own expenses. This rule does not induce optimal litigation decisions. First, plaintiffs will not bring all good suits. Even if the plaintiff can count on the court to decide the case as the plaintiff predicts, the plaintiff will not sue if its litigation costs exceed the amount that it expects to recover. Furthermore, if the plaintiff

---

and less costly enforcement of Rule 11. As we will see, however, attorneys do not need to predict sanctions with certainty for Rule 11 to exert the desired deterrent effect.

[2]Thus, the Court of Appeals for the Seventh Circuit has stated that "Rule 11 is a fee-shifting statute." Hays v. Sony Corp., 847 F.2d 412, 419 (7th Cir. 1988). Burbank (1989), however, criticizes this interpretation of Rule 11 and warns against routine use of expense-shifting as a sanction. He argues that courts should exercise greater discretion in selecting a sanction sufficient to deter in the particular circumstances of the case. This essay shows how courts can tailor the deterrent effect by varying the conditions under which sanctions would be imposed, not by varying the magnitude of the sanctions.

[3]Polinsky and Shavell (1989) stressed the important effect of legal uncertainty caused by judicial error upon the incentive to sue. They do not study, however, how fee-shifting based on the margin of victory can address the problem of judicial error (and other reasons for the unpredictability of judgment).

believes that the court's decision might differ from what the plaintiff expects, either because the court might err or because the plaintiff might err, then the plaintiff will bring some suits that are expected to lose. If the litigation costs are small enough, the plaintiff will find it worthwhile to gamble -- either because the court might err in the plaintiff's favor or because the case might prove to be better than it first appears.

Consider next the British rule, under which the loser pays the expenses of the winner. If the plaintiff could predict the trial outcome without error, then the plaintiff would not bring a bad suit and would never be discouraged by litigation expenses from bringing a good suit, which would guarantee the plaintiff reimbursement of its litigation expenses. If the plaintiff cannot predict the trial outcome with certainty, however, then the plaintiff will not always make litigation decisions consistent with the goals suggested above. First, if litigation costs are small enough, then the plaintiff will bring some bad suits because it might win at trial. Second, if litigation costs are sufficiently large, then the prospect of losing (and bearing the expenses of both litigants) will deter the plaintiff from bringing some good suits because it might lose at trial.

This essay shows how to design fee-shifting rules so as to induce optimal litigation decisions. The fee-shifting should depend not only on which party prevails but also on the margin by which they prevailed. That is, the rule should take into account not only who won but also the degree to which they won easily. This information is useful because a plaintiff who loses by a large margin is less likely to have believed ex ante that the case was good than a plaintiff who loses by a small margin. Similarly, a plaintiff who wins by a large margin is less likely to have believed ex ante that the case was bad than a plaintiff who wins by a small margin.

Given the distribution of the errors in the plaintiff's prediction of the trial outcome, it is possible to design a fee-shifting rule that would induce optimal litigation decisions. Because the optimal rule depends on this distribution, a court can implement such a rule as long as it has some sense of how these errors are distributed. This essay examines the structure of such rules and how they affect litigation decisions.

Section B of this essay presents the formal model of litigation that we shall use to analyze the plaintiff's incentives to bring suit. Section C examines those incentives under the classic fee-shifting rules, which allocate litigation costs according to the identity of the losing party. In Section D, we allow the fee-shifting rule to depend upon the margin of victory as well, and we show how one can thereby provide the plaintiff with the optimal incentives to bring suit. Section E shows that if the court can shift the fees of either litigant in each case, then there exists a whole family of such fee-shifting rules that can present plaintiffs with the optimal incentives. Section F explores some extensions of our model. The concluding remarks in Section G summarize our results and consider the implications of our analysis for the interpretation of Rule 11.

B. THE MODEL

Let x be a parameter describing the defendant's behavior, such as the defendant's level of care, and let x_c represent the value of x that is observed by the court and that forms the basis for the court's judgment. Let X represent the standard of liability: the court will find the defendant liable if and only if $x_c < X$. For example, the court will hold the defendant liable for negligence if the observed level of care is less than the standard of due care. Let D represent the damages that the court will require the defendant to pay the plaintiff if the court finds the defendant liable, where $D > 0$.

Except in Section C.1, we assume that the court errs by a random amount e, so that $x_c = x + e$. The random variable e is distributed according to a cumulative distribution function F(e) defined over the domain $(-\infty, \infty)$. For any particular realization of e, the corresponding F(e) will equal the probability ex ante that the error takes on a value strictly less than this realized value of e. Let e take on positive and negative values, each with some positive probability, so that $0 < F(0) < 1$. For simplicity, let F(e) be continuous and strictly increasing while it ranges in value from 0 up to 1.

Let x_p be the plaintiff's observation of x when deciding whether or not to bring suit. At this stage, we assume that the plaintiff knows the true value of x; that is, $x_p=x$. Section F.1 explains how the analysis can be extended easily to the case in which the plaintiff is imperfectly informed about x. In particular, the court may have information available at trial that is unavailable to the plaintiff.

Suppose that the social objective is to induce the plaintiff to sue if and only if $x_p<x^*$, where x^* is some threshold. This general statement of the objective is reasonable: we cannot have the plaintiff act on anything but its impression of its case, and presumably if we want to deter suits for a certain value of x_p, we also want to deter suits for any higher values. At this stage, we assume that $x^*=X$. That is, the plaintiff should sue if and only if the plaintiff believes that the defendant is liable. Section F.2 shows how the analysis can be adjusted for the case in which x^* does not equal X.

