JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 1 WINTER 2005 NUMBER 2

CONTENTS

BOOK REVIEWS

THE SPONTANEOUS EVOLUTION OF CYBER LAW: NORMS, PROPERTY RIGHTS, CONTRACTING, DISPUTE RESOLUTION AND ENFORCEMENT WITHOUT THE STATE

Bruce L. Benson, Ph.D.[*]

[*] DeVoe Moore Distinguished Research Professor, Department of Economics, Florida State University. E-mail: bbenson@garnet.acns.fsu.edu. This paper was prepared at the invitation of Peter Boettke, as a member of the Working Group on the Law, Economics and Technology of Private Contract Enforcement on the Internet, Critical Infrastructure Project, and for publication in the Journal of Law, Economics and Policy, Vol. 1:2. I want to thank the participants in seminars presented at George Mason University and San Jose State University for helpful comments and suggestions. Discussion with and comments from Patrick Peterson were particularly helpful.

I. INTRODUCTION

Some observers suggest that the appropriate analogy for thinking about order in cyberspace, or the lack thereof, is the "wild west."[1] The fact is that the western frontier was not nearly as wild and lawless as most people believe, however (Anderson and Hill 1979, 2004; Benson 1991b), and the same is true of cyberspace. As Johnson and Post (1996, p. 1389) note, "Cyberspace is anything but anarchic; its distinct rule sets are becoming more robust every day." Indeed, while an accurate picture of the "not so wild, wild west" (Anderson and Hill's recent book title) might be analogous to cyberspace, a better analogy is offered by international commercial law, *lex mercatoria* (Johnson and Post 1996, pp. 1389-90; Benson 2000c), the polycentric merchant-produced law of medieval Europe and its modern counterpart. Cyberspace is unconstrained by geographic boundaries, so like international trade, it cannot be effectively governed by geographically defined legal systems. Yet methods of creating property rights and solving

[1] There are critics of this "wild west" analogy. Some accept that the west was a wild frontier but contend that cyberspace is not (e.g., Johnson and Post [1996]). Others prefer a different analogy, seeing cyberspace as a feudal society (e.g., see discussion in Yen [2002]). As noted below, the contention here also is that there is a better analogy, but it is not feudalism.

assurance problems are evolving within functionally defined cyber communities, just as they have within various international trading communities.

In order to explain how effective cyber law is evolving without the backing of coercive power, the following presentation is divided into three sections beyond this introduction. Section II provides a general explanation for the spontaneous and voluntary "bottom-up" evolutionary process that creates rules of obligation and institutions to encourage recognition of those rules, resolve disputes that arise under the rules, and change the rules as conditions change. These rules of obligations and supporting institutions create property rights in order to resolve externality problems and establish trust or recourse mechanisms in order to alleviate assurance problems. Section III maps observed developments of rules of obligation and supporting institutions in cyberspace into the analysis provided in Section II, illustrating that property rights are evolving to deal with the externalities that have developed on the Internet, and that both trust and recourse are increasingly available as solutions to assurance problems. The concluding Section then explains why geographically defined states, and even international organizations of such states, are unable to bring order to cyberspace, and why they should be discouraged from even trying to do so.

II. AN ECONOMIC THEORY OF SPONTANEOUSLY EVOLVING INSTITUTIONS: PROPERTY RIGHTS, CONTRACTS, AND CUSTOMARY LAW

Rules can be thought of as behavioral patterns that individuals expect each other to follow. The rules one individual is expected to follow influence the choices made by other individuals: like prices, rules coordinate and motivate interdependent behavior. A subset of rules generally do not require explicit codification or backing by coercive threats to induce recognition, because they are widely "shared values" (Voigt and Kiwit 1998) voluntarily adopted by individuals in their interactions with other members of an identifiable (but perhaps changing) group of individuals. As such interactions evolve and change, these "norms" or "customs" also spontaneously evolve (Benson 1999a). There are obviously many other rules beyond such norms that people are expected to follow, however. Some rules are not shared values, for instance, but instead, they discriminate in favor of targeted individuals and are imposed on others by employing coercive threats. Furthermore, many rules and accompanying governance institutions are established through deliberate design rather than evolving spontaneously.[2]

---

[2] The historical importance of deliberately designed rules as actual determinants of behavior is probably much less than is popularly perceived, however, since people rely on norms to govern much of their behavior even when some formal rules of law may appear to apply (de Soto 1989; Acheson 1988; Ellickson 1991; Benson 1989; Bernstein 1992). There are too many uncontrolled margins and unanticipated responses for a rule designer to consider. Nonetheless, such designed rules do influence behavior.

The final section of this presentation considers the role and impact of deliberately designed rules imposed from the top down, but until then, the focus will be on the bottom up evolution of norms underlying property rights and contracting arrangements.

II.1. *The Evolution of Property Rights*

The institutions of property—the network of primary and secondary rules regarding obligations to respect entitlements regarding access to, use of, and transfer procedures for assets[3]—determine how resources will be brought into use. Resources will be used, however, no matter what the property rights system is. Even the lack of specified property rights is a property rights system: it implies that a resource can be used by anyone strong enough to claim access. Indeed, since property rights are an important determinant of the allocation of resources, and therefore at least to a degree, the distribution of wealth, individuals have incentives to develop property rights arrangements to help them achieve their personal objectives. Creation of effective property rights requires persuading or inducing others to recognize rules (accept obligations) to respect the rights, of course. After all, rules are generally not necessary if there are no conflicts to resolve. Furthermore, as David Hume (1751) emphasized over 250 years ago, the primary source of conflict between individuals is scarcity. Scarcity becomes apparent when a conflict over use arises, and the institutions of property evolve as such conflicts are resolved by individuals attempting to find ways to expand personal well-being or "wealth" in the face of this scarcity (Hume 1957; Commons 1924, p. 138; Benson 1994b, 1999a),[4] regardless of whether the resolution is achieved through cooperative procedures such as negotiations, or through violence or threats of violence (coer-

---

In particular, deliberate efforts to impose rules create incentives to find and exploit uncontrolled margins in order to avoid the full consequences of those rules (e.g., Cheung [1974]; Barzel [1989]; Kirzner [1985]; Benson [2002]), and in this context, the search for ways to avoid the rules also can significantly alter the path of the spontaneous evolution of behavior.

[3] Primary and secondary rules are defined as in Hart (1961). That is, primary rules define the obligations that individuals are expected to have, and secondary rules establish the processes that induce recognition, provide adjudication, and facilitate change in primary rules. Some primary rules may not be supported by secondary rules, of course (e.g., some moral norms, conventions, customs, etc.), but such rules may be part of the institution of property anyway.

[4] Note in this regard that wealth does not just mean monetary wealth or even physical possessions; it can include many other sources of satisfaction, such as health, security, loyalty, friendship, family, prestige, and power. Indeed, the relative values that individuals place on material and nonmaterial aspects of wealth are at least partially endogenous (Benson 1999a), since preferences continually change as people undergo the experiences of life (Vaughn 1994, p. 80). In a very hostile environment, for instance, individuals may willingly sacrifice a good deal of potential material wealth in order to obtain more safety or security.

cion). Thus, in a positive sense, property rights are a matter of economic value rather than of legal definition or moral philosophy.

II.1.a. Externalities and the evolution of property rights

Coase (1960) stresses the reciprocal nature of externalities, explaining that they arise when two or more individuals attempt to use the same asset or resource for conflicting purposes.[5] The primary reasons for divergent expectations regarding access to and use of the resource are that the relevant property rights either are not clearly assigned or they are not effectively protected (i.e., rules of obligation either are not recognized or not respected). Therefore, a solution to the conflict may be achieved if one party is persuaded/induced to recognize that the other party has the relevant property rights and/or to respect those rights (follow rules of obligation). To avoid harm to one party, the other party must be harmed. The result of allocation and recognition of property rights is that one party will be able to benefit from using the resource and the other party must bear costs if he or she wants to use it (i.e., bargain to purchase it, or face the possibility of prosecution and liability for damages due to trespass or transfer without bargaining [theft]). Because the allocation and enforcement of property rights determines the distribution of costs and benefits, the effected individuals obviously will have different opinions about how the rights should be assigned and protected. Thus, transactions costs arise as individuals attempt to create or protect property rights, because other individuals must either be persuaded or induced to accept obligations to respect claimed rights. Effective property rights are likely to arise when the benefits of creating them exceed the associated transactions costs. In fact, as Demsetz (1967, p. 350) explains, "Property rights develop to internalize externalities when the gains from internalization become larger than the cost of internalization. Increased internalization, in the main, results from changes in economic values, changes which stem from . . . changes in technology and relative prices."

Suppose that a sufficiently large number of people want to use a resource so that the resource is scarce. That is, the use by some individual or

---

[5] Before Coase (1960), the typical economic analysis of externalities focused on a divergence between the private and social costs of the action: an action by one individual creates a cost born by (or benefit captured by) another individual, so the decision maker presumably does not consider that cost (or benefit) in the decision. In other words, the full social costs (or benefits) differ from the private costs (or benefits) internalized by the decision maker. They are external to the decision. In the case of external costs, for instance, the pre-Coasian analysis started with a charge that A inflicts harm on B, so the policy question was, how should A be restrained? The conclusion generally was that the appropriate policy was taxation to mirror the external costs so the decision maker would act as if the costs were internal. The alternative typically was the imposition of regulatory requirements that imposed constraints on the decision maker's choices. But Coase points out that this traditional approach really obscures the nature of the choice that has to be made.

individuals, A, has a negative impact on the well being of (imposes a negative externality on) another individual (or individuals), B. Under these circumstances, each individual has incentives to access the resource and use it before other users do the same. Assume that B's use is prevented altogether if A is successful in using the resource first, for instance, and visa versa. In this case, each individual has incentives to claim the resource before others do.[6] The potential means of successfully asserting that claim (i.e., creating obligations for others in the relevant community to recognize the claim) will be discussed below. The point here simply is that when A's use of a scarce resource precludes B's use and visa versa, the conflict becomes immediately apparent, as are the incentives to claim property rights. Things are not quite this simple, however.

It is costly to measure an asset's attributes, and the cost of increasingly fine delineation of an asset's attributes rise: "Because transacting is costly, as an economic matter property rights are never fully delineated" (Barzel 1989, p. 1; also see Libecap 1986, pp. 230-231, and North 1990, p. 33). Individuals have incentives to delineate and develop rights to those attributes of an asset that are valuable relative to the cost of delineation and rights establishment.[7] However, when costs mean that attributes are less than fully delineated and claimed, some attributes remain accessible to individuals, e.g., B, other than the person or persons, A, who has claimed rights to the delineated attributes. Whether intentionally or not, uses of unclaimed attributes of the resource by B can negatively affect the well-being of A. Thus, externalities persist. As Barzel (1989, p. 5) explains, for instance,

> The rights to receive the income flow generated by an asset are a part of the property rights over that asset. The greater is others' inclination to affect the income flow from someone's asset without bearing the full costs of their actions, the lower the value of the asset. The maximization of the net value of an asset, then, involves that ownership or ownership pattern that can most effectively constrain uncompensated exploitation.

The argument is more general than just income, however. Interference with the ability to obtain any type of subjective benefit will create incentives to look for ways to constrain such interference, and if cost effective means of doing so are discovered, more attributes will be delineated and property rights will evolve. The bundle of rights associated with a particular resource can be quite large, then, as access, use, and transferability rights can apply to and differ over many different attributes.

In some cases, many people can use a resource without actually precluding others from using the same resource, but in the process each per-

---

[6] See Anderson and Hill's (1990) examination of the rush for land under the Homestead Act, for example.

[7] See note 10 below for an example.

son's use, or benefits from use, may be impaired in some way. That is, each person's use reduces the satisfaction that others obtain from their use. In the case of a free-access commons, for instance, crowding or congestion arises and the resource deteriorates in quality (each individual's use impairs every other individuals' uses to a degree) as the result of over use. This has been called the tragedy of the commons, of course.[8] The tragedy of the commons is not the inevitable outcome of free access, however, because the externalities create incentives to develop alternative property rights. In fact, it is not the likely outcome unless the transactions costs of at least partial privatization (e.g., quota rights like fishing licenses or pollution permits) are so high that such privatization is not warranted. In this context, Johnson and Libecap (1982) contend that there are three types of costs that can be relevant: (1) exclusion costs; (2) internal governance costs when exclusive rights are shared by a group; and (3) imposition of punishment for violating an open access constraint imposed by a strong coercive authority (e.g., the state).[9] If these costs are not prohibitive, property rights should evolve when externalities become significant. For example, Demsetz (1967) uses his theory to explain the evolution of property rights among a community of Indian hunters in Eastern Canada who, in the early 18th century, developed exclusive rights to take beaver furs. Prior to the arrival of European fur traders, beaver were so abundant relative to the native population's demand for beaver furs or meat, that if one individual harvested beaver from the commons, it had no noticeable impact on other individuals' ability to harvest beaver. There was no reason to claim exclusive hunting rights. However, Europeans considered beaver to be very valuable. Therefore, when European fur traders began trading things that the Indians valued highly (European manufactured goods such as metal knives, axes, pots and pans,

---

[8] This terminology was coined by Hardin (1968). In this situation, each user has an incentive to use the resource because she is not fully liable for the cost of doing so. Part of the cost is born by others, so the crowding and over use are negative externalities: all of those with access try to use up the resource before someone else does, and the commons deteriorates, perhaps even being destroyed. Crowding is not the only consequence of common access, however. The resource itself can be used up, but even if it is not completely destroyed, it is used inefficiently so that the quality of the output from its use diminishes over time. This could be offset with appropriate investments in maintenance or improvements, but the individuals with common access to the resource do not have incentives to make such investments, because they cannot exclude others from capturing (e.g., charge for) the resulting benefits. In essence, maintenance of common access property generates external benefits so underinvestment results. Classic examples of commons are ocean fisheries (Gordon 1954; Johnson and Libecap 1982) and wild game such as buffalo (Benson 2004), but many publicly provided goods and services, such as highways (Benson 1994a), courts (Neely 1982), and police services (Benson 1994a) can also be characterized as common pools.

[9] An imposed open access constraint need not be binding if its exclusion costs and internal governance costs are low, however. See Umbeck (1977, 1981a, 1981b) on the development of property rights to mining claims on federal land, Acheson (1988) on property rights established in the Maine lobster fishery by "harbor gangs," and de Soto (1989) on property rights in urban land created by the "informal-sector" squatter communities of Peru.

woolen blankets, guns and ammunition, etc.) for beaver, the Indians' incentives to hunt beaver increased dramatically. This created a common pool problem. Rather than allowing the destruction of the beaver population, Indians began claiming and recognizing exclusive hunting rights. Individual Indian households claimed their own beaver populations along streams, and the community of Indians (related bands) recognized those claims. As a result, the population of beaver leveled off and was maintained, rather than being destroyed. Demsetz also argued that the Indians of the American Southwest failed to develop similar property rights because of the relatively high costs and low benefits from establishing exclusive hunting rights.[10]

II.1.b. Factors influencing the evolution of property rights

Demsetz' (1967) property-rights-in-beaver example emphasizes that a change in relative values (e.g., prices) can lead to a change in property rights (also see, for example, Libecap [1978]). Rights can also evolve because of a technological change that reduces the cost of measuring attributes and/or establishing rights. For instance, Anderson and Hill (1975) explain that property rights in range land in the American West evolved from open range to private range with the introduction of barbed wire, and Ellickson (1993, p. 1330) explains that transferability of private use rights developed, at least in part, because written language reduced the cost of record keeping. Property rights can also evolve with changes in the institu-

---

[10] McManus (1972) looked more carefully at the situation in Canada than Demsetz had. He found that the beaver populations in the area were sharply reduced after the introduction of fur trade, as predicted by the common pool problem that arose, but as Demsetz suggested, the population of beavers ultimately stabilized, with the creation of a property rights arrangement. McManus also examined in more detail the property rights structure the Indians developed, ultimately agreeing with Demsetz that it was efficient, although different than Demsetz' characterization. He noted that the Indians were organized into small bands and that individual members of the band had a recognized right to exclude others from taking furs or meat from their territories for sale, but they did not claim the right to exclude others from killing animals for personal consumption of meat. The fur was expected to be left for its owner, however. In other words, rights to use for direct consumption of meat were still commonly held while rights to use for exchange were exclusive. The common right to meat consumption is explained from a transactions cost perspective: since hunters lived in an uncertain world and faced a real threat of starvation, the common right to kill for one's own consumption was an institutionalized form of mutual insurance. McManus referred to it as the Good Samaritan Constraint on the exercise of exclusive rights. But like modern welfare or social insurance schemes, McManus suggest that this also resulted in "irresponsibility and laziness, and the depletion of beaver." The property rights structure tends to evolve to maximize wealth, however, as he contends that the Good Samaritan constraint reduced the cost of enforcing exclusive rights for use in exchange by reducing the incentives to steal or invade neighboring territories. If the insurance and enforcement benefits of the constraint were larger than the costs and if less expensive forms of insurance were unavailable, then the result would be wealth maximizing (see [Johnsen 1986] for another example of sharing norms as mutual insurance).

tions of policing and dispute resolution (Benson 1999a). Similarly, discoveries of new information that lead individuals to recognize that the resource in question is more valuable than previously believed, can also induce increased measurement and rights delineation (e.g., discovery of gold, as in Umbeck [1977, 1981a, 1981b]; also see Libecap [1978]). In addition, property rights can change because of changes in the relative opportunity costs of violence used to induce recognition of rights, or the relative costs of negotiation used to persuade potentially cooperative individuals to recognize each others' rights (e.g., see Benson [1994b, 1999a, 2004]; Anderson and McChesney [1994], perhaps due to technological changes that reduce the costs of violence or negotiation. Thus, many things can change the value of property rights (Demsetz 1967; Anderson and Hill 1975; Libecap 1986, p. 231; Barzel 1989, p. 74; North 1990, p. 48), and this in turn leads to changes in the rights structure. As Barzel (1989, p. 65) emphasizes in echoing Demsetz (1967), "People acquire, maintain, and relinquish rights as a matter of choice. . . . As conditions change, . . . something that has been considered not worthwhile to own may be newly perceived as worthwhile." Such changes in conditions need not be exogenous, however. In fact, they are not likely to be. When substantial externalities arise, the desire to internalize them creates incentives to carry out research in an effort to develop new technologies and/or new information that will lower the transaction costs of rights creation. Thus, the technological advances and other increases in knowledge that allow property rights to evolve often are endogenous.