Let c_p and c_d be the litigation costs of the plaintiff and of the defendant, respectively, where $c_p>0$ and $c_d>0$. The court observes x_c before reaching judgment and also knows D, F(e), c_p, and c_d when it applies the fee-shifting rule. When the plaintiff decides whether to sue, it knows x_p, D, F(e), c_p, c_d, and the fee-shifting rule. The plaintiff is risk neutral and sues if and only if the expected value of bringing the suit is positive.

C. FEE-SHIFTING RULES BASED ON THE WINNER'S IDENTITY

The four classic fee-shifting rules, as described in Shavell (1982), are characterized by the fact that fee-shifting depends only on the identity of the winning party -- that is, on how x_c compares with X. In addition to the "two-sided" fee-shifting rules described above, the American rule and the British rule, Shavell discusses two "one-sided" fee-shifting rules: (1) under the pro-plaintiff rule, each litigant pays its own costs if the plaintiff loses, but the defendant pays the plaintiff's costs if the plaintiff wins, and (2) under the pro-defendant rule, each litigant pays its own costs if the defendant loses, but the plaintiff pays the defendant's costs if the defendant wins. We shall examine the effects of each rule in turn under two alternative assumptions regarding the plaintiff's ability to predict the outcome of a trial.

1. Judgment Predicted With Certainty

In this section we assume that the plaintiff can predict the judgment with certainty. That is, e is not a random variable; instead, $e=0$ in each case. Consider the effect of each fee-shifting rule upon the incentives of the plaintiff to bring suit.

a. The American Rule and the Pro-Defendant Rule: Under either the American or the pro-defendant rules, the plaintiff will sue if and only if both $x_p<X$ and $c_p<D$. A plaintiff will never bring a bad suit, because a case with $x_p \geq X$ would be bound to lose. Thus, the plaintiff would recover nothing and would be saddled with at least its own litigation costs. A plaintiff might fail to bring a good suit, however, if its litigation costs are sufficiently large. Even a winning suit, with $x_p<X$, would not be worthwhile if $c_p>D$, because the plaintiff would not recover enough to pay its litigation costs.

b. The British Rule and the Pro-Plaintiff Rule: Under either the British or the pro-plaintiff rules, the plaintiff will sue if and only if $x_p<X$. Without uncertainty over the trial outcome, these rules yield the optimal incentives for the plaintiff. A plaintiff will never bring a bad suit, because it would be bound to lose. Thus, the plaintiff would recover no damages and bear at least its own litigation costs. A plaintiff will always bring a good suit, because it would be bound to win, and in this case the plaintiff would not have to bear its own litigation costs.

2. Judgment Predicted with Uncertainty

In this section we assume that the plaintiff cannot predict the trial outcome with certainty. That is, e is a random variable. As we shall see, none of the classic rules can guarantee that the plaintiff will have optimal incentives in all cases.

a. The American Rule: Under the American rule, the plaintiff will sue if and only if:

$$-c_p + \Pr(x_c<X|x_p)D > 0. \qquad (1)$$

Substituting x_p+e for x_c in (1), we find this condition is equivalent to:

$$\Pr(e<X-x_p) > c_p/D. \qquad (2)$$

Thus, if $c_p \geq D$, then (2) cannot hold, and the plaintiff will never sue. If $c_p<D$, however, the plaintiff will sue if and only if x_p is less than some threshold value, which we shall denote as y^*. This y^* is defined by:

$$F(X-y^*) = c_p/D,$$

or equivalently:

$$y^* = X - F^{-1}(c_p/D), \qquad (3)$$

where F^{-1}, the inverse function of F(e), is defined over the domain (0, 1).

The plaintiff will have optimal incentives if and only if it is just indifferent about bringing the marginal suit (in which $x_p=X$), that is, if $y^*=X$. Thus, (3) implies that the American rule will lead to optimal incentives for the plaintiff if and only if $F(0) = c_p/D$. For example, if $F(0)=½$, then the critical value for c_p/D is ½. Note that F(0) is the probability that the court will view the plaintiff's case more favorably than the plaintiff does, so that F(0) is the probability that a case with $x_p=X$ would succeed.

As in the case of prediction with certainty, this rule might discourage a plaintiff from bringing a good suit if the plaintiff's litigation costs are sufficiently large relative to the damages at stake. The possibility that even a good suit can lose, however, aggravates this problem. Specifically, if $F(0) < c_p/D$, then too little litigation results. In these cases, either $c_p \geq D$ or $y^*<X$, and the plaintiff will be discouraged from bringing some good suits because c_p would be too large relative to D.

Once we allow for prediction with uncertainty, moreover, it is no longer true that this rule would discourage all bad suits. The plaintiff might bring a bad suit if the plaintiff's litigation costs are sufficiently small relative to the damages at stake, because even a bad suit might prevail. Specifically, if $F(0) > c_p/D$, then too much litigation results. In these cases, $y^*>X$ and the plaintiff would bring some bad suits, because the likelihood of favorable judgments in these cases would be sufficiently high.

b. The Pro-Defendant Rule: Under the pro-defendant rule, the plaintiff will sue if and only if:

$$-(c_p+c_d) + \Pr(x_c<X|x_p)(D+c_d) > 0. \tag{4}$$

Substituting x_p+e for x_c in (4), we find this condition is equivalent to:

$$\Pr(e<X-x_p) > (c_p+c_d)/(D+c_d). \tag{5}$$

Thus, if $c_p \geq D$, then (5) cannot hold, and the plaintiff would never sue. If $c_p<D$, however, then the plaintiff will sue if and only if x_p is less than some y^*, which is defined by:

$$F(X-y^*) = (c_p+c_d)/(D+c_d),$$

or equivalently:

$$y^* = X - F^{-1}[(c_p+c_d)/(D+c_d)]. \tag{6}$$

Thus, this rule will lead to optimal incentives for the plaintiff if and only if $F(0) = (c_p+c_d)/(D+c_d)$.

Again, as in the case of prediction with certainty, this rule might discourage a plaintiff from bringing a good suit if its litigation costs are sufficiently large. The possibility that even a good suit can lose, however, aggravates this problem. Specifically, if $F(0) < (c_p+c_d)/(D+c_d)$, then too little litigation results. In these cases, either $c_p \geq D$ or $y^*<X$, and plaintiffs will be discouraged from bringing some good suits because the litigation costs that the plaintiff would expect to bear would be too large.

Once we allow for prediction with uncertainty, however, it is no longer true that this rule would discourage all bad suits. The plaintiff might bring a bad suit because even a bad suit might prevail. Specifically, if $F(0) > (c_p+c_d)/(D+c_d)$, then too much litigation results. In these cases, $y^*>X$ and the plaintiff will bring some bad suits, because the likelihood of favorable judgments in these cases would be sufficiently high.

c. The Pro-Plaintiff Rule: Under the pro-plaintiff rule, the plaintiff would sue if and only if:

$$-c_p + \Pr(x_c<X|x_p)(D+c_p) > 0. \tag{7}$$

Substituting x_p+e for x_c in (7), we find this condition is equivalent to:

$$Pr(e<X-x_p) > c_p/(D+c_p). \quad (8)$$

The plaintiff will sue if and only if x_p is less than some y^*, which is defined by:

$$F(X-y^*) = c_p/(D+c_p),$$

or equivalently:

$$y^* = X - F^{-1}[c_p/(D+c_p)]. \quad (9)$$

Thus, this rule will lead to optimal incentives for the plaintiff if and only if $F(0) = c_p/(D+c_p)$. Once we allow for prediction with uncertainty, this rule no longer ensures that all good suits are encouraged and that all bad suits are discouraged.

This rule might discourage a plaintiff from bringing a good suit if the plaintiff's litigation costs are sufficiently large relative to the damages at stake, because even a good suit might lose. Specifically, if $F(0) < c_p/(D+c_p)$, then too little litigation results. In these cases, $y^*<X$, and the plaintiff will be discouraged from bringing some good suits because c_p would be too large relative to D.

Furthermore, this rule might encourage a plaintiff to bring a bad suit if the plaintiff's litigation costs are sufficiently small, because even a bad suit might prevail. Specifically, if $F(0) > c_p/(D+c_p)$, then too much litigation results. In these cases, $y^*>X$ and the plaintiff will bring some bad suits, because the likelihood of favorable judgments in these cases would be sufficiently high.

d. The British Rule: Under the British rule, the plaintiff will sue if and only if:

$$-(c_p+c_d) + Pr(x_c<X|x_p)(D+c_p+c_d) > 0. \quad (10)$$

Substituting x_p+e for x_c in (10), we find this condition is equivalent to:

$$Pr(e<X-x_p) > (c_p+c_d)/(D+c_p+c_d). \quad (11)$$

The plaintiff would sue if and only if x_p is less than some y^*, which is defined by:

$$F(X-y^*) = (c_p+c_d)/(D+c_p+c_d),$$

or equivalently:

$$y^* = X - F^{-1}[(c_p+c_d)/(D+c_p+c_d)]. \quad (12)$$

Thus, this rule will lead to optimal incentives for the plaintiff if and only if $F(0) = (c_p+c_d)/(D+c_p+c_d)$. Again, once we allow for prediction with uncertainty, this rule no longer ensures that all good suits are encouraged and that all bad suits are discouraged.

This rule might discourage a plaintiff from bringing a good suit if the parties' litigation costs are sufficiently large relative to the damages at stake, because even a good suit might lose. Specifically, if $F(0) < (c_p+c_d)/(D+c_p+c_d)$, then too little litigation results. In these cases, $y^* < X$, and the plaintiff will be discouraged from bringing some good suits because the costs that it would bear in the event that the suit loses, c_p+c_d, will be sufficiently large relative to the gain in the event the suit wins, D.

Furthermore, this rule might encourage a plaintiff to bring a bad suit if the plaintiff's litigation costs are sufficiently small, because even a bad suit might prevail. Specifically, if $F(0) > (c_p+c_d)/(D+c_p+c_d)$, then too much litigation results. In these cases, $y^* > X$ and the plaintiff will bring some bad suits, because the likelihood of favorable judgments in these cases would be sufficiently high.

e. Comparison of the Rules: We can now compare the incentive for the plaintiff to sue under the four rules discussed above. We can summarize our results from equations (3), (6), (9), and (12), respectively, as follows:

Rule:	Critical Value for F(0):
American	c_p/D
Pro-Defendant	$(c_p+c_d)/(D+c_d)$
Pro-Plaintiff	$c_p/(D+c_p)$
British	$(c_p+c_d)/(D+c_p+c_d)$

If F(0) is below this critical value, then the plaintiff has too little incentive to bring suit; if F(0) is above, then the plaintiff has too great an incentive.

As long as $c_p<D$, so that plaintiffs bring some suits under each rule, a smaller critical value for F(0) implies that a higher threshold x_p is necessary to discourage the plaintiff from bringing suit. Thus, the incentive to sue is the greatest under the pro-plaintiff rule insofar as (9) implies that the threshold x_p is the highest under this rule. The pro-plaintiff rule implies a threshold x_p that is strictly larger than that under any other rule. Similarly, if $c_p<D$, then we can see

from (6) that the pro-defendant rule implies a threshold x_p that is strictly smaller than that under any other rule. If $c_p \geq D$ instead, then the incentive to sue would be smallest under the American rule and under the pro-defendant rule: we know from (2) and (5) that in this case the plaintiff would never sue under either rule.