II.1.c. The security of property rights

Alchain and Allen (1969, p. 158) propose a revealing definition of ownership rights: "the expectations a person has that his decision about the use of certain resources will be effective." This very non-legal and non-philosophical sounding definition stresses the fact that, in practice, rights (or expectations) are never absolute, and not just because of Barzel's (1989, p. 5) point that delineation and development of rights to attributes is costly. The strength or security of property rights (and therefore expectations) is a function of efforts made to protect or enforce rights claims, and the offsetting efforts to take or attenuate those claims. There are many ways that others might attempt to capture control of resources, such as through warfare or theft, or through the political process by lobbying the legislature to change the rules. Given such threats, people incur costs as they attempt to enhance the security of the rights over the assets that they wish to control. Sherman (1983) discusses individual investments in watching, walling, and wariness, for instance. These activities can include investments of time (e.g., watching, lobbying to prevent political transfers), or in the purchase of labor services (e.g., guards, hired lobbyists), technology (e.g., monitoring equipment for watching, better fences and locks as part of walling), and in

the sacrifice of desirable activities that might increase the vulnerability of property (e.g., avoiding certain dangerous areas, not going out at night). Technological changes can lower such costs by improving the productivity of labor and capital. Thus, incentives to perform research and development of security methods are high when property rights are insecure.

Some of the costs of watching, walling, and wariness activities for individuals can also be reduced significantly through organizational innovations. A community of individuals might be formed, for instance, and voluntarily agree to respect each others' rights. In this context, Vanberg and Buchanan (1990, p. 18) define "trust rules" to be rules of behavior toward others which individuals have positive incentives to voluntarily recognize, and explain that:

> By his compliance or non-compliance with trust rules, a person selectively affects specific other persons. Because compliance and non-compliance with trust rules are thus "targeted," the possibility exists of forming cooperative clusters. . . . Even in an otherwise totally dishonest world, any two individuals who start to deal with each other - by keeping promises, respecting property, and so on - would fare better than their fellows because of the gains from cooperation that they would be able to realize.

Wealth can be enhanced for everyone involved in such trust relationships as property rights are made relatively more secure and relatively more private. As Hayek (1973, p. 107) explains, "The understanding that 'good fences make good neighbours,' that is, that men can use their own knowledge in the pursuit of their own ends without colliding with each other only if clear boundaries can be drawn between their respective domains of free action, is the basis on which all known civilization has grown. Property, in the wide sense in which it is used to include not only material things . . . is the only solution men have yet discovered to the problem of reconciling individual freedom with the absence of conflict." Indeed, the absence of conflict may be the primary objective of some agreements. Such implicit agreements as seen between animals as property claims in the form of hunting ranges are delineated and recognized, thereby reducing conflict (Hayek 1973, p. 75). In other words, the primary goal of some agreements may be to obtain non-material wealth in the form of "peace" or security. The production of more material wealth is also likely to be enhanced in such trust relationships. If property rights are made relatively more secure and relatively more private, as time horizons lengthen, incentives to use the property for production, rather than immediate consumption, increase. Thus, whatever the objective, all organizations function to a substantial degree by delineating various rights that individuals are expected to respect (Barzel 1989, p. 7).

Members of such communities may also agree to cooperate in joint production of property protection against outside threats (e.g., neighborhood watch, pooling funds to employ security or policing services for an area, walling and gating an entire community, traveling in a group). But a significant source of transaction cost may stand in the way of adopting such

rules voluntarily through an explicit or implicit contract: the assurance problem. Will individuals live up to their promises to cooperate, or will they attempt to free ride on the efforts of others in the group? Therefore, consideration of the nature of the cooperative process that can underlay a system of voluntarily recognized and protected property rights requires an examination of potential solutions to the assurance problem. These considerations will simultaneously provide an analysis of the spontaneous evolution of the rules and potential for contracting to voluntarily transfer property rights (e.g., trade), since the same assurance problem stands in the way of all contracts, whether they involve trading commitments to respect each others' property rights and to cooperate in watching, walling and wariness, or they involve trading goods and services.

II.2. *Assurance Problems and Sources of Credibility in Contracting*

There would be no assurance problem if everyone had full knowledge, but such perfect knowledge does not exist anywhere except in some economists' mathematical models. Information is so scarce in the real world that trust or recourse generally must substitute for full knowledge in order to make promises credible. Trust, a willingness to make oneself vulnerable to another even in the absence of external constraints, certainly can evolve to support cooperative interaction such as contracting in the creation of protection arrangements or in trade, as explained below, but it takes time, and under some circumstances it can be limited to relatively small communities. If individuals are going to be willing to deal with others that they do not trust, recourse in the form of credibly threatened sanctions against breaches of promises (perhaps supported by a third party dispute resolution) is a necessary substitute. While trust and recourse are both substitutes for knowledge, they are not perfect substitutes for each other. Tradeoffs in transaction costs mean that under some circumstances trust provides a superior solution to assurance problems, while recourse may be more desirable under other conditions (Benson 2001a). Furthermore, there are alternative institutional mechanisms for the provision of recourse, and they also are imperfect substitutes. In order to illustrate this, let us consider some of the differences between alternative institutional sources of trust and recourse.[11]

---

[11] Obviously, the legal systems of nation-states are potential sources of recourse. Unfortunately, in many countries these legal systems do not provide consistent and predictable recourse. Despite the tremendous degree of government failure all over the world, however, many consultants and academics contend that the solution to the assurance problem must come from the state. In writing about law in the newly independent countries of the former Soviet Union, for instance, Ioffe (1996, p. 95) maintains that legislation of commercial law "must now be comprehensive [due to] . . . the emergence of gaps in the law [and] the restructuring of the former Soviet economy which requires new legal regulation"; and later that "the commercial code, not taken literally, must encompass all forms of economic activity, both in production and trade." Arguments such as these fail to recognize that trust is an alternative to recourse,

II.2.a. Building trust

Many of the concepts from game theory are useful in demonstrating the gains from cooperating and defecting in various contexts, and therefore, in thinking about determinants of trust (e.g., see Axelrod [1984]; Ellickson [1991]; Ridley [1996]; Vanberg and Congleton [1992]).[12] Game theory demonstrates that cooperation can arise through repeated interactions, for instance (Axelrod 1984). Perhaps, for instance, at some point, different individuals decide to claim adjoining properties. Facing the likelihood of repeated interaction, they form relatively tentative bilateral relationships based on reciprocity incentives, implicitly promising to recognize each others' territorial claims. Because the long-term reciprocal response is uncertain, a repeated-game situation does not guarantee unconditional cooperation even with tit-for-tat threats to reinforce the positive incentives associated with remaining on good terms with the other party(ies) (e.g., relatively secure property rights, the potential to focus resources to produce wealth rather than violence). The dominant strategy still depends on expected payoffs, frequency of interaction, time horizons, and other considerations (Tullock 1985, p.1073; Ridley 1996, pp. 74-75; Rutten 1997). Furthermore, in emerging economic and social activities, repeated-dealing arrangements must be established. For instance, McMillan and Woodruff (1998), in their study of emerging trade in Vietnam explain that an entrepreneur tends to be very cautious when considering a potential trading partner.[13] As a result, building trust can take time, of course, and that is obviously one of the drawbacks of exclusively relying on trust relationships.

Individuals may be able to gain the trust of others relatively quickly by offering some sort of bond or hostage (Williamson 1983). When an unknown party posts a bond with some trusted third party (e.g, a reputable bank) as a guarantee that his promises are credible, for instance, he may be able to overcome a lack of trust. Similarly, individuals can invest in signals

---

and that there are non-state sources of recourse. In this context, recognize that there are many analogies between emerging markets and institutions in geographic space, and the emerging Internet market and institutions (Benson 2000c). These similarities are alluded to below.

[12] North (1990, p. 15) explains that game theory "does not provide us with a theory of the underlying costs of transacting and how those costs are altered by different institutional structures." An understanding of the evolution of rules and property rights really requires consideration of the factors that lead to a transition from one institutionalized game setting to another and another and so on, as suggested below, rather than the analysis of a particular game. Thus, game theory can only serve as a supplement to the more fundamental institutional analysis outlined here.

[13] A Vietnamese entrepreneur often visits the plant of the firm he is considering in order to see if the facility appears to be permanent and efficient. He inspects the output of the plant, asks other trusted traders if they have dealt with or know about the potential partner, and so on. The information gathered can never be perfect but if it is positive, a small trade is often arranged. If that one works out, the next one is larger. It is only after several deals that the transactions reach a level that involve a substantial commitment.

that demonstrate a commitment to high quality (Klein and Leffler 1981; Shapiro 1982, 1983; Diamond 1989). As an example, consider Nelson's (1974) explanation of advertising of experience goods. He notes that such advertising serves two primary functions for the rational buyer, and neither of these functions focus on the provision of direct information about the quality of commodities that are advertised: first, "advertising relates brand to function" and provides information about the general uses of the product, but second and more important, the volume of advertising is a signal to buyers that shows the extent of committed investment by the seller. What matters most to a rational buyer is not what advertising says about quality, but simply that it is a recognizable investment in non-salvageable capital: brand name. Investments in other non-salvageable assets (e.g., elaborate store fronts, charitable contributions, community service) can serve the same function. Essentially, investments in non-salvageable assets are offered as a bond to insure credibility. Information transmission is a key to the success of such a business strategy (e.g., see Milgrom, North, and Weingast [1990]), as buyers must be aware of such commitments. If they are, then, as Klein and Leffler (1981) explain, the marginal cost to buyers of measuring such specialized or non-salvageable investments must also be less than the prospective gains: "If the consumer estimate of the initial sunk expenditure made by the firm is greater than the consumer estimate of the firm's possible short-run cheating gain" he or she will tend to trust the seller. When effective recourse is not available or is relatively costly, individuals who want to enter into cooperative relationships such as trade have very strong incentives to make such investments. Time is required to build reputations through these kinds of processes, of course, so new entrants may have to suffer through a considerable period of losses before they can expect to see investments in reputation building pay off. Indeed, since the payoff to such investments are delayed and very uncertain, incentives to make them tend to be relatively weak, and the emergence of cooperation based on such sources of recourse also can be quite slow.[14] Some individuals may have reputations that they have developed in other activities that

[14] Much of this uncertainty is due to the state, however. As Pejovich (1995) notes, "The arbitrary state undermines the stability and credibility of institutions, reduces their ability to predict the behavior of interacting individuals, raises the cost of activities that have long-run consequences, and creates conflicts with the prevailing informal rules. . . . [M]ost countries in Eastern Europe [and many other parts of the world] are arbitrary states." When property rights are insecure due to potentially arbitrary and/or opportunistic behavior by government (e.g., changes in tax policy to capture the quasi-rents that arise with investments in reputation), incentives to invest in reputation or to count on future dealings are weak and the kinds of private sanctions discussed below are likely to be relatively weak. But that also means that the state cannot be relied upon to provide consistently effective recourse, as traders clearly recognize (even if policy "experts" do not). McMillan and Woodruff's (1998) interviews of entrepreneurs in Vietnam show that despite their frequent reliance on informal sanctions (tit-for-tat, exit, spreading information about non-cooperative behavior), these entrepreneurs do not want the state to get involved in contract enforcement because they do not trust the state.

can be transported into the new situation. Firms with international reputations may enter an emerging market and become established very quickly, for instance.

Markets for reputation can also develop. For instance, a firm or other organization might develop a reputation for honestly assessing the quality of other firms' products or services. When someone wants to enter a market and quickly establish a reputation, he or she can pay to have a product or service performance inspected/tested and "certified" by this reputable assessment organization (Carter and Manaster 1990; Anderson, Daly, and Johnson 1999). Moody's, Standard & Poor's, Underwriters Laboratory and the Good Housekeeping Seal of Approval come to mind. Similarly, the American Automobile Association inspects and rates motels.

Reputation can be purchased in other ways as well. One possibility is to contractually affiliate with a recognized organization which requires that all its affiliates meet specified standards for their products or services. Best Western motels are one such example. Locally owned franchises of regional or national chains of motels, restaurants, and retail outlets provide other examples. All such arrangements involve non-salvageable investments since failure to maintain quality will result in loss of the reputation signal that has been purchased.

A related phenomenon is the growth of specialists who gather and sell independent assessments on products and services. Consumer Reports is an obvious example, but there are many others (e.g., restaurant reviewers; movie and Broadway critics; travel magazines and books; rankings of colleges and graduate schools). Others collect information about potential buyers. Credit reporting agencies provide information about potential debtors to potential lenders, for instance (see Klein [1992] for relevant discussion). Individuals who want to build reputations for quality and/or reliability may pursue endorsements from such independent evaluators.

II.2.b. From trust and reputation to recourse

Most arguments about the inability of private parties to cooperate without the backing of a coercive power are explicitly or implicit prisoners' dilemma arguments. As suggested above, the one-shot prisoners' dilemma analogy does not characterize many kinds of interactions. When repeated dealing arrangements are valuable, each individual has implicit threat of punishment if the other party fails to live up to a promise, commits fraud or behaves opportunistically: the tit-for-tat response. As more bilateral relationships are formed in recognition of the benefits from cooperation, a loose knit group with intermeshing reciprocal relationships often begins to develop. The fact is that individuals are generally involved in several "communities" as described by Taylor (1982, pp. 26-30), wherein "the relations between members are direct and . . . many-sided" (also see Ellickson [1993]), and in such communities a tit-for-tat becomes a less significant

threat. An exit threat becomes credible when each individual is involved in several different games with different players, in part because the same benefits of cooperation may be available from alternative (competitive) sources (Vanberg and Congleton 1992, p. 426). When competitive alternatives within a community of transacting individuals make the exit option viable, Vanberg and Congleton (1992, p. 421) suggest that one strategy that can be adopted is unconditional cooperation until or unless non-cooperative behavior is confronted, and then imposition of some form of explicit punishment of the non-cooperative player as exit occurs. They label such a strategy as "retributive morality," and the "blood-feuds" of primitive and medieval societies provide examples of such behavior. This practice of retributive morality strengthens the threat to non-cooperative behavior. However, the fact is that retributive morality or the blood feud played a much less significant role in primitive and medieval societies than is popularly perceived (Benson 1991a, 1994a). After all, such violence is risky, and there is an even better alternative.

When individuals cooperatively engage in successful bilateral relationships within an evolving web of such relationships, others are likely to notice their cooperative behavior and attempt to initiate mutually beneficial relationships with them. In other words, when information about cooperation spreads, such behavior in one relationship can serve as investment in building a reputation for fair dealing, and this reputation can attract more opportunities. Importantly, however, when information spreads about non-cooperative behavior, all of the beneficial relationships that the non-cooperative individual enjoys within the community can be put in jeopardy. All members of the community have an exit option, and therefore they may cut off all relationships with someone who has proven to be untrustworthy in dealings with anyone else in the group. This means that there is a low cost option to retributive morality: unconditional cooperation whenever an individual chooses to enter into some form of interaction, along with a refusal to interact with any individual who is known to have adopted non-cooperative behavior with anyone in the group and the spread of information about untrustworthy people. Vanberg and Congleton (1992) refer to this response as "prudent morality," and given that reputation information spreads quickly within a group, the consequences of retributive and prudent morality become quite similar. If everyone spontaneously responds to information the result is social ostracism, a very significant punishment, even though it is not explicitly imposed by a single retributive individual. Essentially, each individual's reputation is "held hostage" by every other individual in the evolving group, a la Williamson (1983), and reputation is an ideal hostage. It is highly valued by the individual who has invested in building it, so a credible threat of destroying it can be a significant deterrent, and the threat is also credible because the reputation hostage has no value to the hostage holder and the cost of destroying it (spreading truthful information) is low. That is, it is a non-salvageable asset (an asset that might be built

relatively quickly by offering bonds or investing in even more non-salvageable assets, as noted above, thus increasing the value of the hostage). Such a threat of ostracism can be a very powerful source of recourse, where the "third party" providing the threatened sanction is the community of individuals who receive and respond to the information about non-cooperative behavior.

The spontaneous development of social ostracism illustrates another point. As an informal community evolves from a web of bilateral trust relationships, group-wide norms also evolve. Note, in this context, that it is not the existence of "close-knit" communities that generates group-wide norms, as some have contended. Instead, norms and communities can evolve simultaneously as each affects the other: the evolution of norms of cooperation lead to the development of a web of interrelationships that can become a "close-knit" community, and the development and extension of such a community in turn facilitates the evolution of more effective norms (Benson 1999a). Thus, as Vanberg and Congleton (1992, p. 429) conclude, perceptions "of what is moral vary with relevant differences in exit costs. At the high-cost end of the spectrum, moral justification for tit-for-tat and retributive behavior seems to be fairly common, whereas prudent morality gains in importance as we move to the low-cost end."