Between the British rule and the American rule, however, it is ambiguous which rule implies the higher threshold x_p. Once we allow for prediction with uncertainty, it is no longer clear whether the British rule or the American rule offers the plaintiff the greater incentive to sue. Comparing (3) and (12), we find that the incentive is greater under the British rule if and only if:

$$(c_p+c_d)/(D+c_p+c_d) < c_p/D,$$

or equivalently:

$$c_p/c_d > D/(c_p+c_d). \qquad (13)$$

Most important, none of the four rules can ensure a threshold such that $y^*=X$ for all cases. Under each rule, plaintiffs bring some bad suits and fail to bring some good suits. That is, once we allow for prediction with uncertainty, a fee-shifting rule must use more information to ensure optimal incentives to bring suit.

D. THE OPTIMAL ONE-SIDED FEE-SHIFTING RULE

In this section we drop the requirement that fee-shifting be based solely on the identity of the winner, and allow the court to take into account the margin of victory. We show that as long as $F(0) < (c_p+c_d)/D$, there is always some one-sided fee-shifting rule that implies optimal incentives to bring suit. This rule allows only one side to recover its litigation expenses from the other if that side does well enough, that is, better than some threshold. As we shall see in Section V, as long as litigation costs are sufficiently large relative to the amount of damages at stake, there is a whole family of two-sided fee-shifting rules that can present plaintiffs with the optimal incentives.

Consider the American rule: we can distinguish between those situations in which this rule leads to too much litigation and those in which it leads to too little litigation. If $c_p/D < F(0)$, then $X<y^*$ and plaintiffs sue too often. If $F(0) < c_p/D$

< 1, then $y^* < X$. In this case (and in the case of $D \le c_p$), plaintiffs do not sue often enough.

1. Excessive Incentive to Litigate Under the American Rule

If $c_p/D < F(0)$, then the American rule leads to excessive litigation, and one might consider the British rule as an alternative. As we have seen in Section III.B.5, however, the British rule would create even greater incentives to bring suit in those cases in which (13) holds. A better solution would be a "refined" pro-defendant rule:

Proposition 1: If $c_p/D < F(0) < (c_p+c_d)/D$, then there exists a refined pro-defendant rule that would create the optimal incentives for the plaintiff to bring suit. This rule would specify that the plaintiff pays the defendant's litigation costs if $x_c \ge Y$, where:

$$Y = X + F^{-1}\{[c_p+c_d-F(0)D]/c_d\}. \tag{14}$$

Proof: Under the proposed rule, the plaintiff would sue if and only if:

$$-(c_p+c_d) + \Pr(x_c<Y|x_p)c_d + \Pr(x_c<X|x_p)D > 0.$$

Thus, the plaintiff will sue if and only if $x_p<y^*$, where y^* is defined by:

$$-(c_p+c_d) + F(Y-y^*)c_d + F(X-y^*)D = 0. \tag{15}$$

Thus, to ensure that $y^*=X$, we must set Y such that:

$$-(c_p+c_d) + F(Y-X)c_d + F(0)D = 0. \tag{16}$$

Note that the left-hand side of (16) is monotonically increasing in Y as long as $0 < F(Y-X) < 1$. If we let Y approach ∞, then $F(Y-X)$ goes to 1, so that the left-hand side of (16) must be positive, and plaintiffs would sue too often. If we let Y approach $-\infty$, then $F(Y-X)$ goes to 0. In this case, the left-hand side of (16) must become negative, so that plaintiffs will not sue often enough, if and only if the following inequality also holds:

$$F(0)D < c_p+c_d. \tag{17}$$

Thus, there exists a Y that solves (16) if and only if (17) holds; otherwise, there would always be too much litigation. Assuming such a Y exists, solving (16) for

Y yields the expression for Y given in the proposition. ■

Remark: Note that if litigation costs are sufficiently small relative to the damages at stake, so that (17) fails to hold, then no fee-shifting rule can provide optimal incentives for the plaintiff. In this case, even if the defendant had exercised due care (so that x=X), the plaintiff's expected payoff from trial would exceed the litigation costs of both parties. Even if the plaintiff were certain that it would have to pay the fees of both parties, these costs would not deter the plaintiff from bringing such a suit. In such cases, fee-shifting is an inadequate sanction, and a court would have to impose additional penalties to deter such suits. For example, under Rule 11 courts can impose fines or nonmonetary sanctions in addition to fees and costs. See Schwarzer (1985, pp. 201-04). Sufficiently large sanctions would replace c_d in (16) and (17) with some greater penalty so that (17) holds and a solution for Y in (16) exists.

Note also that if Y is greater than or equal to X, then the plaintiff would pay the defendant's fees only if the plaintiff loses. The rule in Proposition 1, however, can offer optimal incentives to the plaintiff consistent with Y≥X if and only if:

$$[c_p+c_d-F(0)D]/c_d \geq F(0),$$

or equivalently:

$$(c_p+c_d)/(c_d+D) \geq F(0). \tag{18}$$

Recall that if and only if (18) does not hold, then the classic pro-defendant rule (which is a special restricted case of our pro-defendant rule, with a constraint Y=X imposed) leads to too much litigation. In this case, then Y<X is necessary to reduce the plaintiff's incentives. This proposed rule is even more pro-defendant than the classic pro-defendant rule, so that in some cases the proposed rule shifts the defendant's fees to the plaintiff even though the plaintiff prevails.

As with condition (17), however, a court could substitute a greater sanction for c_d in (18) such that the inequality in (18) would hold. Because (18) is a stronger condition than (17), a greater sanction would be necessary to satisfy (18) than to satisfy (17). A sufficiently large sanction could provide optimal incentives for the plaintiff, even under a rule subject to the constraint Y≥X.