Many group-wide norms are simply commonly adopted trust rules that apply for all interactions in a web of relationships. As this web of relationships becomes a community, other rules can arise. Called "solidarity rules" by Vanberg and Buchanan (1990, pp. 185-86), these are expected to be followed by all members of the group, because individual sacrifices associated with obeying solidarity rules produces shared benefits within the group (Vanberg and Buchanan 1990, p. 115). Solidarity rules are things like "do not behave recklessly and put others at risk." However, they can also involve rules about individuals' obligations in cooperative production of rule-enforcement functions. Rules like "inform your neighbors about individuals who violate trust rules," and "do not cooperate with individuals who behave in a non-cooperative fashion with someone else," are solidarity rules in the sense that production of information and ostracism create benefits for everyone in the group by deterring non-cooperative behavior.

Significant limits on abilities to reason and to absorb knowledge means that individuals are not able to use conscious reason to evaluate every particular option in the array of alternatives that are available (O'Driscoll and Rizzo 1985, pp. 119-22; Hayek 1937, 1973). Therefore, rational individuals will often find it beneficial to voluntarily conform to a community's rules in an almost unthinking way. In this context, as Ridley (1996, p. 132) notes, "Moral sentiments . . . are problem-solving devices [that evolve] . . . to make highly social creatures effective at using social relations [by] . . . settling the conflict between short-term expediency and long-term prudence in favor of the latter." People conform to all sorts of faddish and ritualistic behaviors, and even though they may appear to have

nothing to do with evolving moral sentiments, they actually may have similar functions: facilitating cooperation. After all, while individuals want to identify and exclude non-cooperative players, they also have strong incentives to identify themselves as cooperative (Ridley 1996, p. 139). Outward conformity to a group's fads and rituals can serve as a signal of willingness to cooperate in order to be in a position to reap the rewards from participation in joint production and other forms of interaction within the evolving group. As Ridley (1996, p. 188) explains, "We are designed not to sacrifice ourselves for the group but to exploit the group for ourselves."

Incomplete knowledge, scarcity, and transactions costs mean that someone alleged to have violated a trust or solidarity rule may not be guilty, so "disputes" over guilt or innocence arise. Indeed, confrontations can arise under two different conditions in an expanding or increasingly dynamic group. In addition to disputes regarding alleged violations of norms, disputes can also arise when property rights that are not clearly defined become valuable and conflicting claims to those rights are asserted. Some (and in a close-knit community, most) disputes can be solved by direct bargaining, but transactions costs can prevent successful bargaining in some cases. Therefore, third-party dispute-resolution institutions are desirable in order to reduce the chances of violent confrontation (a dispute resolution process that can create considerable costs for other members of a community), and to increase the chances that the community can survive so its members can prosper from the mutually beneficial interactions it supports. Public courts are one source of such dispute resolution, but there are many other options as well. Contracting parties can specify some sort of alternative dispute resolution (ADR), be it mediation, arbitration, or some combination of the two. Individuals may also choose ADR after a dispute arises, even if they have not specified the option in a formal contract, possibly because they want to maintain a good relationship or because other members of the community apply social pressure.

II.2.c. Third party dispute resolution

Voluntary acceptance of ADR means that the selected third party must be acceptable to both disputants, so "fairness" is embodied in the dispute-resolution process. There are a wide variety of potential sources of ADR. Specialists can be selected from organizations like the International Chamber of Commerce (ICC), the American Arbitration Association (AAA), or any number of other private dispute resolution providers, including private ADR firms and for-profit "courts."[15] Mechanisms for ADR (e.g., arbitrator and mediator) selection actually vary widely, but they all are designed to

---

[15] See for example, Phalon (1992), Ray (1992), and Benson (1998d) for discussion of the developing private-for-profit court industry in the United States.

guarantee the selection of an unbiased third party who will apply the norms that the parties share within their relevant community.[16]

In general, the choice of an ADR provider is made without requiring explicit agreement by the two parties while still allowing for prescreening, and possibly more than one level of screening (Benson 1999b, 2000a, 2000b). For example, one common selection method involves a pre-approved list of professional mediators or arbitrators determined by contracting parties (or their community organization, as suggested below), so if a dispute arises, a person is chosen from the list by some preset mechanism (e.g., random selection, rotating selection, selection by a third party such as a governing board of a trade association). Empirical evidence indicates that selection of arbitrators for a pre-approved list is based on the reputation of the arbitrators for impartiality and expertise in contractual matters that might arise (Ashenfelter 1987; Bloom and Cavanagh 1986). Another common arbitrator selection system gives the parties the resumes of an odd numbered list of arbitrators from a larger pre-selected group (e.g., pre-selected by a community such as a trade association, as noted below, or provided by an organization like the ICC or the AAA), with each party having the power to successively veto names until one remains. Thus, a second level of screening is added at the time of the dispute, contributing "to the legitimacy of the arbitrator and his award in the eyes of the parties" (Bloom and Cavanagh 1986, p. 409). Since the parties are given the arbitrators' resumes, they have information about experience, training, the nature of awards given in the past, and so on. A similar practice provides the parties with a list and resumes of an odd number of potential arbitrators from a pre-approved list, with each disputant having the power to veto one less than half and rank the others, and the arbitrator who is not vetoed by either party and has the highest combined rank is chosen. Both sides of the dispute may also provide a list of a fixed number of mediators or arbitrators with each being able to veto any or all of the names on the other party's list; if all

---

[16] There also is considerable variation in the institutional arrangements themselves. Some communities rely almost exclusively on mediation backed by social pressure to voluntarily reconcile differences (e.g., see Benson [1991b] for discussion of the Quakers and some of the other religious based groups in early America). Others appear to rely more heavily on arbitration. The preceding statement includes the word "appear" because it may be that mediation efforts are informal, so they are not easily observed, while arbitration arrangements are more formal and open to public observation. Arbitration often can be quite "public" when community backing is required, for instance, as among the Yurok and other Native American communities of Northern California during the early nineteenth century (Benson 1991a), Anglo-Saxon communities (Benson 1994a), and historical commercial communities (Benson 1989). Of course, arbitration can also be very private, as in many modern commercial situations, for reasons such as those discussed in Benson (1999a, 2000b). Many communities probably employ a combination of mediation and arbitration. Diamond merchants mandate that disputes go through a conciliation (mediation) process before they can go to arbitration, for instance, and most disputes are actually resolved through this consolidation process (Bernstein 1992). Arbitration arises only when mediation clearly cannot achieve a solution.

names are vetoed each provides another list and the process is repeated (clearly, this procedure requires that both parties want to arbitrate, so they do not continue to provide unacceptable names). All such systems are intended to guarantee the appointment of an third party without requiring explicit agreement by the disputants while still allowing for prescreening, and possible more than one level of screening, of the potential mediators or arbitrators.

Biased rulings are not likely in such a competitive environment where potential arbitrators are chosen beforehand by the trading community (e.g., as in the diamond traders associations) or where both parties have the power to reject judges proposed by the other party. Furthermore, successful arbitrators will be those who consistently apply the norms that members of the relevant community expect to be applied. Indeed, by choosing an arbitrator/mediator attempting to build a reputation of trustworthiness, strong incentives are created for those aspiring to be chosen as third-parties in dispute resolution to avoid the appearance of bias. The chosen arbitrator/mediator must convince individuals in the group that a judgment should be accepted voluntarily, and that he has no coercive power to enforce it. More importantly, an appearance of bias will damage the individual's reputation. The ruling can therefore be backed by an implicit threat of ostracism, although in general, dispute resolutions are likely to be accepted because individuals recognize the benefits of behaving in accordance with community members' expectations, not because they fear ostracism (Pospisil 1971; Benson 1989, 1991a).

II.2.d. Recourse through contractual associations

Both commitments and threats can be made more credible, and some uncertainty can be eliminated, if individuals with mutual interests in long-term interactions form "contractual" groups or organizations rather than waiting for trust or reputation institutions to evolve more slowly into informal communities. Potential contractual arrangements are numerous, including the implicit contracts of family bonds and ethnic or religious networks, clubs and other social organizations, and in the commercial area, arrangements such as indirect equity ties through pyramidal ownership structures, direct equity ties, interlocking directorates, and trade associations. As Khanna and Rivkin (2000) explain, for instance, business groups are actually "ubiquitous in emerging economies" (as evidence, they cite a large number of studies about groups such as *grupos* in Latin America, business houses in India, and *chaebol* in Korea). Many of these associations may form for reasons other than the development and enforcement of norms, but once they develop, the cost of adding such functions is relatively low. In addition to creating strong bonds that facilitate interaction, an affiliation with such a group can be information generating in that it can imply a credible signal of reputable behavior.

A contractual organization can provide a formal mechanism to overcome frictions in communication, insuring that information about any individual's non-cooperative behavior will be transmitted to others in the contractual community. Then group membership can include a contractual obligation to boycott anyone who fails to follow the group's contractually accepted rules: specifically, any non-cooperative party will be automatically expelled from the organization. Such automatic ostracism penalties make the reputation threat much more credible (Williamson 1991, p. 168).

These groups can also lower transactions costs by establishing their own unbiased dispute resolution arrangements. Because not all allegations of non-cooperative behavior are necessarily true, they may have to be verified. Within some organizations a single mediator or arbitrator or panel of mediators or arbitrators is chosen for a set period to arbitrate all disputes between members. Thus, prescreening by the formal group occurs as ADR providers are chosen from a competitive pool by the association through its membership approved selection process. For instance, in the diamond industry, arbitrators are elected from the organization's membership (Bernstein 1992, pp. 124-25). In many religious organizations, the religious leadership (e.g., priests, elders) provides this function. Those selected are likely to have considerable standing (reputation) within the community. They have strong incentives to maintain their own reputation for fairness, and are not likely to be biased or easily corruptible. These services do not have to be produced internally, the group may contract with external mediation and/or arbitration specialists.

When a dispute involves new and unanticipated issues, an ADR supplier may be required to determine what rules should be applied to the situation. In such context, Lew (1978, p. 589) explains that "Owing no allegiance to any sovereign State, international commercial arbitration has a special responsibility to develop and apply the law of international trade." The "law" that dominates international trade has, for the most part, evolved through contracting and the use of arbitration, as explained below.

II.2.e. Customary Law

In modern societies, the most visible types of rules are the "laws" designed and imposed by those with authority in nation-states, but as noted above, there are other rules (e.g., habits, conventions, norms, customs, traditions, or standard practices) that are much more important determinants of behavior in many aspects of human activity. A key distinguishing characteristic of such rules is that they are initiated by an individual's decision to behave in particular ways under particular circumstances. As Hayek (1973, p. 97) emphasizes, adopting a behavioral pattern creates expectations for others who observe it and this creates an obligation to live up to those expectations. Furthermore, as Mises (1957, p. 192) explains, when individuals who interact with one another observe each others' behavioral patterns

they often emulate those that appear desirable so such behavior and accompanying obligations spread. In other words, these rules evolve spontaneously from the bottom up rather than being intentionally designed by a legislator, and they are voluntarily accepted rather than being imposed. No central "authority" with coercive powers is necessary to produce the resulting cooperative social order, as obligations are largely self-enforcing: it pays for each party to behave as expected in order to be able to expand wealth over the long run through mutually beneficial interaction.

Pospisil (1971, 1978) distinguishes between "legal" arrangements that evolve from the top down through command and coercion, which he calls "authoritarian law," and systems of trust and solidarity rules that evolve from the bottom up through voluntary interaction, which he refers to as "customary law" (also see Fuller [1964, 1981]).[17] Such a norm-based cooperative arrangement often can be characterized as a "legal system" following Hart's (1961) definition of law, since, as implicitly suggested above, it has primary rules of obligation (e.g., recognized norms), and it can be backed by secondary rules or institutions of recognition (e.g., reciprocities, mechanisms to spread information about reputations, ostracism, mutual

---

[17] The term, customary law, is problematic, of course, because, as Pospisil (1978) explains, it has more than one definition. The term can refer to rules that are not codified and have been relied upon by the members of a group, unchanged "from time immemorial." Customary underpinning of the common law are often treated in this way by judges, for instance. This definition is highly questionable whenever a careful study of the origins of customary law is performed, however, because customary norms can actually evolve quite rapidly (Pospisil 1971; Benson 1989; Trakman 1983). A second and more complete definition was used by the Commentators of Roman law in the thirteenth and fourteenth centuries. The Commentators also emphasized *longa conuetudo* or long use, a questionable criterion for reasons just noted, but "Their second criterion, however, seems to be much more significant for scientific research. *Opinio necessitatis*, the requirement that, to be regarded as customary, a law must be backed by the people's 'conviction of its indispensability' and desirability . . . , brings out the basic characteristic of the term" (Pospisil 1978, pp. 63-64). In other words, the vast majority of the people in a group view a customary law to be binding and desirable, so the law is "internalized" as if through a voluntary contract. Thus, such a law guides everyone's actions within a group and makes behavior relatively certain or predictable. Fuller (1981, p. 213) explains that,

> To interact meaningfully men require a social setting in which the moves of the participating players will fall generally within some predictable pattern. To engage in effective social behavior men need the support of intermeshing anticipations that will let them know what their opposite members will do, or that will at least enable them to gauge the general scope of the repertory from which responses to their actions will be drawn. We sometimes speak of customary law as offering an unwritten code of conduct. The word code is appropriate here because what is involved is not simply a negation, a prohibition of certain disapproved actions, but of this negation, the meaning it confers on foreseeable and approved actions, which then furnish a point of orientation for ongoing interactive responses.

This view of customary law is adopted here, with the added condition that the customary norms are supported by processes of adjudication and change, a la Hart (1961), as discussed below. After all, Hart (1961, p. 97) includes "customary practice" as one possible "authoritative criteria" of legal validity (i.e., "rule of recognition"). Hart probably is using the term as in the first definition listed above, of course, but the contention here is that general acceptance implies validity, as do institutionalized means of spreading information about misbehavior, and of ostracism.

insurance, cooperative policing),[18] adjudication (e.g., negotiation, arbitration, mediation), and change (e.g., innovations in behavior followed by observation, emulation and conformity, contracting, dispute resolution).

For an obligation to achieve the status of a "customary law" it must be recognized and accepted by the individuals in the affected group. In other words, a strong consensus rule applies, and as a result, customary law tends to be quite conservative in the sense that it guards against mistakes. Nonetheless, flexibility and change often characterize customary law systems (Popisil 1971; Benson 1989, 1998b). For instance, if conditions change and a set of individuals decide that, for their purposes, behavior that was attractive in the past has ceased to be useful, they can voluntarily devise a new contract stipulating a new behavioral rule. Thus, an existing norm (custom) can be quickly replaced by a new rule of obligation toward certain other individuals without prior consent of, or simultaneous recognition by, everyone in the group. Individuals entering into contracts with these parties learn about the contractual innovation, however, and/or others outside the contract observe its results, so if it provides a more desirable behavior rule than older custom, it can be rapidly emulated. Contracting may actually be the most important source of new rules in a dynamic system of customary law (Fuller 1981, p. 157).[19] For instance, many innovations in commercial law

---

[18] Positive incentives to recognize trust rules are strong because they arise voluntarily through mutually-beneficial interactions. Incentives to violate such rules can arise under some circumstances, of course, but negative incentives also arise through the threat of spreading of information about misbehavior and the resulting ostracism. Since solidarity rules produce benefits for everyone in the group, free-rider incentives arise, just as with any other jointly produced products. However, free riding is successful only to the extent that a free rider cannot be excluded from consuming benefits. Thus, as solidarity rules develop, the scope of the ostracism solidarity rule itself is likely to expand to include "do not interact with anyone who does not obey other solidarity rules." Therefore, solidarity rules are not public goods, as non-free riders are the only members of a group who are likely to retain membership in a customary law community.

[19] Contractual and customary processes can easily become intertwined. As Fuller (1981, pp. 224-25) explains, "If we permit ourselves to think of contract law as the 'law' that parties themselves bring into existence by their agreement, the transition from customary law to contract law becomes a very easy one indeed." Indeed, Fuller (1981, p. 176) argues that a sharp distinction between custom and contract is inappropriate:

> if problems arise which are left without verbal solution in the parties' contract these will commonly be resolved by asking what "standard practice" is with respect to the issues in question. In such a case it is difficult to know whether to say that by entering a particular field of practice the parties became subject to a governing body of customary law or to say that they have by tacit agreement incorporated standard practice into the terms of the contract.
>
> The meaning of a contract may not only be determined by the area of practice within which the contract falls but by the interaction of the parties themselves after entering the agreement. . . . The meaning thus attributed to the contract is, obviously, generated through processes that are essentially those that give rise to customary law. . . . [In fact,] a contract [may be implied] entirely on the conduct of the parties; . . . the parties may have conducted themselves toward one another in such a way that one can say that a tacit exchange of promises has taken place. Here the analogy between contract and customary law approaches identity.

have been initiated in contracts and dispersed quickly through the relevant business community (Benson 1989, 1998b).

Alternatively, as conditions change, the inadequacy of existing customary rules can be revealed when a dispute arises. Negotiation is probably the primary means of dispute resolution for members of a customary law community, reinforcing the contention that contracting is a primary mechanisms for initiating rapid change in customary law. If direct negotiation fails, however, the parties to a dispute often turn to an arbitrator or a mediator. Since a dispute can suggest that existing rules are unclear or insufficient, new customary rules can be and often are initiated through third-party dispute resolution (Fuller 1981, p. 110-11; Lew 1978, pp. 584-89; Benson 1989, 1998b). Unlike public court precedent, such a decision only applies to the parties in the dispute, but, if it effectively facilitates desirable interactions, the implied behavior can spread rapidly through the community, becoming a new rule.

No community evolves in complete isolation. Anthropological and historical evidence suggests that intra-group conflict has been an almost ubiquitous characteristic of human history. Since a key function of customary-law communities is to establish and secure private property rights, and such rights are insecure if outsiders are able to "invade" and take the property, one joint product of a cooperative group is likely to be mutual defense. In fact, an external enemy can strengthen group cohesion (Wesson 1978, p. 184; Ridley 1996, p. 174), leading some to actually suggest that norms are important because they enable groups to be sufficiently united to deter their enemies, not because they allow people to create order (coordination)

In essence, individuals are able to establish their rules of obligations toward one-another through practice and observation or through negotiation and explicit agreement. Thus, customary legal arrangements may be predominantly contractual. In fact, one reason for development of contracts in a customary law system is that individuals often base their expectations of how others will act, and determine how they should act, through observation of passed events. The resulting norms tend to be backward looking. Third party dispute resolution also tends to be backward looking. Arbitrators often justify their decisions by placing them in the context of past practices, for instance, in order to maintain a continuity in the law: custom and tradition rule. Therefore, if these are the only means of legal change, customary law might evolve very slowly. Certainly, customary law does tend to be conservative, but contracts provide a source of forward-looking voluntary legal change that can produce rapid but beneficial alterations in the status quo.

The expanding use of contract and development of contractual arrangements is, in fact, a natural event in the evolution of customary law. As customary legal arrangements evolve and are improved upon, they tend to become more formal, and therefore, more contractual. In addition, as a group develops and expands so that the trust relationships that characterize small group interaction do not apply, conflicts are avoided by explicitly stating the terms of the interaction *a priori*; that is, by contracting. A carefully constructed and enforceable contract can substitute for trust. Thus, with the evolution of contracts and enforceable dispute resolution mechanisms, the original bases for trust rules become relatively less important and the group can grow beyond the bounds of bilateral trust and even reputation mechanisms. Indeed, inter-group interaction can arise, given inter-group acceptance of contracts and dispute resolution.

within their groups (Alexander 1987). Clearly, norms that support the production of mutual defense evolve, and an important part of an individual's belief system will be a "communitarian" one (e.g., tribalism and "a concept of them and us" where individuals are expected to aid in the defense of the "community"). In fact, however, these communitarian norms simply evolve along with intra-group norms of cooperation as part of the overall objective of creating an environment conducive to the pursuit of subjective well-being.

II.2.f. Polycentric governance

Inter-group competition does not have to involve violence. Voluntary associations imply both the ability to voluntarily join a group, given acceptance by existing members, and to voluntarily withdraw. Inter-group movements are a distinct possibility (e.g., see Pospisil [1971]; Umbeck [1981a]; and Benson [1991a]). Indeed, individuals have incentives to "migrate" to the group which best facilitates pursuit of their objectives. Furthermore, individuals in groups that gain cooperative members enjoy more opportunities for mutually beneficial interaction so they have incentives to compete to attract or hold membership (the basis of customary law groups may be geographic proximity, but it also may be kinship, functional proximity as in a trade association or the "business community," religion, or any of a number of factors that create repeated dealings and/or reputation effects). As a result, members of customary legal systems have incentives to imitate desirable institutions and rules developed elsewhere. Competition and emulation lead to standardization of many rules and institutions across similarly functioning groups, although differences may remain, reflecting preferences of various groups' members.

A group does not necessarily have to expand to expand opportunities for beneficial interaction. If individuals want to interact, but only on some dimensions, or if they want to maintain different sets of rules for different dimensions of interaction, then parallel "localized" mutual support groups may be maintained while a "second order of clustering" (Vanberg and Buchanan 1990, p. 189) is established, facilitating a relatively limited scope for interaction.

A group whose members insist on strictly imposing their own morality and penalties on outsiders would probably be unable to initiate beneficial inter-group interaction. Thus, if people wish to simultaneously facilitate inter-group interaction and impose laws that differ substantially from the norm in other groups, they have strong incentives to inform outsiders of the differences in order to avoid conflict and minimize the difficulty of maintaining non-standard laws. Part of the reciprocal agreements with other groups may be the explicit recognition of differences in laws and procedures for treating conflicts. This in turn implies that as inter-group interactions expand, a hierarchical jurisdictional arrangement may become neces-

sary. For example, each localized group may have jurisdiction over rules for and disputes between its members. Rules for inter-group interactions can differ from both groups' internal rules, although they certainly do not have to, and disputes between members of a confederation of different groups are settled by some "higher" confederation level adjudication process (e.g., see Pospisil [1971] and Benson [1994a]). Note that these are not "higher courts" where disputes can be appealed from within-group dispute resolution mechanisms (e.g., they are not a supreme authority). Rather, this is a jurisdictional hierarchy defining the role of each adjudication process and allowing for increasingly more distant interactions (as in Pospisil [1971] and Benson [1994a]). This allows for differences between the law applied within groups and between groups (Pospisil 1971); a monopoly in law is not required.

A judgment involving an inter-group dispute will have to be considered to be a fair one by members of both groups. Thus, an equal number of individuals from each group might serve as an arbitration board (e.g., see Benson [1994a]), or a mutually acceptable third party (i.e., an arbitrator or mediator with a reputation for good judgment) might be chosen (e.g., see Pospisil [1971] and Benson [1989]). This provides another reason for the tendency toward standardization of rules across parallel groups with similar functions, at least for those functions carried out in the process of inter-group interactions.

Some individual members of each group must recognize the potential benefits of inter-group interaction and be willing to bear the cost of initiating institutional innovations. Furthermore, the resulting innovations must involve more than just dispute resolution, because such interaction faces an assurance problem. Individuals must feel confident that someone from the other group will not be able to renege on a promise and then escape to the protection of that other group. After all, at least initially, repeated game and reputation effects are localized within each group, and there is limited potential for a boycott sanction. Thus, for second order clustering to develop, some sort of inter-group insurance or bonding arrangement becomes desirable, along with an apparatus for inter-group dispute resolution. For instance, as inter-group interaction develops the mutual support group can become a surety group as well (Friedman 1979; Solvason 1992, 1993). Membership in a group then serves as a signal of reputable behavior, and if a member of a group cannot or will not pay off a debt to someone from the other group, the debtor's group will. The individual then owes his own group members so the boycott threat comes into play once again.

Limits to the extent of an inter-group network of cooperation are determined by the relative costs and benefits of information about other groups and their legal systems. The costs of establishing inter-group legal arrangement depend in part on how "distant" the groups are from one another, where distance can be in terms of geographic space, or in terms of the behavioral norms that are relevant to the groups. Thus, extensive interac-

tion between starkly different groups may not arise. However, these limits are stretched as individuals become members of several groups. After all, as Mises (1957, p. 257) explains, "Man is not the member of one group only and does not appear on the scene of human affairs solely in the role of a member of one definite group. . . . The conflict of groups is not a conflict between neatly integrated herds of men. It is a conflict between various concerns in the minds of individuals." Thus, the relatively limited jurisdictions of some customary communities are not as constraining as they might appear to be. A person may simultaneously belong to many groups that have well established customs (and be subjects to the commands of several rule-making authorities, e.g., as in a formal federalist system of government), so being in one community does not preclude dealing with people in other communities. A person may belong to a trade association, a homeowners association, a religious group, a fraternal organization, and so on, for instance, each with its own rules and governance institutions. The membership of all of these communities can differ, although considerable overlap may also occur, so individuals may deal with other individuals on some dimensions but not on all dimensions. Indeed, in any complex society, there are many distinguishable systems of rules and institutions, and yet people from many of these different systems interact regularly without having to call upon any legal authority. Thus, inter-group cooperation appears to be the norm rather than the exception, and it appears to be quite widespread. And with good reason: as Gluckman (1955, p. 20) suggests, "multiple membership of diverse groups and in diverse relationships is . . . the basis of internal cohesion in any society." An all inclusive legal system would eliminate the benefits of competition and emulation and undermine the incentives for innovation, as Berman (1983, p. 10) explains:

> It is this plurality of jurisdictions and legal systems that makes the supremacy of law both necessary and possible. . . . The very complexity of a common legal order containing diverse legal systems contributes to legal sophistication. Which court has jurisdiction? Which law is applicable? How are legal differences to be reconciled? . . . The pluralism of . . . law, . . . has been, or once was, a source of development, or growth—legal growth as well as political and economic growth. It also has been, or once was, a source of freedom.[20]

The competitive/cooperative relationship between consensual customary legal systems is driven by the desire to facilitate voluntary mutually-beneficial interactions rather than a desire for legal sovereignty. Thus, many different customary systems can co-exist and interact. An under-

---

[20] Berman's "or once was" phrase recognizes that diverse legal systems are increasingly being subjugated by authoritarian legal systems. Indeed, while consensual legal arrangements tend to be characterized by internal stability, they face a significant external threat to stability. The size of consensual groups and second order clusters are constrained by transactions costs, and in many cases such organizations have been unable to resist takeover by groups cooperating in the production of violence. This issue is explored below.

standing of customary law requires that individuals and their organizations be the points of reference rather than "society" as a whole: "there may then be found utterly and radically different bodies of 'law' prevailing among these small units, and generalization concerning what happens in 'the' family or in 'this type of association' made on the society's level will have its dangers. The total picture of law-stuff in any society includes along with the Great Law-stuff of the Whole, the sublaw-stuff or bylaw-stuff of the lesser working units" (Llewellyn and Hoebel 1961, p. 28).

Customary law can be geographically extensive and functionally decentralized (i.e., specialized), in contrast to the law of geographically defined states which tends to be functionally centralized and geographically constrained. Thus, customary law can have different sized jurisdictions for different functions. In some areas of law, economies may be considerably more limited than any state, so existing political entities are too large geographically (e.g., this applies for many aspects of criminal law [Benson 1998d]) or functionally (e.g., many aspects of domestic commerce may be most effectively governed by diverse trade associations rather than by the state [Benson 1995; Bernstein 1992]). In other areas of law, such as international commerce, some of these economies appear to be greater in geographic scope than any existing nation can encompass, although many also are narrow in functional scope, as international trade associations may be the most efficient source of rules and governance for many groups of traders.

II.2.g. An analogy for cyber law: the polycentric customary law of international trade

The vast majority of contract disputes in international trade are resolved through negotiation, perhaps with the help of mediators, but in the event that a voluntary solution cannot be achieved, virtually all international trade contracts have clauses that refer any dispute to arbitration (Lew 1978, p. 589; Berman and Dasser 1990, p. 33; Casella 1992, p. 1).[21] International arbitration is attractive relative to national courts for a number of procedural reasons,[22] but arbitration is also attractive because it provides a

---

[21] Indeed, even though some state-owned enterprises are prevented by various state laws from accepting arbitration, most are forced by trading partners to agree to private arbitration if they want to enter into international contracts (Böckstiegal 1984, pp. 17-19). This was even true of enterprises from the countries of Eastern Europe under communism, for whom arbitration was the "exclusive method of dispute settlement in business relations with other socialist countries and also the standard method in contracts with business partners in non-socialist countries" (Böckstiegal 1984, p. 15).

[22] The choice of an arbitration institution does involve a choice of procedural rules (Böckstiegal 1984, p. 23), and in this context, specialization by arbiters selected for their expertise and reputation (Ashenfelter 1987) means that arbitration typically is a faster, less formal, and less expensive procedure than litigation, in part because the parties do not have to provide as much information to the arbitrator to

means of supporting the contracting parties' choice of legal jurisdiction. Böckstiegal (1984, p. 23) points out that:

> When, in relation to arbitration, judges, arbitrators or authors speak of the 'applicable law' they do not always mean the same thing. The term is sometimes used rather indifferently with regard to three separate questions: What is the applicable procedural law? What are the applicable conflict of law rules? What is the applicable substantive law? All three of these questions have their definitive relevance in any kind of international commercial arbitration.

A large number of international trade associations have their own conflict resolution procedures, using arbitrators with special expertise in trade matters of concern to association members, including the community's "practices and usage" (i.e., customary rules) [over three decades ago, for example, Lazarus, et al. (1965) discussed more than 120 such tribunals], but other sources of arbitration are also available. The ICC's arbitration institution provides a substantial list of arbitrators with expertise in international commerce, for instance, as do nation-specific organizations around the world (e.g., the AAA, the Hungarian Chamber of Commerce) which are eager to provide arbitrators for international trade disputes. *Ad hoc* arbitration is also widely used (Böckstiegal 1984, p. 21), and in this regard, there is a rapidly growing market in private dispute resolution services provided by for-profit firms, at least in the United States (Ray 1992; Phalon 1992; Benson 1998d, pp. 113-16). Procedural rules vary across these institutions, some of which offer different procedures depending upon the wishes of the disputing parties, but more importantly, a contract can also specify the substantive law and conflict-of-law rules under which any dispute should be resolved. A contract might designate the contract law of the seat of arbitration or of some other national legal system be applied, for instance. The "usual way" of determining the relevant substantive law for international commercial arbitration, however, is to decide cases "exclusively on the interpretation of contracts and the relevance of trade usages so that very little depends on the question of the applicable [national] law" (Böckstiegal 1984, pp. 27, 23). Lew's (1978, p. 581) detailed analysis of available records (also see Trakman [1983]; Draetta, Lake, and Nanda [1992]) reveals that in principle, "[t]he answer to every dispute is to be found *prima facie* in the contract itself. What did the parties intend, what did they agree and what did they expect?" When an arbitrator cannot discover the parties' intent in the contract, however, the focus turns to consideration of what the

---

avoid an error in judgment as they would to a non-specialized judge or jury (Benson 1989, 1999b, 2000a). Another benefit arises when court time is allocated by waiting, since delay often can be devastating to a business and arbitration services can be purchased in a market or provided by a trade association without such delay (Benson 1989, 1995). Other potentially important procedural benefits include the facts that, if desired, privacy can be maintained (Bernstein 1992), and that arbitration is generally less "adversarial" than litigation, so it is more likely to allow continuation of mutually-beneficial repeated-dealing relationships (Benson 1989, 1995).

parties expected or should have expected, and in this regard, international arbitrators generally intentionally "denationalize" their awards, making them acceptable by showing their consistency with accepted traditional "practices and usage" (customary law) of the relevant business community (Lew 1978, pp. 582-85). Contracts might explicitly state that the practices and usages of a particular commercial community (e.g., a trade association, an informal group of traders who deal in the same products) should be applied, or this may simply be understood. Business custom provides the default rule, at any rate, as arbitrators apply the customary rules which are commonly recognized within the "private international law systems from which the parties come" (Lew 1978, p. 585), unless a particular nation's substantive law has been specified in the contract.[23]

In this regard, there are many different commercial communities, and their customary laws can be quite different. The wide variety of activities and relationships that exist in a modern world mean that many rules that are effective for one type of transaction or one group may not be effective for another. Consider the diamond traders discussed by Bernstein (1992) and the oil traders discussed by Trakman (1983), for instance. The products being traded within these commercial communities are very different, suggesting that very different contractual issues are likely to be relevant, but the trading communities are also very different. Diamond merchants share common ethnic and religious backgrounds, creating an environment of mutual understanding (e.g., of common trade practices and usage) and trust, for example, thus reducing the need for highly technical and specific contracts. On the other hand, oil traders display much greater ethnic and religious diversity as well as differences in motivations (a number of oil producing states have nationalized production, for instance, so political considerations can have major impacts of decision-making), possibly reducing the level of common understanding and undermining trust relationships, thus dictating much more specific and complex contracts. These two commercial groups are likely to share many of the same rules but they are also likely to be some important differences in traditions and practices. Indeed, imposition of the diamond merchants contractual rules and governance institutions on the oil traders would probably lead to much higher transactions costs for these traders, including more contract disputes, while imposition of the oil traders rules on the diamond merchants would add unnecessary complexity and costs to their contracting process. Thus, as Cooter (1994, p. 216) explains, decentralized lawmaking is actually desirable in the increas-

[23] The same is often true within domestic commerce as trade association mediators or arbitrators apply the association's own rules rather than those of the government of the territory within which the commercial transactions take place (Benson 1995). Indeed, historically, as trade evolved beyond small close-knit groups formed on the basis of trust and reputation, "legal systems" arise as a substitute for more informal arrangements, but these legal systems generally are not the product of nation-states (Benson 1989, 1998b, 1998c, 1995).

ingly complex international economy. The benefits of specialization that are anticipated in economic activity also can arise in the legal arena, making polycentric law more effective than monocentric (monopolized) law.

While various commentators have suggested analogies for the analysis of law in cyberspace, including the western frontier and feudalism (Yen 2002), the contention here is that the best analogy is international commerce (also see Johnson and Post [1996, pp. 1389-90], and Benson [2000c]). Since international commerce operates under a polycentric system of customary law, the hypothesis is that similar arrangements can provide effective governance in cyberspace.[24] To see if this hypothesis might be supportable, let us consider evidence of the evolving property rights, contracting arrangements (sources of trust and recourse), and customary law in cyberspace.

III. SPONTANEOUSLY EVOLVING INSTITUTIONS IN CYBERSPACE

The purpose of this section is to illustrate that the economic theory of spontaneously evolving institutions of property, contract, and customary law that was detailed in the preceding section can predict and explain developments in cyberspace. This will be done by applying the analysis in each of the subsections and sub-subsections of Section II in corresponding subsections and sub-subsections in Section III to explain recent and/or currently-arising developments in cyberspace. Cyberspace is a very dynamic environment, however, so any attempt to describe its institutions is going to be incomplete and inaccurate. Many of the innovations in the technology of property rights delineation and in institutional development discussed below may not even be relevant by the time this paper is published. The innovations may prove to be ineffective, for instance, or new even more effective developments may displace them.[25] Thus, the specific examples discussed below should be considered as simply an incomplete sampling of

---

[24] Many nation-states create rules and institutions in an attempt to influence international commerce, of course, but generally in order to limit beneficial trade through protectionist activities, rather than trying to enhance the ability of international traders to cooperate. Organizations of states such as the European Union and NAFTA also write rules and create institutions that deal with international trade, but most of these organizations of states have been formed to reduce their member states' powers to limit trade, rather than to provide the rules and institutions that traders themselves employ to facilitate cooperation.