2. Insufficient Incentive to Litigate Under the American Rule

If $F(0) < c_p/D$, then the American rule leads to insufficient litigation. A better solution to this problem than the British rule would be a "refined" pro-plaintiff rule:

Proposition 2: *If $F(0) < c_p/D$, then there exists a refined pro-plaintiff rule that would create the optimal incentives for the plaintiff to bring suit. This rule would specify that the defendant pays the plaintiff's litigation costs if $x_c<Y$, where:*

$$Y = X + F^{-1}\{[c_p-F(0)D]/c_p\}. \tag{19}$$

Proof: Under the proposed rule, the plaintiff would sue if and only if:

$$-c_p + \Pr(x_c<Y|x_p)c_p + \Pr(x_c<X|x_p)D > 0.$$

Thus, the plaintiff would sue if and only if $x_p<y^*$, where y^* is defined by:

$$-c_p + F(Y-y^*)c_p + F(X-y^*)D = 0. \tag{20}$$

Thus, to ensure that $y^*=X$, we must set Y such that:

$$-c_p + F(Y-X)c_p + F(0)D = 0. \tag{21}$$

Note that the left-hand side of (21) is monotonically increasing in Y as long as $0 < F(Y-X) < 1$. If we let Y approach ∞, then $F(Y-X)$ goes to 1, so that the left-hand side of (21) must become positive, so that plaintiffs would sue too often. If we let Y approach $-\infty$, then $F(Y-X)$ goes to 0, so that the left-hand side of (21) must be negative, and plaintiffs would not sue often enough. Thus, there exists a Y that solves (21). Specifically, solving (21) for Y yields the expression for Y given in the proposition. ■

Remark: Note that if Y is less than or equal to X, then the defendant would pay the plaintiff's fees only if the plaintiff prevails. The rule in Proposition 2, however, can offer optimal incentives to the plaintiff consistent with $Y \leq X$ if and only if:

$$[c_p-F(0)D]/c_p \leq F(0),$$

or equivalently:

$$c_p/(c_p+D) \leq F(0). \tag{22}$$

Recall that if and only if (22) does not hold, then the classic pro-plaintiff rule (which is a special restricted case of our pro-plaintiff rule, with a constraint $Y=X$ imposed) leads to too little litigation. In this case, then $Y>X$ is necessary to increase the plaintiff's incentives. This proposed rule is even more pro-plaintiff than the classic pro-plaintiff rule, so that in some cases the proposed rule shifts the plaintiff's fees to the defendant even though the plaintiff loses.

3. Example with a Uniform Distribution

Suppose e is uniformly distributed in the interval (-E, E), for some $E>0$. In this special case,

$$F(e) = \tfrac{1}{2}(1 + e/E) \tag{23}$$

for e in the interval (-E, E), and $F(0) = \tfrac{1}{2}$. Solving (23) for the inverse function, we find:

$$F^{-1}(p) = (2p-1)E \tag{24}$$

for p in the interval (0, 1). Using this particular inverse function in (14) and (19), we can derive the optimal fee-shifting rules.

Proposition 1 implies that if $c_p/D < \tfrac{1}{2} < (c_p+c_d)/D$, then the optimal rule specifies that the plaintiff pays the defendant's litigation costs if $x_c \geq Y$, where:

$$Y = X + [(2c_p+2c_d-D)/c_d - 1]E,$$

or equivalently:

$$Y = X + (2c_p+c_d-D)E/c_d. \tag{25}$$

This rule operates only against losing plaintiffs (that is, $Y \geq X$) if and only if $D \leq 2c_p+c_d$. Proposition 2 implies that if $\tfrac{1}{2} < c_p/D$ instead, then the optimal rule would specify that the defendant pays the plaintiff's litigation costs if $x_c<Y$, where:

$$Y = X + [(2c_p-D)/c_p - 1]E,$$

or equivalently:

$$Y = X + (c_p-D)E/c_p. \tag{26}$$

This rule operates only against losing defendants (that is, $Y \leq X$) if and only if $c_p \leq D$. Note also that the effect of greater uncertainty over the trial outcome is ambiguous: in both (25) and (26) an increase in E could cause Y to either rise or fall, depending upon whether Y was greater or less than X.

E. THE FAMILY OF OPTIMAL FEE-SHIFTING RULES

In this section we examine the possibility of fee-shifting rules under which either side may reimburse the other for its litigation costs. We find that there exists a whole family of fee-shifting rules of this general form that can ensure that the plaintiff brings suit if and only if the case is sufficiently strong:

Proposition 3: If $F(0) < (c_p+c_d)/D$, then there exist a family of two-sided fee-shifting rules that would create the optimal incentives for the plaintiff to bring suit. Each such rule would specify that the defendant pays the plaintiff's expenses if $x_c<Y$ and that the plaintiff pays the defendant's expenses if $x_c \geq Z$, where:

$$-(c_p+c_d) + F(Y\text{-}X)c_p + F(Z\text{-}X)c_d + F(0)D = 0. \tag{27}$$

Proof: Under the proposed rule, the plaintiff would sue if and only if:

$$-(c_p+c_d) + Pr(x_c<Y|x_p)c_p + Pr(x_c<Z|x_p)c_d + Pr(x_c<X|x_p)D > 0.$$

Thus, the plaintiff will sue if and only if $x_p<y^*$, where y^* is defined by:

$$-(c_p+c_d) + F(Y\text{-}y^*)c_p + F(Z\text{-}y^*)c_d + F(X\text{-}y^*)D = 0. \tag{28}$$

Thus, to ensure that $y^*=X$, we must set Y and Z such that (27) holds. Note that the left-hand side of (27) is monotonically increasing in Y as long as $0 < F(Y\text{-}X) < 1$ and in Z as long as $0 < F(Z\text{-}X) < 1$.