[25] Some also may prove to be too costly because of actions taken by national governments. Some organizational efforts to reduce externalities might be seen as anti-competitive, for instance, and therefore violations of antitrust laws. Similarly, some governments, such as those in China, Saudi Arabia, and Iran, are attempting to undermine privatization efforts that limit these governments' abilities to monitor and filter Internet traffic. Some Western democracies do the same, in the name of national security (e.g., consider the FBI's Carnivore program which might be thwarted by some of the new developments discussed below).

the kinds of things the entrepreneurial participants in the dynamic and evolving socioeconomic and institutional landscape of cyberspace are trying or considering as they search for ways to internalize externalities and facilitate cooperative interaction.[26]

III.1. *Evolving Property Rights in Cyberspace*

The Internet originated in the 1960s with ARPANet, a decentralized computer-based communications network set up by the United States Department of Defense and the National Science Foundation. It has developed into a global technological system of computer networks with an estimated 900 million users.[27] In addition, the types of communication that flow over the Internet have exploded as new applications have been developed (e.g., Telnet, FTP, Gopher, LISTSERV, Usenet, EDI, E-mail, and so on). One of the most influential developments, the World Wide Web, is only 15 years old (it was created in 1990 by Tim Bernesers-Lee at CERN). Thus, many uses of the Internet are relatively recent, and are clearly evolving. For instance, Lucking-Riley and Spulber (2001, p. 55) report various estimates and projections of the magnitude of business-to-business (B2B) e-commerce. A 1999 estimate by the Gartner group put the level of B2B e-transactions at $90 billion, while a 2000 estimate by Jupiter Communications concluded that these transactions amounted to $336 billion (out of a total of $11.5 trillion). The accuracy of such estimates is questionable because e-commerce is a global phenomenon and much of it probably is not reported to any government authority. Nonetheless, the estimates reveal that B2B e-commerce is substantial, and probably growing. Business-to-consumer (B2C) trading (retailing) is highly visible, and this component of e-commerce has attracted much more attention from governments and the media (and probably from academia). Yet, the Gartner Group's estimate of B2C e-transactions was $16.7 billion in 1999, including both retail sales and brokerage fees for online financial transactions, compared to the $90 billion in B2B transactions (Lucking-Riley and Spulber 2001, pp. 55-56). Growth is also evident in B2C e-trading as Bakos (2001, p. 69) reports that such trade reached $45 billion for 2000, for instance, and Tedeschi (2004b) cites a Shop.org report that estimated 2003 retail sales for U.S. retailers (not including sales of travel services) on line to be $72 billion (5.4 percent of

---

[26] In this regard, note that in an effort to be relatively timely, many of the specific examples of technology are drawn from recent popular press stories. No effort has been made to verify the claims made by reporters, as these developments are intended to simply illustrate the kinds of things that are being tried and considered, rather than to claim that they will actually provide the lasting solutions to externality and credibility problems that arise in cyberspace. In other words, they demonstrate the predicted search for solutions, but not necessarily the discovery of long-term solutions.

[27] This figure is from ClickZ Stats: the Web Worldwide (http://www.clickz.com/stats/web_worldwide/), February 2005, citing the Computer Industry Almanac.

total U.S. retail receipts for the year). B2B e-commerce is likely to continue to be larger and less apparent than B2C e-commerce for the reasons suggested below. Before turning to these developments, and their institutional underpinnings, let us briefly describe the cyberspace itself. As Crews (2003, p. 1) notes, "the Internet wasn't originally designed to be a mass commercial and consumer medium that it is today. If one were to design a commercial network today from the bottom up, it would probably" look very different. The technological configuration of the Internet has raised the cost of some forms of institutional development, but it does not preclude all such innovations.

The Internet is not simply a technological system. It also is an implicit agreement, or perhaps more accurately, "a loosely-organized international collaboration of autonomous, interconnected networks" (Internet Engineering Taskforce, RFC 2026) that allows bits to flow among computers using a particular language or "protocol." It may be useful to consider the result in terms of layers (Crawford 2005). At the bottom there is a physical infrastructure (e.g., cable, satellites, routers, DSL, WiFi). Above that there is a logical "protocol" layer (TCP/IP, HTTP). The next layer up consists of applications (e.g., browsers, e-mail, VoIP), and at the top there is a content layer (text, music, speech, images). The bottom two layers separate mechanical transportation of bits from the protocol. The protocol divides the bits into packages that can be reassembled after transport. Internet hosts (computers that store data that is included in the Internet) have domain names that are translated into IP addresses. The TCP/IP protocol allows very heterogeneous infrastructure networks to transport data from one IP address to another. Routers are computers that link the component networks. They contain routing tables, and they mechanically look up parts of the address in a data package in order to send the package on to another router that is closer to the final destination. Thus, the technical part of data transmission is very mechanical.

The "free access" characteristic of the Internet applies to the two lower levels. For instance, the logical layer is allowed to run across all (or at least enough of) of the infrastructure layer to reach any IP address in the network of networks. That is, the installers of cables and wireless connections and routers are expected to mechanically and indiscriminately allow TCP/IP to work. Furthermore, the logical layer presumably does not discriminate between applications, so anyone can send any kind of information over the Internet. The early tradition of free access to the Internet also presumed that, like the infrastructure and the logical layers, the applications layer would not discriminate against particular kinds of content. Thus, free access to the applications layer presumed that everyone could send anything to everyone else. Importantly, however, the top two layers are not actually global at all. They are developed and installed by individuals and organizations such as firms and governments, in order to use the lower layers. If developers of an application choose not to freely allow any and all content

to be deposited into an address or addresses in their part of the network of networks, the Internet can still function as a global communications system. Everyone is still free to send whatever content they want to send, but some individuals or organizations might refuse to accept delivery of some contents. Therefore, even though the Internet does require free access to the bottom two layers to function as a global communications system, given its current technological configuration, it does not require free access into all IP addresses by all types of applications of all types of content.

III.1.a. Externalities in cyberspace and the evolution of property rights

Recall that externalities arise when two or more parties want to use the same scarce resource for conflicting purposes. Externalities are an inevitable result of common access to the Internet, a system comprised of scarce resources. Indeed, as more and more uses of these resources are discovered, and more and more people enter the commons, it is not at all surprising to find increasingly costly externality problems arising. People want to use this resource to lower their costs of buying and/or selling goods and services, to transmit and store data, to communicate with friends and colleagues, and for entertainment, but some people also get pleasure from causing harm to others. The net is sufficiently large to support tremendous numbers of activities, but crowding (i.e., conflicting uses) is becoming an increasingly significant problem.

An obvious example is spam. The "bulk-mailing" of large numbers of e-mails advertising Viagra and other prescription drugs, pornography sites, and a huge array of other goods and services, crowd the Internet, slowing traffic, and more importantly, they crowd e-mail inboxes, raising the cost to recipients who do not want to deal with "junk mail." As Stross (2005, p. 1) puts it, recipients of e-mail bear the costs: "[i]t is nominally free, of course but it arrives in polluted form. Cleaning out the stuff once it reaches our inbox, or our Internet service provider's, is irritating beyond words . . . . This muck . . . is a bane of modern life." A December 2004 survey suggested that Internet users were spending an average of 10 working days per year dealing with spam, and at least some industry analysts estimate that the 2005 cost of spam to business due to lost productivity and additional network maintenance costs will be around $50 billion for the year (Zeller 2005, p. 3). This is not surprising because spam accounts for an estimated 80 percent of all e-mail traffic (Zeller 2005, p. 2). Spammers go beyond just filling in-boxes and slowing Internet traffic. They also "commandeer personal computers as zombie spam transmitters," by using a virus to install programs on other people's computers so they will unknowingly act as free relays while simultaneously making it more difficult for others to break through the "cloak of anonymity" to find out the true source of the spam (Zeller 2005, p. 5). Some estimates suggest that 50 percent or more of the

spam flowing through the Internet is handled by "hijacked machines" (Zeller 2005, p. 5).

Most spam involves efforts to convince consumers to purchase various products or services. While spammers probably consider such bulk mailings to be legitimate business practice, as an effective and low cost form of advertising which clearly results in enough sales to cover their costs, they obviously do not consider the external costs they generate for other web users, and especially for recipients who do not want to receive the spam. After all, despite laws passed by various governments that make some of these activities illegal, they are not likely to be held liable for these costs,[28] which, as suggested above, are becoming very high. Not all spam consists of annoying advertising, however. Phishing is a hacker-coined term which originally referred to the act of stealing AOL accounts, but phishers are expanding their targets. June 2004 saw 492 different mass e-mailings attempting to convince customers of Citibank that they should provide confidential financial information to them (e.g., account numbers), along with 285 mass mailings to eBay users (Gallagher 2004). Since then, similar mailings have falsely represented themselves as coming from many other financial institutions. Spam of this kind does more than simply crowd e-mail inboxes. It is an intentional effort to harm (impose costs on) others. Viruses and worms are similar in this regard.

A 2002 survey conducted by the FBI and the Computer Security Institute reports that virus attacks caused losses of $49.97 million over a twelve month period during 2001 and 2002, and other surveys put the cost much higher, ranging up to $12 billion (Krim 2002, p. 3). Estimates suggest that the "love bug" virus of 2000 cost the global economy $2.62 billion alone (Powell 2005, p. 1). The Coordination Center of the Software Engineering Institute reports that attacks on business and government computers doubled in 2001 over the 2000 level, with an additional 26,829 attacks during the first quarter of 2002 (another doubling over the previous year) (Krim

---

[28] The U.S. "Can Spam" Act went into effect in January, 2004, for example, but spam has actually increased considerably since then. Zeller (2005, p. 2) suggests that spam accounted for about 50 to 60 percent of the e-mail traffic before the law was instituted, for instance, but that afterward the percentage jumped to 80 percent. This increase in spam actually was predicted by critics of the bill, since the law does not outlaw spam. Instead, it gives bulk e-mailers the right to send spam as long as they follow certain rules. Critics note that before the Act was passed, the legal status of spam was uncertain, but after the Act, it became clear that spam was legal in the eyes of the U.S. government (Stross 2005, p. 2). More importantly, even if a country like the U.S. were to outlaw spam, spammers would simply move their operations into other national jurisdictions, already a common practice (Zeller 2005), in part because some governments, like China, promise not to interfere with the spamming operations. Indeed, such countries see spam operations as desirable because they generate local employment and stimulate local economic activity, so local laws actually protect websites advertised through spam broadcasts (Zeller 2004, p. 5). Thus, as Susan Getgood, senior vice president of U.S. marketing at SurfControl LLC, noted "Clearly the content hasn't changed at all [as a result of the Can Spam Act]. . . . Spammers didn't waste a minute to make it look like they were complying" (quoted in Krim [2004, p. 2]).

2002, p. 3). Underreporting of such attacks is likely since many businesses and government agencies do not want to reveal their vulnerability. 2003 has been called "the year of the worm" by some computer experts (Thompson 2004, p. 2).[29] The year started with the Slammer worm infecting almost 75,000 servers in just ten minutes in January. Among the most spectacular of the infestations that followed was the Blaster worm which infected hundreds of thousands of computers, and Sobig.F, which, at one point, resulted in one out of every 17 e-mail messages being transmitted carrying a copy of Sobig.F. One estimate put the global cost to business of 2003's computer viruses at $55 billion (Crews 2004, p. 2). Mydoom.A followed in January of 2004, however, and spread even faster than Sobig.F (at its peak, one in five e-mail messages contained a copy of Mydoom.A). The worldwide cost of these worms, for cleanup and lost productivity, clearly is many billions of dollars. The people who intentionally create such attacks obviously do not consider the tremendous costs they impose on others. Perhaps more accurately, they do not consider these consequences to be costs at all, because attackers apparently get personal pleasure from their actions. Such attacks often have a specific target against which the programmer has a personal grudge, but they impose tremendous costs on millions of innocent bystanders in the process. The Blaster worm was developed to attack Microsoft, for instance (it exploited flaws in Windows and attempted to bombard a Microsoft website with data), while Mydoom.A reprogrammed computers to attack the website of another software firm, SCO.

Like viruses, Adware and Spyware are programs that are actually installed on a person's computer by someone else, but their purpose is not to disrupt or destroy. Those who deploy these programs want users to continue their Internet activities, as they are secretly monitored. These programs create very real external costs, however. For one thing, they tend to overburden PC's, making them operate slower as they respond to the processes prompted by hidden programs. In addition, Adware monitors online activity in order to display pop-up ads to users who show interest in particular kinds of services or products. Large numbers of popup ads create considerable cost themselves, as they typically must be deleted before the user can continue whatever he or she is doing. The extent of the spread of Adware is difficult to determine, because users generally do not know that it

[29] Technically, viruses and worms are different (Thompson 2004, p. 4). A virus that arrives on a computer cannot start itself. Therefore, a person has to be fooled into starting it. Thus, a virus typically arrives in some sort of disguise, perhaps made to look like an MP3 file, when in fact it is a program which, when activated, reprograms the computer to do something new. It typically has an unusual suffix because it is not an MP3 file. On the other hand, worms do not require a person to activate them. They also generally do not alter or destroy data on the computer. Instead, a worm rapidly multiplies, often by sending copies to every address in each victim's address book, generating so much traffic that Internet servers cannot handle the flow. The distinction between worms and viruses is breaking down, however, as a worm can also carry a virus that is deposited in each computer while it is e-mailing itself to new targets (Thompson 2004, p. 4).

has been installed. In all likelihood, however, such programs are becoming ubiquitous. For instance, one Adware vendor, Claria, reportedly had 29 million users running its Adware products on their computers in 2004 (O'Brien and Hansell 2004, p. 1), up from 1.5 million in 2000. The company's 2003 revenues from firms using popup ads were 90.5 million, which produced a profit of $35 million.

Spyware is even more "parasitic." These programs are installed on computer hard drives by piggybacking on software programs that people intentionally download, or by being sent through security gaps in web browsers when users visit certain websites. Spyware automatically records and discloses everything the person does on line. It can be used to learn passwords, as well as information facilitating identity theft, such as account and credit card numbers, social security numbers, and other information a user keys into his or her computer. Indeed, identity theft has become an issue of considerable concern, whether the thefts are accomplished through low-tech activities like phishing, or high tech activities such as distributing spyware and intentionally hacking into targeted databases containing information such as credit card and social security numbers.

The Federal Trade Commission reports that an estimated 27.3 million Americans suffered the theft of their identities over the 5-year period from April 1998 to April 2003. Over a third of these thefts occurred during the final 12 months of the period (O'Brien 2004, p. 2). Internet-based identity thefts represented a substantial portion of these identity thefts and resulted in tremendous costs. In the final 12 months of the period, financial institutions and businesses lost and estimated $48 billion to these identity thefts. In addition, theft victims lost an estimated $5 billion in out-of-pocket expenses required to reestablish their financial identities, but the psychological and time costs were probably much higher. Clearly, the Internet has reduced the relative certainty of many non-cyber asset property rights[30] because those assets are used in online transactions.

III.1.b. Factors influencing the evolution of property rights in cyberspace

The economic model of property rights suggests that investments in the creation and security of private property rights should increase as a consequence of rising external costs in cyberspace. This is clearly occurring. Laws that threaten punishment for trespassers, thieves and others who violate claims represent one way to establish property rights. However, this deterrence approach requires that punishment threats be credible. While

[30] For example, these include individual credit and payment arrangements, but not IP addresses and computers.

some governments (e.g., U.S. Can Spam Act[31]) make such threats, violators can locate outside those governments' jurisdictions and cloak their identities. This renders such laws ineffective in creating anything close to secure rights (this issue is discussed at greater length in the concluding section).

Alternatively, watching, walling, and wariness may protect individual pieces of property (Sherman 1983). New technological developments lower the costs to sort (watching) and exclude (walling) unwanted data from IP addresses and personal computers, and to require authentication from senders (wariness). These developments work much like the invention of barbed wire, which lowered the costs to define property boundaries and to exclude other people's cattle from land in the American West (Anderson and Hill 1975). Other important factors include innovations in reputation establishment methods in cyber commerce; in contractual arrangements; and in other organizational methods for institutionalizing incentives, such as developments in dispute resolution techniques. This subsubsection focuses on technological innovations that allow individuals to exclude others from access to their addresses and computers. Most of the institutional changes will be considered below.[32]

Access to a person's computer and/or IP address can be limited if the benefits of doing so in the minds of the computer owner exceed the costs. A growing numbers of technological innovations prevent entry of unwanted data by screening and filtering content. Efforts by parents to monitor and restrict their children's access to content exemplifies one such considerable development. For instance, some parents express increasing concern regarding with whom their children communicate via online instant messaging and chat rooms.[33] Physical methods of monitoring such activities prove very costly and are often ineffective, but technological methods of monitoring are increasingly available.[34]

Many paid online services, such as AOL and MSN, offer instant messaging and chat rooms, and these services offer parents control tools that can constrain children's use of these options. AOL subscribers can create restricted accounts that limit chat and instant messaging opportunities, as well as web browsing and e-mail opportunities that parents consider inappropriate. For example, the AOL "Kids Only" default category blocks all instant messaging, while the "Young Teen" category does not allow the exchange of images, files, voice, or videos. A "Mature Teen" category is also available. It limits chat and web browsing but not instant messaging.

---

31 See note 28 for discussion.

32 Note, however, that institutional and technological innovations cannot be completely separated, as some of technological innovations are being developed and implemented through evolving contractual associations.

33 These online tools allow Internet users to communicate virtually instantaneously.

34 The following discussion of parental options for limiting access to their children's computers draws from Magid (2004).

Individual parents can customize all of these categories, even controlling when and how long their children can use AOL. Similarly, MSN parental control categories include Teen, Preteen, and Young Child, with varying degrees of parental limitations and customized settings. Parents can block all use of Microsoft Messenger, or restrict access to people on a specified contact list. MSN also allows parents to block file downloads, and can e-mail parents reports about their children's online activities.

Many other Internet access providers (ISPs) have similar options for parents. Of course, children can use browsers not provided by their ISP to circumvent such controls. Even though AOL and other providers offer some options for controlling these activities as well, individual parents can do more if they remain concerned. In particular, parents can purchase and install a number of parental control software packages. For instance, CyberPatrol can block instant messaging, and its Chat Guard feature filters specified strings of characters such as objectionable words, phone numbers, and names. Cybersitter fulfills similar functions, and also allows parents to record the text of instant messaging exchanges. It also cannot be disabled without a password (unlike recording features in Yahoo and MSN messaging software, for instance). Spectorsoft records all e-mail, instant-messaging, and chat conversations and keeps parents informed of their children's online activities by e-mail. Of course, children can always find ways around constraints if they try hard enough, whether the constraints are imposed in geographic or cyber space. Still, if parents can limit significantly access to a child's computer address, even when children do not want such limits, this suggests that individuals who do not want to receive certain kinds of content themselves can also dramatically limit access to their addresses.

Software to filter incoming messages is developing rapidly, and both individuals (e.g., Anti-Virus programs purchased from Norton, McAfee, and other firms) and contractual arrangements use them. Indeed, the same ISPs that offer parents methods to control access to their children's computers, also offer various methods to exclude spammers, Adware popups, viruses, etc. For instance, AOL reported that it blocked 500 billion spam e-mails intended for their subscribers during 2003, or roughly 15,000 spams per AOL member (AOL is blocking spam. *Washington Post,* January 5, 2004, p. E02). Casual observation of television advertising by such ISPs suggests that these filtering features are very attractive to consumers; therefore, they are becoming an increasingly important focus of competition.

E-mail services like Yahoo Mail, Hotmail, Eudora, and Outlook also provide spam filters. Hotmail reports catching 3.2 billion spam messages per day (Stross 2005, p. 1), while Eudora version 6.0's filter is said to have stopped 97 percent of the user-directed spam after it was released in September of 2003 (Hafner 2004a, p. 2). Such filters are not perfect, of course. As Barzel (1989, p. 5) stresses, property rights are never likely to be completely and perfectly delineated, given the costs of doing so relative to the

benefits. Indeed, the Eudora filter is less accurate than it was when it was first released because spammers have altered their messages to avoid detection. This filter, probably one of the most effective, uses a method called Bayesian scoring (Hafner 2004a, p. 2). It looks at words in an e-mail message and assigns each word a probability of being part of a spam. It then calculates a cumulative probability to estimate the total likelihood that the message is spam. The user can set the level of probability with which he or she is comfortable, and the program will divert all messages with a probability above the chosen level.

Like other filters, the Bayesian scoring method does not filter all spam, and it can divert messages that are not spam. Over time, spammers adjust by filling messages with low probability words. For instance, a spam message may have a large number of what appear to be random words at the bottom, or the ad might be surrounded by poetry. As a result, the large numbers of words the program does not consider likely spam content swamps the likely spam words. Thus, the filters have to change over time as the spam does.

One way to adjust the filter is to allow individual users to do so. If spam gets through, the user can designate it as such, and the filter will block future messages of that kind. Filter programs can also divert messages to a "junk" mailbox that the user can check occasionally to make sure that desired messages are not being diverted. The user can tag desired messages as valid, so they will not be diverted in the future. In addition, filter providers continue to experiment in an effort to refine the filtering process, just as spammers experiment to undermine it. As a result, this filtering software has become very effective at catching a large majority of spam before it ever gets into e-mailboxes (Zeller 2005, p. 4).

Others ways to exclude unwanted e-mail are also available. For instance, some e-mail account providers, including Earthlink and Mailblocks, use "Challenge-response" systems (Krim 2004, p. 1). This technology sends an automatic challenge from an e-mail recipient to the sender. The challenge asks for a response of some sort (e.g., supply a generated password, or answer some question) to verify that the sender is a real person. Since most spam is sent in bulk by computers, it will not reply to the challenge. As a result, the system will not verify the message, and will exclude it. These systems have proven very effective at eliminating spam (Crews 2003, p. 3).

Growing consumer demands for security and the competition to offer more secure e-mail and browsing systems no doubt, in part, contribute to the increased availability of such technology. In 2004, Microsoft reported that it was experimenting with this technology as a possible added protection for the more than 100,000 million users of its Hotmail and MSN mail servers (Krim 2004, p. 1). Also, Microsoft purchased a number of other software firms specializing in security software (e.g., Sybari Software, a specialist in programs to protect business computers networks from viruses,

worms, and other threats, and GeCAD Software, an anti-virus firm [Associated Press, 2005, p. 1]) and reportedly spends about $2 billion per year on computer security research and development (Markoff 2005, p. 1). Further, the increased use of the Firefox browser apparently induced Microsoft to rush Explorer 7.0 to the market. Explorer 7.0 offers new anti-spyware technology and improved anti-virus features, (Markoff 2005, p. 1).

Firms also offer E-mail checking services. For example, a PermitMail user provides the service with a list of permitted sender addresses, and if an incoming e-mail is not on the list, it is held.[35] The sender receives a message indicating that the mail is being held, and that he or she can fill out a form asking to be put on the user's permit list. The sender must provide a name, and can write a message explaining why he or she should be added to the list.

Markets are also rapidly developing other kinds of identification and authentication processes. For instance, businesses like banks, brokerage firms, travel sites, and other online merchants, have started requiring authentication in an effort to distinguish between their legitimate customers and spam, and to protect their transactions from potential identity theft. One way to do this is to issue customers a "hardware token," a small device (small enough to attach to a key chain) that displays a six digit number which changes once a minute. When customers access their accounts, they type in their user names and passwords and also enter the numbers displayed on their tokens. Several banks in Europe and Australia already use tokens, and mandate them for their online customers. A number of U.S. banks have issued tokens to their corporate customers and to their own employees and are considering the technology for their retail banking services. U.S. Bancorp plans to experimentally test the system, while E*Trade Financial offers this devise as an option to its customers this year (Kingson 2004, p. 1). Apparently, virtually every bank in the U.S. is seriously considering the technology.

Other technologies with similar authentication applications include smart cards (plastic cards with embedded microprocessor chips); biometrics (identification of people by their voice, physical characteristics, or fingerprints); and shared secrets (customers must answer a question that, at least in theory, only they know the answer to). In a similar vein, several ISPs have begun to "stamp" customers' outgoing messages with a "digital signature." The digital signature includes customers' domain names, and uses strong cryptography to prevent signature alteration or counterfeiting. Recipients can then look for such a signature on incoming mail using DomainKeys software. An organization of ISPs supports this identification process in an agreement to prevent spammers from operating through their services. This organizational development is discussed below.

---

[35] Permit mail is a service of Internet Light and Power, Inc. See http://www.ipermitmail.com/. Similar services are offered by other firms.

In yet another option, firms also consider pricing e-mails. For example, an Australian firm, CashRamSpam.com, already offers such a payment system (Krim 2004, p. 2). Customers set a fee for incoming e-mail from unknown sources. The firm collects the fee, keeps 10 percent, and passes the rest to the customer. Refusal to pay means that the e-mail will be deleted, so a high fee is expected to eliminate all unwanted e-mail, while a lower fee might provide the customer with a source of income if paid by some advertisers. Whether this system will catch on is unclear, but Microsoft Research is considering an alternative involving payment in time rather than for money (Stross 2005, p. 3). The Microsoft system forces a sender to use the sending computer's resources to solve a computational puzzle. Since each puzzle is unique to a message it creates a "stamp." "The puzzle uses an intricate design involving the way a computer gains access to memory and resists a quick solution by speedy processors, requiring about ten seconds. It is not so long that you'd notice it for the occasional outgoing message, but if you have eight million Viagra messages queued up, good luck in getting each one stamped." (Stross 2005, p. 3). An individual can then set his or her e-mail program to filter all incoming messages from unfamiliar senders according to their digital stamp (proof of having completed an assigned problem).

In addition, the free and easy sharing of files can reduce significantly the value of so-called intellectual property rights, particularly to music and films. As a result, firms in these industries substantially pressure governments to protect such rights, lobbying legislatures and pursuing court cases. Yet, if a solution exists, technological innovations rather than government will likely solve these intellectual property rights issues. This should not be surprising. After all, as Cox and Crews (2005, p. 2) explain, "The online intellectual property problems faced by copyright owners do not exist in isolation. They are rooted in the Internet's (perceived) anonymity [and free access] - but so are spam, cybersecurity, spyware, and authentication of transactions. Indeed, to a large extent, we don't have an intellectual property problem; we have an anonymity/authentication[/free-access] problem."

In this context, "digital rights management" (DRM) technology, in combination with pricing schemes, now provides a method of protecting copyrights to digital content distributed on line. A DRM system involves a technological container format for "packaging" the product (Einhorn and Rosenblatt 2005). For instance, a music album may contain the music and tract titles, as well as a set of rules that the software and hardware music players must follow in order to play the material. This set of rules can include a price per play of each song, or for a set number of plays. Alternatively, a consumer might subscribe to a service, perhaps for rentals or demos, through which chosen song titles remain valid as long as the subscription is maintained.

More significantly, DRM address issues of redistribution (Cox and Crews 2005, p. 2). Sellers can contract with consumers to meet whatever

preferences they may have regarding transferability, but consumers must pay for the right to reproduce. For example, encrypted controls can limit how users make and distribute copies of the digital files (Einhorn and Rosenblatt 2005, p. 2). As with the pricing arrangements suggested above, DRM offers a tremendous number of options regarding copying. For instance, DRM can charge a different price for a downloadable/reproducible movie or music album than for the same film or music that cannot be reproduced. Formats supporting DRM include Windows Media and Advanced Audio Coding (AAC).[36]

The DRM technology is certainly not foolproof. Programmers are already finding ways to beat it. However, reasonable prices create incentives for a large portion of relevant Internet users to purchase access rather than invest in methods to beat the DRM technology. For instance, Apple Computer introduced its Internet Music Store in April, 2003, using AAC compression technology and a "light-handed" DRM system called FairPlay (Einhorn and Rosenblatt 2005, p. 4). Music buyers can transfer songs to Apple iPod players, and can burn unlimited numbers of CDs, but the technology limits downloads to three other hard drives. The next version of Music Store will have a number of additional features. In addition, RealNetwork's Rhapsody allows "all-you-can-eat" streaming for a $9.95 per month fee, but charges 79 cents for individual burns. Other copyright-respecting music services including MusicMatch and Napster. MusicMatch offers 99 cent downloads, as well as a subscription service that permits on-demand steaming and playlist sharing with friends. Non-subscriber friends can play the first 20 tracks on each playlist they receive, but only up to three times before they have to pay for individual downloads or subscribe. The re-launched Napster charges 99 cents per downloadable track or burnable individual song, as well as other subscription services (Einhorn and Rosenblatt 2005, pp. 4-7).

Thus, DRM provides a way to contract between suppliers of digital music and film and their consumers. Such contracts give consumers specific use rights while raising the cost of uncompensated redistribution. In addition, norms regarding "fair use" (i.e., "reasonable usage") of downloading and transferring are evolving through the competitive market process. Different firms offer various alternatives. Those that survive and prosper will be the firms that meet the expectations of sufficient groups of consumers while compensating the producers of music and movies for their products.

---

[36] See entries in TechEncyclopedia on TechWeb for discussion: http://www.techweb.com/encyclopedia/defineterm.jhtml?term=drm; http://www.techweb.com/encyclopedia/defineterm.jhtml?term=AAC; http://www.techweb.com/encyclopedia/defineterm.jhtml?term=Windows+Media.

III.1.c. The security of property rights in cyberspace

In the past, individuals had little choice but to accept messages from all sources, inspect them, and reject data only after it was received. By contrast, the current Internet appears to be transforming. In this modern-day system, address holders can affirmatively decide to accept messages only from identifiable sources. In today's environment each individual can decide, or contract with someone else (e.g., an ISP, an employer, a specialist in screening e-mail) to decide when and with whom a connection to an address will be allowed. Clearly, the trend is to require verification from others. Individuals and organizations consider the costs to their time and of spam, viruses, adware, spyware, and other computer threats higher than the benefits to unlocked in-boxes. As a result, they allow a message to enter only if the sender is identified as trustworthy.

Developments in financial transactions also reflect efforts to secure information sent from one party to another. For instance, customers of several credit card providers (e.g., Citibank, MBNA, Discover[37]) can request a temporary account number for use in online purchases. While these numbers are linked to the customer's account, they expire after one use or after reaching a customer-specified spending limit (Bayot 2004, p. 1). Depending on the provider, customers can obtain these numbers over the phone, from the company's website, or by downloading number-generating software that works either upon request or when the software detects a cardholder is at an online retailer's checkout page. Furthermore, the temporary number will be useless to a hacker who attempts to steal numbers from a company database that stores credit card numbers. While these temporary numbers limit the potential for identity theft and hacking with spyware, they also can prevent retailers from renewing purchases such as magazine subscriptions without customer notice.

Identification and authentication provide means to privatize rights. Still, while they exclude others from an individual's in-box, they simultaneously raise concerns about another right: the right to privacy and anonymity. Encryption technology may provide a potential solution to this dilemma. Encryption ensures that information accessible in computers or sent over the Internet cannot be read without an encryption key. Encryption and public keys code information and verify sources while protecting an individual's offline identity.[38] In real space, this process sounds cumbersome and even infeasible, but in cyberspace it is virtually costless and very fast.

---

[37] See http://www.citibank.com/us/cards/tour/cb/shp_van.htm, http://www.discovercard.com/deskshop, and http://www.mbnashopsafe.com.

[38] See Friedman (2001) for detailed discussion of encryption and public key technology as a means of maintaining privacy while simultaneously providing for authentication.

The encryption process involves a pair of "keys" (mathematical algorithms), which are inverses of each other. The sender retains a private key, and others decode the message with a "public key." However, the "public" key does not imply that anyone and everyone have it. Instead, it is a decoding key that the private key holder provides to select other individuals. Indeed, only a person who has a public key is able to decode the message, so the sender must provide the intended recipient with the public key.[39] The transmission remains private since it is coded and verifies a unique sender's identity since only the person who has the private key can send the message, but the technology cannot verify the credibility of the sender. The recipient may not know who sent a message with an attached public key. Thus, an assurance problem may stand in the way of such "verifiable" transactions without some additional reason to trust the sender. Still, the market may also solve this other types of assurance problems.

III.2. *Assurance Problems and Sources of Credibility in Cyber Contracting*

Securing property rights is not enough, if one is to enjoy the potential benefits of low-cost communication in cyberspace. Individuals must have others with whom to communicate, and for many kinds of communication, this requires the development of trust and/or recourse. This is one reason for the relatively rapid growth in and relatively large size of B2B trading on the Internet (as compared to B2C trading, discussed above). Much of e-commerce probably involves B2B trading between firms with a history of offline repeat-dealings. These trades simply move online in order to lower transactions costs. They already have trust relationships. Others may not have previously established trust relationships but may belong to the same business community (e.g., trade association). Alternatively, firms may have pre-existing offline reputations that they take with them when they move online. Of course, some new businesses develop online, but these new firms must establish trust relationships or have access to recourse if they are going to flourish in B2B e-commerce.

Similar points apply to online B2C trading. Many online retailers have pre-existing offline reputations. These firms prove relatively more likely to establish profitable online businesses and do so relatively quickly. New retailers who operate only online face higher costs as they attempt to establish credibility. For instance, a 2004 survey of online retailers found that 93 percent of firms with online websites as well as offline catalog business reported a profit, as did 85 percent of traditional retailers (i.e., firms with real-space retail stores) with established websites (Tedeschi 2004b, p. 1). On the other hand, only 67 percent of retailers who sell exclusively

[39] Technically, the public key can be sent along with the message to the recipient.

through the Internet reported profits, and many of them suffered substantial losses the year before. As a group, they "still struggl[e] to achieve profitability." (Tedeschi 2004b, p. 1). Survival of such firms clearly requires that they solve the assurance problem. The same holds true for individuals who may want to engage in many other kinds of online interactions.

III.2.a. Building trust in cyberspace

When two strangers initiate an interaction, such as a trade, the typical process involves several small steps rather than an immediate large commitment. The two strangers start by attempting to gather information about the potential partner, and if nothing negative is discovered, they make a small commitment (e.g. a small trade). If that proves successful, additional transactions occur and can get larger, but substantial commitments will not occur until a strong trust relationship develops (e.g., see McMillan and Woodruff [1998]). Investments in establishing such relationships can take some time, and delayed payoffs are uncertain. The relatively weak incentives to build trust relationships suggests these cooperative interactions may be slow to emerge.

As noted above, while repeated dealing produces trust, other investments in reputation building can also achieve results. Firms that exclusively trade on the Internet are not in a position to invest in some of traditional firms' non-salvageable assets, such as elaborate store fronts (elaborate web pages might be a substitute, but they are not likely to be seen as large investments). Still, advertising remains an option to these firms. Online advertising is ubiquitous, of course. Spam is cheap but probably adds nothing positive to firms' reputations. Indeed, many Internet users probably view spam advertising as a signal of unreliability, but many online services survive on advertising revenues, much like television networks. Since users probably perceive such advertising as very inexpensive to firms, it may not be an effective method of reputation building.