If we let both Y and Z approach ∞, then both F(Y-X) and F(Z-X) go to 1, so that the left-hand side of (27) must become positive, and plaintiffs would sue too often. If we let both Y and Z approach $-\infty$, then both F(Y-X) and F(Z-X) go to 0. In this case, the left-hand side of (27) must become negative, so that plaintiffs would not sue often enough, if and only if (17) holds. Thus, there exists a pair (Y, Z) that solves (27) if (17) holds. If (17) does not hold, then there would always be too much litigation.

If such a solution exists, then there exists a whole family of solutions (Y, Z). We can solve (27) for either Y or Z. The solution for Y may be expressed as a monotonically decreasing function of Z:

$$Y(Z) = X + F^{-1}\{[c_p + c_d - F(0)D - F(Z\text{-}X)c_d]/c_p\}. \tag{29}$$

Similarly, the solution for Z may be expressed as a monotonically decreasing function of Y:

$$Z(Y) = X + F^{-1}\{[c_p + c_d - F(0)D - F(Y-X)c_p]/c_d\}. \qquad (30)$$

Thus, increases in either Y or Z would substitute for increases in the other. ■

1. Comparative Statics

Note that the plaintiff's expected payoff from bringing the marginal suit, the left-hand side of (27), is monotonically increasing in F(0) and D. Furthermore, as long as both $0 < F(Y-X) < 1$ and $0 < F(Z-X) < 1$, this payoff will strictly increase in Y and Z, but strictly decrease in c_p and c_d.[4] Thus, if the fee-shifting rule uses both thesholds, Y and Z, to optimize the plaintiff's incentives to bring suit, then each threshold -- holding the other constant -- will decrease in F(0) and D but increase in c_p and c_d.

That is, ceteris paribus, the optimal policy becomes more pro-defendant if either F(0) or D increases, but more pro-plaintiff if either c_p or c_d increases. If the award that the plaintiff expects to recover increases, then the optimal rule -- to offset the increased incentive to bring suit -- must increase the probability that the plaintiff would have to bear the parties' litigation costs. If on the other hand, these litigation costs increase, then the optimal rule -- to offset the reduced incentive to bring suit -- must decrease the probability that the plaintiff would have to bear them.

2. Fee-Shifting Subject to Constraints on Y and Z

Each of the classic fee-shifting rules is a special restricted case of our two-sided fee-shifting rule, with (Y, Z) constrained to take on a particular pair of values: $(-\infty, \infty)$ for the American rule; (X, X) for the British rule; $(-\infty, X)$ for the pro-

---

[4]As long as the court threatens to impose both c_p and c_d on the plaintiff in the marginal suit, its payoff will decrease in both parameters. In particular, if there is some positive probability of $x_c \geq Y$ when $x_p = X$, then this payoff will be strictly decreasing in c_p. Similarly, if there is some positive probability of $x_c \geq Z$ when $x_p = X$, then this payoff will be strictly decreasing in c_d also.

defendant rule; and (X, ∞) for the pro-plaintiff rule. As we have seen, these constrained rules provide optimal incentives only in very special circumstances. Similarly, constraints such as $Y \leq X$ or $X \leq Z$ would also restrict the circumstances under which the two-sided fee-shifting rule could provide optimal incentives. These restrictions, however, prove to be less severe than those imposed by the classic fee-shifting rules.

The constraint $Y \leq X$, for example, would imply that the fee-shifting rule can be no more pro-plaintiff than the classic pro-plaintiff rule. Recall that even the classic pro-plaintiff rule leads to too little litigation if and only if condition (22) does not hold. Thus, fee-shifting subject to the constraint $Y \leq X$ can provide optimal incentives if and only if (22) holds. We can express condition (22) as follows:

$$c_p[1 - F(0)] \leq F(0)D. \qquad (31)$$

Condition (31) states that in the marginal case, the plaintiff's expected gain from adjudication must at least equal its expected liability for its litigation costs under the pro-plaintiff rule.

If the plaintiff's litigation costs are too great relative to the damages at stake, then (31) fails to hold. In such cases, one would have to supplement such restricted fee-shifting with other policies in order to induce plaintiffs to bring suit in all appropriate cases. For example, in a particular category of cases, one could provide for subsidies or less costly procedures to reduce the plaintiff's c_p, or provide for treble damages to increase the D recovered, or reduce the plaintiff's burden of proof to increase F(0), so that (31) can hold.

Similarly, the constraint $X \leq Z$ would imply that the fee-shifting rule can be no more pro-defendant than the classic pro-defendant rule. By analogous reasoning, fee-shifting subject to this constraint can provide optimal incentives if and only if (18) holds. We can express condition (18) as follows:

$$F(0)D \leq c_p + c_d[1 - F(0)]. \qquad (32)$$

Condition (32) states that in the marginal case, the plaintiff's expected gain from adjudication cannot exceed its expected liability for litigation costs under the pro-defendant rule.

If litigation costs are too small relative to the damages at stake, then (32) fails to hold. In such cases, one would have to supplement c_d under such restricted fee-shifting with some other sanctions for plaintiffs. Similarly, one could raise the plaintiff's burden of proof in a particular category of cases, in order to reduce F(0), and thereby induce plaintiffs to bring suit only in appropriate cases.