Even though Internet advertising appears inexpensive, high levels of advertising could prove effective, and it appears many firms believe this. For instance, AOL paid $435 million for Advertising.com, a firm that sells ads on a network of websites, and Digitas bought Modem Media in a $200 million stock transaction (Ives 2004, p. C2). TNS Media Intelligence, an organization that tracks advertising spending, reports that firms spent $5.6 billion on online advertising during the first nine months of 2004 (Ives 2004, p. C2). This represents a 25.8 percent increase over the same period in 2003. However, despite the increase, online advertising constitutes only about 5.5 percent of total advertising spending ($102.5 billion) for the 2004 period. This "stubbornly small portion" of advertising spending approximately equals the amount spent on radio advertising (Ives 2004, p. C2). It suggests these investments probably signal firm reliability less effectively than offline advertising (e.g., celebrity endorsements, television ads during

prime time, etc.). As a consequence, many cyber-firms also resort to physical universe advertisements. Firms trying to establish themselves in Internet markets now commonly advertise in television and magazines. So, cyberspace advertising builds firm reputation less effectively than real space advertising because it spreads information very rapidly and cheaply and thus fails to signal a substantial investment in firm reputation to consumers. Consumers interpret firms' willingness to invest substantially in advertising as a proxy for reliability and quality.

In an additional development, online specialists now supply information regarding firms, and some companies seek endorsements from these sources. For instance, some firms send free products to prominent reviewers on sites like Epinions.com and Slashdot.org, even though these reviewers carry no official status or credentials. These forums have developed self-enforcing reputation measures for reviewers and contributors. Some, such as Epinions, actually pay small fees to reviewers based on reader reactions to the reviews (e.g., do they click over to the company's web page or go on to read other reviews?) (Thompson 2003, p. 2). Epinions also allows users to comment on individual reviews as well as products and enlists teams of experienced site-users to monitor and detect any producer efforts to "plug" a product. Similarly, Slashdot gives each user a "karma" rating based on how frequently a person contributes and how other users rate those contributions. The karma rating determines user access to some site privileges.

EBay has developed perhaps the most widely cited and studied online reputation mechanism.[40] EBay acts as an online intermediary through which sellers post auctions and buyers bid. It obtains revenues from seller fees following a successful auction. EBay has developed an innovative feedback mechanism that facilitates reputation formation and reputation-based sanctions. Following an auction, the buyer and the seller can give each other a "grade" (+1, 0, -1) and provide a comment. EBay then displays several aggregations of both seller and buyer grades. This aggregation is an overall rating that adds the grades from the person's entire eBay history; the percent positive; the date the person first registered on the site; a summary of recent reviews; and the entire feedback record.

Reputation has become very valuable to both buyers and sellers. Sellers with good reputations obtain higher prices, can expand their sales, and survive to sell again and again (e.g., Dewally and Ederington (2003); Resnick, et al. (2003); Cabral and Hortacsu (2004)). A large portion of eBay's traders are repeat players. In fact, it has been estimated that around 500,000

---

[40] A number of empirical studies of eBay's reputation mechanism have been performed. An incomplete sampling includes Bajari and Hortacsu (2003), Cabral and Hortacsu (2004), Dewally and Ederington (2003), Houser and Wooders (2003), Jin and Kato (2004), Lucking-Reiley, et al. (2000), McDonald and Slawson (2002), Melnik and Alm (2002), Resnick and Zeckhauser (2002), and Resnick, et al. (2002, 2003).

people make full or part-time livings through online auction sales (Murphy 2004, p. 1).[41] Increasingly, sellers with a recognized reputation status also act as agents for less frequent traders. In 2003, an estimated 30,000 traders used eBay's trading assistants program, and several new "store-front" firms emerged as consignment operations "specifically to take in merchandise to sell on eBay" (Alexander 2003, p. 1). These firms compete for business on the basis of price. AuctionBytes, an Internet newsletter, compares prices among consignment shops and provides sellers with centralized information about consignment alternatives. Buyer reputation also matters, and eBay also posts feedback on buyers. This allows sellers to avoid buyers with reputations for being difficult in a market with many potential buyers (Murphy 2004, p. 2).

Other devices offer individuals alternatives to reputation-building in cyberspace. For instance, individuals can purchase certifications of quality and/or performance. In response to complaints of seller fraud in travel auctions, eBay began to require all sellers of vacation packages, cruises, lodging, and air travel to register with SquareTrade (Tedeschi 2004a, p. 2). SquareTrade, a privately owned seller-verification and dispute resolution company, certifies a seller only if he or she verifies the company's name, contact information, and location.

Sellers often independently seek certification. Certification providers like Comic Guaranty LLC, Professional Sports Authenticator, and numerous others have developed reputations for specializing in the inspection and grading of specific types of items in real space (e.g., comic books, sports cards). Their certifications also carry tremendous weight in the cyberspace market. For example, Dewally and Ederington (2003) examine the impact of Comic Guaranty LLC's quality certification on comic books sold through eBay. Certified comic books command a 50 percent higher price than uncertified comics, on average. Further, they price higher regardless of the seller's eBay reputation (reputation also significantly influenced price, as suggested by other studies, such as Resnick, et al. [2003], and Cabral and Hortacsu [2004]).

Other certification providers also have developed online in order to provide cyber firms with their "seals of approval" regarding various aspects of quality or performance. VeriSign Inc. is a leading supplier of encryption technology and public key arrangements. In addition to encryption/public-key services, VeriSign also provides a digital certificate "verifying that messages sent with a public key are sent by the entity to whom VeriSign distributed that key, an audit service that monitors the entity's use of and continued security of their public key infrastructure (guaranteeing that this entity is the only one with access to the private key for example) and a 'le-

[41] EBay, with an estimated 114 million users, is by far the largest of these auction sites, but there are others, such as Ubid.com, Bidville.com, ePier, and auction sections for Amazon.com, Yahoo, and Overstock.com.

gal' authority to revoke or suspend a certificate in the event that an entity does not pass an audit" (Hadfield 2000, p. 28). A VeriSign customer gets a "trustmark," which is posted on his or her website. Clicking on the trustmark moves the user to VeriSign's secure server, which displays the current information and status of the customer's digital certification. This does not completely solve the assurance problem, of course, since users must believe that the site to which they have been transported actually is VeriSign's website. They must trust VeriSign. However, as Hadfield (2000, p. 29) notes, such certification options take "a commitment problem which arises at thousands or even millions of websites and folds them back to a commitment problem for a single entity: VeriSign Inc. . . . Fundamentally, this structure moves the commitment problem from an entity (the individual e-commerce website) that faces incentives for security breach (because it is costly to maintain security or because there are gains to be had from distributing information that is suppose to be kept secret) to an entity that faces incentives for security maintenance." After all, the value certification companies like VeriSign offer is their ability to provide secure systems and their reputation for auditing and revoking certification from customers who fail to meet their security requirements.

The e-commerce market developed similar certification procedures to protect other consumer interests and privacy concerns. A group of Internet firms, including Microsoft and AOL, organized to form the Online Privacy Alliance. This group, in conjunction with the Electronic Frontier Foundation[42] and The Boston Consulting Group, started TRUSTe, a non-profit corporation. This corporation established a set of practices regarding respect for user privacy, and provides a trustmark to firms that adopt those practices (Hadfield 2000, p. 30). TRUSTe audits firms to make sure that they adhere to the practices. Certified firms have a seal that, when clicked, takes users to the firms' privacy statements. TRUSTe also offers a "click-to-verify" seal that takes the user to TRUSTe's secure server where the seal is authenticated. TRUSTe monitors compliance through regular reviews and submits user information containing identifiers that are then tracked through the firm's system. In addition, it has a "watchdog" site where users can report privacy-policy violations and other concerns. TRUSTe makes these reports available to users of TRUSTe's website. TRUSTe also maintains a dispute resolution process to resolve complaints by users who feel that their private information has been misused. Such dispute resolution processes are discussed below.

Certification of quality and performance standards also exist in cyberspace (Hadfield 2000, pp. 32-35; Kesan 2003, p. 101). For instance, the American Institute of Certified Public Accountants and the Canadian Institute of Chartered Accountants (AICPA/CICA) offer a WebTrust program.

---

[42] A non-profit organization funded by founders of Lotus Development Corporation and Apple Computers that promotes freedom of expression in cyberspace.

This combined group establishes procedures for auditing online business practices regarding privacy, security, and the handling of complaints about quality and performance. WebTrust issues an Enrollment Identification (EID) to firms that obtain a favorable report from a WebTrust licensed CPA or CA. The EID allows these firms to apply for certification by a private firm like VeriSign. In return, private firms, such as VeriSign, agree to manage a WebTrust seal. Clicking on the seal takes the user to the certificate and the accounting report. Periodic audits ensure continued compliance. Firms with WebTrust seals also must agree to submit unresolved consumer complaints to online binding arbitration by WebTrust-approved third-party dispute resolution agents. Similarly, BBBOnline offers a "Reliability seal" that certifies an online business as "reliable" and "trustworthy,"[43] along with a three-stage dispute resolution process discussed below.[44]

Certification seals are non-salvageable assets, and such investments provide a potential method of quickly building reputation. Non-salvageable assets provide a source of recourse against those who have invested in them. Information about wrongful behavior can devalue the asset. Such information may include certification withdrawal or changes in consumer willingness to deal with the wrongdoer in the face of negative information (e.g., as on eBay or from Epinions reviewers). Other methods of punishment are also available to individuals in cyberspace, but as reputation mechanisms (including certification) develop, such methods are likely to become relatively less important.

III.2.b. From trust and reputation to recourse in cyberspace

Recall Vanberg and Congleton's (1992) distinction between "retributive morality" and "prudent morality." They suggest that when individuals have an exit (or non-play) option and feel that someone else is not behaving appropriately, physical retribution becomes attractive. However, low information costs and reputations made valuable by competitively available

---

[43] See BBBOnLine, *About the Reliability Program* at http://www.bbbonline.org/reliability/index.asp.

[44] Some observers have suggested that the small portion of websites displaying seals certifying that they meet reliability and/or privacy standards suggest that these private "industry self-regulatory" efforts are insufficient to protect consumers, and that they require supplementation by government. For instance, see Kesen (2003, p. 103) and Bergerson (2001, p. 1543). Such criticisms fail to recognize that there are many ways to establish reputations, however, and the purchase of quality or performance certification is not likely to be the most cost-effective means for most firms. This is particularly true for firms with reputations that have been created over years of success offline who then move online as well. Once a reputation is established, certification is not required. When consumer concerns arise over issues that are relatively unique to cyberspace, however, firms clearly do respond. For instance, only 2 percent of the commercial Internet sites posted privacy policies in 1998, but two years later, 62 percent of these sites had such postings (Department of Commerce, 2000, p. 39).

alternatives, create different incentives. In this context, victims of wrongdoing prefer to spread information so that others in the relevant community also refuse to deal with the offender. This strategy generates a spontaneous ostracism sanction.

First, consider cyber versions of retributive morality. Physical retribution is not a reasonable option in cyberspace, but other punishment alternatives exist. Victims of fraud get "tremendously frustrated," as Ina Steiner, publisher of AuctionBytes.com, notes. "There's a sense of urgency that victims have, and it just does not synch-up with the time that it takes law enforcement to pursue these matters" (quoted in Schwartz [2004, p. 1]). As a result, some victims strike back.[45]

Victims have found a number of different options to attack Internet wrongdoers (Hafner 2004b; Schwartz 2004). For instance, some individuals monitor eBay and other auction sites, looking for fraud. When they discover what appears to be a fraudulent seller, they can warn bidders, report the activity to the auction service or the police (not necessarily a very effective option, as explained below), or make extremely high bids, thus ending the auction and preventing someone else from being victimized.[46] Some individuals even send e-mails to fraudulent sellers that attach surveillance software. The software obtains information about sellers' activities and reports them to the ISP or the auction service. In addition, fraudulent sellers commonly hijack a legitimate user's auction service account. If retribution-seeking individuals discover this, they notify the user whose ac-

---

[45] It is impossible to determine how many individuals engage in such activities since many of them are careful to maintain their anonymity, just as perpetrators do, in order to retain their Internet privileges (e.g., keep their account on eBay or some other auction service). This discussion should not be taken to infer that fraud is rampant in online auctions. EBay's estimates suggest that about one-hundredth of one percent of the sales on its site are fraudulent, although some observers suggest that the portion is somewhat higher (Hafner 2004b, p. 3). Clearly, however, eBay takes large numbers of precautions in an effort to prevent fraud. The company has some 800 employees around the world, to investigate and prevent fraud (Hafner 2004b, p. 1). Thus, the company routinely discovers fraudulent sales before the transactions are completed and warns the winning bidders not to go forward with the transactions (Hafner 2004b, p. 3). It also has warnings against unsafe practices, such as sending money through Western Union or going off the eBay site to complete a transaction, and such warnings are quite visible on the site. Indeed, efforts by eBay and other auction services to combat fraud include "innovating insurance, reputation mechanisms, dispute resolution process, certification and bonding and escrow devices" (Hadfield 2000, p. 33). Some of these activities have already been discussed, and dispute resolution processes are examined below.

[46] In many cases, the signs of fraud are very obvious to someone who knows what to look for (Hafner 2004b, p. 3). For instance, fraudulent sellers virtually always ask for payment through Western Union, and they frequently suggest going off the auction site to complete a transaction. Furthermore, they often offer to sell an item for a much lower price than a good quality item of that type generally commands. Often there is no record of feedback from previous sales unless the fraudulent seller has hijacked a legitimate user's account, and then an examination of the account's history is likely to reveal warning signals. For instance, a fraudulent sale is suggested when the item being sold is substantially different from the type of goods or services that the account history suggests are made by the seller.

count has been compromised. These actions are generally not directed at the offender who actually victimized the person seeking retribution, but they apparently generate substantial satisfaction for the victim with the possible spillover benefit of protecting another victim.[47] Low cost means of Internet communication offer a superior option in many cases, as information-based prudent morality responses can lead to boycott sanctions.

Independent providers can facilitate information flows and prudent morality sanctions. One site, iLevel.com, provides consumers the opportunity to register complaints against companies. Then, iLevel conveys the complaints to the companies. ILevel gives a company with a complaint 30 days to respond to the consumer. If the consumer remains unsatisfied after 30 days, iLevel posts the complaint and any company response on its website. Consumers check the website to see if companies they deal with have histories of unresolved complaints. Other sites offer even more extensive information to consumers. The eWatch site monitors and tracks companies' online reputation "in thousands of locations in cyberspace including online media, chat rooms and bulletin boards" (Hadfield 2000, p. 34). This site also offers services to companies. Its CyberSleuth program attempts to determine the source of negative online statements about companies and recommends solutions. The preceding sub-subsection provides numerous examples of other kinds of information mechanisms available on the Internet. Many of these systems provide means for individuals to spread positive information in an effort to build reputations, but like iLevel, they also provide ways for people to disperse negative information when they have been mistreated. The result can be significant.

Consider what happened to Intuit, the producer of TurboTax software, when it displeased some customers in 2003. Shortly after the release of the newest version of TurboTax in January of 2003, angry customers began to flood Internet forums, such as Extremetech.com, CNET.com, and Slashdot.org. Customers primarily aimed grievances and criticisms at the antipiracy features of the software (Thompson 2003, p. 1). These features made it very difficult for customers to load the program on more than one computer and also created the belief that Intuit was tracking users. Intuit quickly responded by e-mailing angry customers to assure them that it was not spying. Also, recent versions of TurboTax allow users to load the software easily on to more than one computer.

---

[47] Individuals are also taking actions against other types of online misbehavior. Some hackers release programs to repair damages done by viruses, for instance, and some "private crusaders cruise Internet chat rooms for pedophiles and report their findings to law enforcement - or even expose them online" (Schwartz 2004, p. 1). The director of the Counter Pedophilia Investigative Unit (a group of former hackers who search for predators in cyberspace) notes, however, that "law enforcement is ill-equipped to handle . . . tracking cybercrime, and particularly pedophilia" (quoted in Schwartz [2004, p. 2]).

Negative information can alter behavior very quickly when the target of the information wants to maintain its reputation and survive in the long run. If the target of the information does not respond effectively, "punishment" also spontaneously materializes. For instance, Cabral and Hortacsu's (2004, pp. 1-2) empirical analysis of eBay's reputation mechanism concludes that sales, prices, and survivability vary significantly across sellers, and depend on feedback records. Negative feedback produces significant "punishment." They report that "the growth rate of a seller's transactions drops form about 7% per week to about -7% following the first negative feedback . . . . We also find that . . . a 1 % level increase in the fraction of negative feedback is correlated with a 9 % decrease in price. . . . Moreover, a 1 % level increase in the fraction of negative feedback is correlated with a 1 to 2 % increase in the probability of exit." Naturally, a substantial amount of positive feedback can counter a small amount of negative feedback (Resnick, et al. 2003).

Complete ostracism does not occur immediately in eBay-like settings. After all, one negative feedback incident could represent a misunderstanding or even a false accusation. Indeed, as Cabral and Hortacsu (2004, p. 19) note, "Many times, when an eBay seller receives a negative comment, there is a 'war of words' between the seller and the buyer who places the negative. During this 'war of words,' the two parties can give several negatives to each other within a period of two or three days." Such contradictory information can be difficult for other parties to assess, so it can harm both parties, even though one or both might be innocent of any real wrongdoing. Not surprisingly, dispute resolution mechanisms are springing up all over cyberspace and offer alternatives to destructive "wars of words."

Most disputes are resolved through direct negotiation. Software developments lower transaction costs to negotiate in cyberspace and enhance the chances of successfully reaching agreements. Clicknsettle.com exemplifies one such software. Negotiators commit to an automated negotiation procedure managed by clicknsettle.com. If the bids come within 30 percent of each other, the program delivers an agreed upon price that splits the difference. "This process allows negotiating parties to attempt to reach agreement without revealing bids, and hence information, to each other. This preserves the value of any private information in the event there is no settlement, meaning that the act of negotiation itself does not alter either parties' position. It also decreases the scope for strategic bidding and hence the cost of a negotiation to one or both parties" (Hadfield 2000, pp. 53-54).