3. Example with a Uniform Distribution

Suppose again that e is uniformly distributed in the interval (-E, E). If (17) holds, then we can use the F(e) in (23) to derive the family of optimal rules for this example. After rearranging terms, we find that Proposition 3 and (23) together imply:

$$Yc_p + Zc_d = (c_p+c_d-D)E + (c_p+c_d)X, \tag{33}$$

where Y and Z both lie in the interval (X-E, X+E). Solving (33) for Y yields:

$$Y = X + [(c_p+c_d-D)E - (Z-X)c_d]/c_p, \tag{34}$$

and solving (33) for Z yields:

$$Z = X + [(c_p+c_d-D)E - (Y-X)c_p]/c_d. \tag{35}$$

Note that the effect of an increase in the uncertainty of the trial outcome has an ambiguous effect on the (Y, Z) locus: an increase in E may move the locus to higher values of Y and Z or to lower values, depending on whether c_p+c_d is greater or less than D.

F. EXTENSIONS

In this section, we present some extensions of the preceding model. First, we relax the assumption that the plaintiff knows x with certainty. Second, we relax the assumption that the social objective is for plaintiffs to sue if and only if they believe the defendant is liable. Third, we relax the assumption that the plaintiff knows D with certainty.

1. Plaintiff Uncertainty Regarding x

Until now we have assumed that $x_p=x$. We can assume instead that not only the court but also the plaintiff observes x with error. In particular, let $x_p = x+e_p$ and $x_c = x+e_c$. In this case, $x_c = x_p+e$, as before, where now $e = e_c-e_p$. The model yields the same results as before, with the random variable e now reinterpreted to represent the difference between the court's error and the plaintiff's error.

2. Other Social Objectives

The preceding analysis took the policy goal to be to induce the plaintiff to sue if and only if $x_p<x^*$ where $x^*=X$. We can, however, allow for x^* other than X. For example, courts may identify a class of suits that society would not want to discourage even though their claims may be unlikely to succeed. There may be some public good flowing from suits that fall short of success along some particular dimension x.[5] Thus, courts may set $x^*>X$ to encourage such litigation.

This extension requires only minor modifications in the preceding analysis. Again, we would set Y (or Y and Z) so that the plaintiff is just indifferent about bringing the marginal suit, but now we regard the case in which $x_p=x^*$ (rather than $x_p=X$) to be the marginal suit. Thus, x^* would replace X where appropriate. For example, one would substitute x^* for X in (14), (19), (25), (26), (29), (30), (34), and (35). Note also that $F(X-x^*)$ replaces $F(0)$, $F(Y-x^*)$ replaces $F(Y-X)$, and $F(Z-x^*)$ replaces $F(Z-X)$ in the preceding discussion. A higher x^* implies a rule more likely to put the burden of litigation costs on the defendant, whereas a lower x^* implies a less pro-plaintiff rule, with a lower Y (or a set of lower Y and Z values).

3. Plaintiff Uncertainty Regarding Damages

For simplicity, we have assumed that the plaintiff knows the damages that a court will award, D, with certainty. In reality, plaintiffs face uncertainty over D

---

[5]Controversy over Rule 11 focuses on the risk that sanctions will discourage the use of innovative legal theories. See, e.g., Nelken (1986) and Note (1987). Courts can respond to these concerns by allowing an attorney greater leeway in making a losing argument when it is a novel "argument for the extension, modification, or reversal of existing law," in the language of Rule 11. Wilder (1986) argues that courts have in fact been sensitive to the chilling effect of sanctions and have used Rule 11 cautiously.

as well as over x_c. The model can be extended to the case of uncertainty if we let D represent the expected damages conditional on a finding of liability. If this D is independent of x, then much of our analysis will be unchanged.

The damages awarded, however, may be correlated with the court's findings on x. Suppose for example that a court is more likely to award punitive damages if it finds that the defendant was grossly negligent.[6] In this case, the expected damages D anticipated by the plaintiff will depend on the plaintiff's observed x_p, and some of the preceding equations would have to be modified. As long as the court knows the function $D(x_p)$, however, and the plaintiff's expected payoff from litigation still declines in x_p, as is likely, the court can set fee-shifting thresholds much as before. The optimal thresholds can ensure that the plaintiff is just indifferent about bringing the marginal suit, but willing to bring better suits and unwilling to bring worse suits. Thus, such an extension would still preserve the general thrust of our results.

G. CONCLUDING REMARKS

The preceding analysis has shown how fee-shifting based on the margin of victory gives courts an additional instrument that permits them to fine-tune the plaintiffs incentives to bring suit. If the courts are instead restricted to the classic fee-shifting rules, which simply turn on the identity of the winning party, they cannot insure that plaintiffs have the optimal incentives. Under each of these rules, plaintiffs will bring some suits lacking in merit and fail to bring some meritorious suits.

The authors of Rule 11 intended their rule to deter frivolous suits, and

[6]One might provide for such additional damages in some categories of cases to induce plaintiffs to bring good suits that they would otherwise not bring because of high litigation costs. If courts make damages a function of x, rather than a fixed D as we have assumed, then they have another policy instrument that can supplement a two-sided fee-shifting rule. This instrument can be particularly useful when that rule cannot ensure optimal incentives for the plaintiff, as in the case discussed in Section E.2 when $Y \leq X$ and condition (31) fails to hold.

accordingly one can interpret Rule 11 as an example of the type of fee-shifting rule analyzed in this essay. Rule 11 is a two-sided fee-shifting rule that reserves all sanctions for parties that lose by wide margins. In terms of the model presented in this essay, Rule 11 restricts the thresholds, Y and Z, so that $Y < X < Z$. Given this constraint, Rule 11 fee-shifting could not ensure that plaintiffs have optimal incentives, even if litigation costs are large enough relative to the damages at stake to ensure condition (17) holds. Fee-shifting under Rule 11 can provide optimal incentives if and only if the inequalities in (31) and (32) both hold strictly, that is, if and only if:

$$c_p[1 - F(0)] < F(0)D < c_p + c_d[1 - F(0)]. \qquad (36)$$

That is, in the marginal case, the plaintiff's expected gain from adjudication must fall between its expected liability for litigation cost under the pro-plaintiff rule and its expected liability under the pro-defendant rule. Nevertheless, Rule 11 is an improvement over either the American rule or the British rule, which subject fee-shifting to still more severe restrictions.

Rule 11, however, also allows for sanctions other than the award of attorneys' fees, and these other sanctions may supplement the use of fee-shifting. Courts may use these sanctions when litigation costs are too small relative to the damages at stake for the second inequality in (36) to hold. In these cases, the prospect of Rule 11 fee-shifting would be insufficient to deter all frivolous lawsuits, and additional penalties would be appropriate. Rule 11 provides courts with no instrument, however, to handle cases in which the plaintiff's litigation costs are so great relative to the damages at stake that too little litigation results even under the pro-plaintiff rule -- that is, those cases in which the first inequality in (36) fails to hold.

Most important, this essay suggests that courts can adjust the Rule 11's deterrent effect not only by varying the magnitude of the sanctions, but also by varying the threshold that will trigger the sanctions. The comparative statics results in Section E.1 offer some guidance for courts exercising the discretion that they enjoy under Rule 11. Courts should give Rule 11 an interpretation generous to plaintiffs in cases in which the litigation costs are large relative to the damages at stake. As Section E indicates, this interpretation may mean that courts use Rule 11 more sparingly against plaintiffs, or that they use it more aggressively against defendants, or both. Conversely, if the damages at stake are large relative to the litigation costs, then courts should invoke Rule 11 more readily against plaintiffs and less frequently against defendants. To the extent that other social objectives militate in favor of plaintiffs bringing some particular types of lawsuits, however, Section F.2 suggests a more pro-plaintiff policy.

As the examples in Sections D.3 and E.3 indicate, an increase in the plaintiff's uncertainty over the trial outcome has an ambiguous effect on the optimal fee-shifting rule: it can militate in favor of either a more pro-plaintiff or a more pro-defendant rule. We have assumed, however, that the plaintiff is risk neutral. Given a risk-averse plaintiff, an increase in risk would magnify the deterrent effect of any given fee-shifting rule. This effect suggests that courts should give Rule 11 a more pro-plaintiff interpretation in the face of greater legal uncertainty or when dealing with particularly risk-averse plaintiffs.

We have assumed also that the plaintiff sues only if litigating to judgment is worthwhile. In reality, parties usually settle out of court before going to trial, and plaintiffs may bring suits solely to extract a settlement offer. See Bebchuk (1988, 1991); Rosenberg and Shavell (1985). Fee-shifting rules affect the parties' expected payoffs from adjudication, and thereby influence how the parties divide the surplus gained by avoiding the litigation costs imposed by a failure to settle insofar as that failure leads to trial and judgment. Because fee-shifting rules can set the "threat point" of the settlement bargaining and thereby influence the settlement amount, they can still improve the plaintiff's incentives to bring suit by tailoring that threat point to the circumstances of the case.

A comprehensive normative analysis would also consider the effect on the defendant's incentives ex ante and the total administrative costs of the legal system. Fee-shifting rules affect not only the plaintiff's incentive to bring a negligence suit, for example, but also the cost to the defendant of such a lawsuit. Through both these effects, such rules affect a potential defendant's incentive to take care. Ultimately, one would determine the optimal incentives for plaintiffs to bring suit by weighing the costs of suboptimal primary behavior by defendants against the social costs of lawsuits. Thus, the optimal fee-shifting rule would depend upon more than just the rule's effects on the plaintiff's incentives to bring suit. Further research is necessary to address the issues raised by these more complex social objectives, which fall outside the scope of our preliminary analysis in this essay.

REFERENCES

Bebchuk, Lucian A. (1991), "The Credibility and Success of Suits Known to be Made Solely to Extract a Settlement Offer," Harvard Law School, mimeo.

Bebchuk, Lucian A. (1988), "Suing Solely to Extract a Settlement Offer," Journal of Legal Studies, Vol. 17, 437-450.

Burbank, Stephen B. (1989), "The Transformation of American Civil Procedure: The Example of Rule 11," University of Pennsylvania Law Review, Vol. 137, 1925-1967.

Levinson, Sanford (1986), "Frivolous Cases: Do Lawyers Really Know Anything at All?" Osgoode Hall Law Journal, Vol. 24, 353-378.

Nelken, Melissa L. (1986), "Sanctions Under Amended Federal Rule 11 -- Some 'Chilling' Problems in the Struggle Between Compensation and Punishment," Georgetown Law Journal, Vol. 74, 1313-1369.

Note (1987), "Plausible Pleadings: Developing Standards for Rule 11 Sanctions," Harvard Law Review, Vol. 100, 630-652.

Polinsky, A. Mitchell and Steven Shavell (1989), "Legal Error, Litigation, and the Incentive to Obey the Law," Journal of Law, Economics, and Organization, Vol. 5, 99-108.

Rosenberg, David and Steven Shavell (1985), "A Model in Which Suits Are Brought for Their Nuisance Value," International Review of Law and Economics, Vol. 5, 3-13.

Schwarzer, William W. (1988), "Rule 11 Revisited," Harvard Law Review, Vol. 101, 1013-1025.

Schwarzer, William W. (1985), "Sanctions Under the New Federal Rule 11 -- a Closer Look," Federal Rules Decisions, Vol. 104, 181-206.

Shavell, Steven (1982), "Suit, Settlement, and Trial: A Theoretical Analysis Under Alternative Methods for the Allocation of Legal Costs," Journal of Legal Studies, Vol. 11, 55-81.

Wilder, Nancy H. (1986), "The 1983 Amendments to Rule 11: Answering the Critics' Concern With Judicial Self-Restraint," Notre Dame Law Review, Vo. 61, 798-818.