Another technological negotiation tool, "One Accord" software, essentially serves as a mediator. It facilitates negotiation settlement by obtaining information from each party and combines the information to generate settlement terms that are best for both parties. "One Accord takes the foundation of mediation and then adds to it both analytical rigor and technological power" (Ware and Cole 2000, p. 593; also see Katsh, Rifkin, and Gaitenby 2000, pp. 722-23). Indeed, third party mediators can use programs like One

Accord to aid them in their efforts to help disputants reach agreements. Not surprisingly, third party dispute resolution procedures are developing rapidly, as are technological aids to facilitate these dispute resolution options.

III.2.c. Third party dispute resolution in cyberspace

In December 1998, eBay asked representatives of the Center for Information Technology and Dispute Resolution at the University of Massachusetts to conduct a pilot project to see if mediation could effectively resolve disputes over auction transactions (Katsh, Rifkin, and Gaitenby 2000, p. 708). A link on eBay's customer service page informed users of available assistance for auction-related disputes. Both buyers and sellers filed complaints through the site. While customers filed substantial numbers of complaints, "the number of complaints filed suggest[ed] a rather low level of disputing relative to the overall number of transactions." (Katsh, Rifkin, and Gaitenby 2000, p. 724). When mediators receive complaints, they e-mail the other party, provide information about the mediation process, ask for information related to the dispute, and inquire about willingness to participate in mediation. Not all second parties agree to participate, but a large majority do. Ignoring the possibility of resolving a dispute implies "risk to their future online life and even to their economic wellbeing" (Katsh Rifkin, and Gaitenby 2000, p. 728). In addition to the reputation mechanism, eBay may exclude users. EBay does not often exercise this power but buyers and sellers are aware of it (Katsh, Rifkin, and Gaitenby 2000, p. 731). This eBay mediation experiment is just one of several taking place online.

"Internal Neutral," the first wholly online mediation service, uses video conferencing rather than e-mail to maintain the semblance of traditional face-to-face mediation, as conducted in real space (Ware and Cole 2000, p. 593). In face to face negotiations, the mediator can take advantage of nonverbal cues. Broadband and faster modem technology infuses online disputes with traditional offline mediation tools. However, other types of online communication, such as e-mail, offer alternative benefits that may make them superior mediation methods in some circumstances (Gibbons, Kennedy, and Gibbs 2002, p. 37).

Mediators also take advantage of these options. For instance, "The mediator can dedicate discrete time to each communicative transaction, thus reducing mediator costs. Party time and mediator time will be active productive time rather than merely sitting . . . waiting for the next stage in the mediation process . . . . Asynchronous communication does not require complex feats of scheduling so that the parties and the mediator are together at the same time . . . . Finally, the mediator may privately caucus with either or both parties without artificially interfering with the flow of the mediation. These characteristics save both time and expense while promoting efficiency in the mediation" (Gibbons, Kennedy, and Gibbs 2002, pp. 42-43). In addition, programs such as One Accord provide technological aids

to mediation. Other programs include Cybersettle, a patented double-blind bidding process (like clicknsettle, discussed above) used by over 475 insurance companies; SmartSettle, a six-stage process that uses graphic displays to assist parties to visualize negotiation progress; and Legalspace, a process that imposes structure on mediation and explains and demonstrates parties' issues and concerns (Gibbons, Kennedy, and Gibbs 2002, pp. 65-67).

Parties now frequently refer to computerized ADR communication as Online Dispute Resolution (ODR). "ODR technology may be so influential . . . as to almost become the 'fourth party' . . . . ODR ranges from mediation, which aims at encouraging the parties to reach an amicable voluntary resolution of their disagreement, to binding arbitration that imposes on the parties a legally enforceable arbitral award through the reasoned decision of arbitrator who applies the private law created by the parties to the dispute" (Gibbons, Kennedy, and Gibbs 2002, p. 40). Arbitration actually involves a much less complex communication process than mediation, so arbitration software is "much less of a challenge" to develop (Katsh, Rifkin, and Gaitenby 2000, p. 721). Thus, where sufficient pressure can induce disputants to accept an arbitration ruling (e.g., strong ostracism threats, removal of a valuable certification seal), arbitration may prove more attractive than mediation. Some independent organizations offer ODR while market providers like eBay offer but do not necessarily mandate ODR options. Still other ODR arrangements arise out of contractual agreements backed by formal exclusionary threats.

III.2.d. Recourse through contractual cyber associations

As noted, many certification providers (e.g., TRUSTe, BBBOnLine, PricewaterhouseCooper's BetterWeb, WebTrust) also provide dispute-resolution services to resolve disagreements between certified firms and their customers. For instance, firms operating under the WebTrust seal must submit to binding arbitration if they cannot satisfy a customer's complaint through direct negotiation. WebTrust developed its dispute resolution criteria with the National Arbitration Forum, a private provider of arbitration services. Other arbitration services also can resolve WebTrust disputes, but they must follow the criteria developed by WebTrust and the National Arbitration Forum (Hadfield 2000, p. 33). Parties may choose online arbitration.

By contrast, BBBOnLine developed a different arrangement.[48] When BBBOnLine receives a complaint, first it addresses complaint legitimacy. If the complaint appears to have possible merit, it moves to the organization's Privacy Policy Review Service (PPRS). The PPRS asks both the consumer and the firm to provide their evidence and arguments and renders

[48] This discussion draws on Kesan (2003, p. 102-3).

a verdict. If one of the parties finds the verdict unacceptable, the dispute goes to the Privacy Review Appeals Board (PRAB). The PRAB reviews the case, makes a ruling, and announces its decision. This decision is final. Certified firms are expected to adhere to the ruling. Refusal can result in expulsion from the certification program (Kasen 2003, p. 97). Other certification firms also threaten seal removal if firms do not accept a dispute-resolution ruling.[49]

Contractual organizations can also sanction wrongdoers such as spammers and virus writers. As Stross (2005, p. 1) notes, "We can now glimpse what once seemed unattainable: stopping the flow at its source. The most promising news is that companies like Yahoo, Earthlink, America Online, Comcast, and Verizon have overcome the fear that they would prompt antitrust sanctions if they joined forces to reclaim the control they have lost to spammers."[50] These firms belong to a new organization, the Messaging Anti-Abuse Working Group (MAAWG), which shares anti-spam techniques and tries to get other e-mail providers to join. The organi-

[49] The Internet Corporation for Assigned Names and Numbers (ICANN) also is frequently cited as an example of private governance, including dispute resolution. For instance, see Kasen (2003), and Hadfield (2000), from whom the following discussion is drawn. When the Internet was first developed, IP numbers and associated names were assigned and a list was maintained by a single individual (Jon Postel of the University of Southern California), and then by an organization under contract with the National Science Foundation (the Internet Assigned Numbers Authority). In 1991-92 the National Science Foundation solicited bids from private companies to take over the domain-name registration and IP-assignment process, and Network Solutions, Inc. was granted an exclusive contract. Pressure to introduce more competition and to privatize management of the Internet's infrastructure led the U.S. government to finally transfer responsibility to ICANN in 1998. A non-profit corporation, ICANN took over in September of 1998, and began accrediting organizations, including private corporations, to register secondary domain names, thus introducing competition into the process. ICANN also has developed a Uniform Dispute Resolution Policy (UDRP) to resolve disputes about the rights to domain names. ICANN has accredited several dispute resolution providers, beginning with the National Arbitration Forum and the World Intellectual Property Organization in December 1999, the dispute.org/eResolution consortium in January of 2000 (an arrangement that was terminated in November of 2001), CPR Institute for Dispute Resolution in May 2000, and Asian Domain Name Dispute Resolution Centre, in February 2002. Market rates are charged for dispute resolution and the providers publish their decisions on their websites as they inform ICANN. ICANN follows the resolutions' instructions, by recognizing, canceling, or transferring domain names. While ICANN is a private organization, and while it does rely on private providers of dispute resolution, it should be recognized that it was designed by the U.S. government, and it is subject to considerable political pressure: "As a result, we cannot consider ICANN a technical management corporation, but rather a political organization with well-specified constituencies and power groups" (Kasen 2003, p. 120). Thus, in contrast to various privately-initiated organizations, such as eBay and certification providers, ICANN is not an example of the bottom up development of rules and institutions.

[50] Some have raised concerns about collusion by the Group members, but as Crews (2004, p. 9) stresses, "policymakers should resist the urge to intervene, and allow alliances for purposes of security without fear of antitrust or competitive scrutiny." In fact, collective action by ISPs is not new. ISPs have been blocking known spammers listed in the Mail Abuse Prevention System's "Realtime Blackhole List" and other similar directories, for instance. Actual anti-competitive actions are unlikely, given the high value Internet users put on the freedom to communicate widely, but in a secure environment.

zation intends to go beyond filtering incoming mail to screen outgoing mail as well. For example, "Port 25 blocking" requires all outgoing mail from a service provider to go through the ISP's mail server in order to identify and cut off high-volume batches of identical mail. This blocking protocol attempts to prevent subscribers from running their own mail servers in order to send out large quantities of spam.

In addition, these ISPs also have started to digitally stamp outgoing mail, using strong cryptography to prevent counterfeiting. This digital signature tells recipients an email's origin. The outgoing digital stamp may evolve into a type of certified mail for cyberspace. Customers concerned about spam subscribe to a self-policing ISP that allows them to determine whether incoming mail is from another self-policing member of the working group. Of course, spammers can always move to other ISPs. Still, if a group of ISPs contract to exclude mail from ISPs that do not certify or pledge to prevent spam mailings from their sites, these contracting ISPs protect their own customers from spam. In addition, individuals who are not spammers will tend to move to self-policing ISPs, since non-policing ISP memberships will substantially curtail their communication possibilities. Similarly, ISPs that want to provide services to consumers other than spammers will have to join the group, or at least adopt similar (or superior) procedures, in order to survive. As a result, spammers may find that the only people they conceivably can spam are other spammers. If this cooperative security arrangement proves successful, such ISPs may determine that other problem-creating non-spammer customers should be constrained or rejected. An ISP might best serve its particular market segment by rejecting customers who fail to install and maintain anti-virus software or other security software that incorporates timely developments. Certification procedures can also develop when markets value verification that senders use such software. As Crews notes, "contractual agreements among major players could constitute a critical element of tomorrow's more secure cyber-infrastructure." (2004, p. 9).

Much of "law" concerns security or "public safety." In cyberspace, this safety issue does not concern physical safety, but rather, it concerns safety from harms or losses caused by spam, viruses, worms, fraud, identity theft, etc. Many online service providers and organizations want to be perceived as cyber-places where the risk of such losses are low. Thus, individual service providers, their customers, and various cyber organizations, such as MAAWG, interact to develop cyber-safety "law."

III.2.e. Customary Law in Cyberspace

Implementers of the eBay mediation program discussed above discovered that eBay represents a legal jurisdiction as well as a marketing arrangement. "As we encountered disputants and observed them as they participated in our process, we began to see eBay not from eBay's perspective,

which assumes that eBay is the equivalent of a landlord with little power over how a transaction is finalized, but from the user's perspective. The more we saw of this, the more we became persuaded that disputants were, indeed, participating as if they were 'in the shadow of the law.' The law whose shadow was affecting them, however, was eBay's law rather than the shadow of any other law."[51] (Katsh, Rifkin, and Gaitenby [2000, p. 728]).

Parties agreed to participate in mediation "at a very high rate" because of eBay law. Their primary concern was to maintain their eBay reputations. As Katsh, Rifkin, and Gaitenby (2000, p. 729) explain,

> EBay's response to this public safety problem was not to install a police force to deal with problems after they occurred but to use an information process to try to prevent disputes from occurring. Since the public safety problem largely focused on unknown and perhaps untrustworthy sellers and buyers, eBay put in place a process for sellers and buyers to acquire reputations as trustworthy parties .... Protecting one's feedback rating looms large in any eBay user's mind. As one guidebook to eBay points out, "on eBay, all you have is your reputation."
>
> .... While online auctions try to limit potential liability by creating distance between the auction site and those doing business in the auction site, the site owners are the designers and administrators of the process of creating identities and establishing reputations. This is a formidable power and, while it might appear that the auction site owners are merely making a process available and then letting users employ it, there are terms and conditions governing these data collection and data distribution processes, and these rules are made and administered by eBay and other proprietors of auction sites.

Of course, eBay management does not simply create eBay's rules based on arbitrary prerogative. Many rules develop as a consequence of interactions with users. If users find an arrangement unattractive, it will not last in its initial form. Recall, for instance, eBay's response to customer complaints about fraud by travel auction sellers. As a result of these complaints, eBay introduced a well-received new rule that required all travel service sellers to register with SquareTrade, the privately owned seller-verification company (Tedeschi 2004a, p. 2). Even eBay's successful dispute resolution process began as an experiment intended to see how eBay users would react to its availability. Indeed, eBay law is not and cannot be imposed through coercion. After all, other auction service providers offer eBay customers many alternative choices. As Post (1996, p. 167) notes, "Mobility - our ability to move unhindered into and out of these individual

---

[51] The terminology, "in the shadow of the law" is generally attributed to Mnookin and Kornhauser (1979, p. 968), who suggested that bargaining occurs in the shadow of the law because the legal rules give each party "certain claims based on what each would get if the case went to trial. In other words, the outcome that the law will impose if no agreement is reached ...." Since then it has often been contended that ADR also operates in the shadow of the law, where law implies state-made law (statutes, precedent). See Benson (1998a) in this regard, however. As Katsh, Rifkin, and Gaitenby (2000, p. 728) imply, non-state made law (i.e., customary law) also casts a shadow. They acknowledge that other sources of law might cast some shadows too, but they are not very significant if they do. Recourse to state-made law and public courts is rarely even mentioned.

networks with their distinct rule-sets - is a powerful guarantee that the resulting distribution of rules is a just one." EBay dominates the market because its rules and procedures produce a legal environment conducive to trust-building and voluntary exchange.

III.2.f. Polycentric cyber governance

EBay law, and, more generally, the alternative legal arrangements of auction sites, are far from unique to cyberspace. As Gibbons, Kennedy, and Gibbs (2002, p. 41) note, "many traditional businesses have learned that existing institutions such as contract law (private law making) and its corollary alternative non-judicial dispute resolution (private adjudication or ADR) may be used in new and creative ways. Both traditional and ebusiness synergistically couple the efficiency and flexibility of private law and private adjudication with the technological and communicative nature of cyberspace, achieving, in many instances, an economically optimal result." The same holds true of non-commercial groups. Numerous "online legal cultures containing what might be considered to be legal doctrine and legal processes already are emerging in many online 'places'" (Katsh, Rifkin, and Gaitenby 2000, p. 726).

When AOL filters e-mail to reject mail from a blacklisted address it promulgates a rule about spam (Post 1996, p. 169). Subscribers who believe it is a good rule remain in the AOL community, subject to AOL law. Those who believe it is a bad rule (perhaps because it infringes on the spammer's freedom of speech, or the right of individuals to receive the spam advertising) are free to leave the community and enter another ISP's jurisdiction. If AOL survives in the market with this rule, the implication is that its subscribers prefer it (or more accurately, the set of rules that characterize AOL law) to available alternatives. As Post (1996, p. 169) suggests, in cyberspace, users are free to "vote with their electrons." Furthermore, "Cyberspace is not a homogeneous place; groups and activities found at various on-line locations possess their own unique characteristics and distinctions, and each area will likely develop its own set of distinct rules" (Johnson and Post 1996, p. 1379).

Cyberspace contains many boundaries that do not correspond to the political boundaries of geographic space. These boundaries can slow or block the flow of information. Various technological intangibles mark these boundaries. These include: distinct names and addresses, required passwords, entry fees, and various visual cues created by software that distinguishes one part of cyberspace from another. "The Usenet newsgroup 'alt.religion.scientology' is distinct from 'alt.misc.legal,' each of which is distinct from a chat room on Compuserve or America Online which, in turn, are distinct from the Cyberspace Law Institute list-server or Counsel Connect. Users can only access these different forums through distinct ad-

dresses . . . , often navigating through login screens, the use of passwords, or the payment of fees." (Johnson and Post 1996, p. 1395).

These borders separate various activities and allow different Internet communities to develop their own distinct sets of rules, which evolve over time. Indeed, rules can change very quickly with such rapid online communication and information flows. Online firms and membership clubs can control participation and even prevent outsiders from learning about their activities. Behavior that may be acceptable in one cyber-community may not be tolerated in another community. Enforceable rules can establish who can enter a community, and under what conditions they can copy or redistribute data. They can exclude violators. While hackers may breach many of these boundaries, "[s]ecuring online systems from unauthorized intruders may prove an easier task than sealing physical borders from unwanted immigration [or smuggling]" (Johnson and Post 1996, p. 1397). Cyber law clearly is polycentric law.

A great deal of Internet activity crosses cyber boundaries. E-mail goes from a sender in one ISP to an inbox serviced by another ISP. Web searches take individuals all over cyberspace. Individuals download information from distant locations, enter eBay to trade and then leave, perhaps to go to a list-server or a chat room, and so on. The ease with which individuals can move from one jurisdiction to another creates important implications. First, individuals can belong to many different customary law communities as long as they behave according to the rules of each community. Second, "[i]n Cyberspace, . . . any given user has [a] more accessible exit option, in terms of moving form one virtual environment's rule set to another's, thus providing a more legitimate 'selection mechanism' by which differing rule sets will evolve over time." (Johnson and Post, 1996, pp. 1398-99).

Individuals can compare the rules offered by different communities and choose those that best meet their preferences. Competition between firms and organizations that provide similar services either leads to similar rules when all consumers have similar preferences, or leads to different rules when consumers have divergent preferences. Furthermore, technological changes that alter the ability to travel and/or create boundaries may require significant changes in both boundaries and rules. Finally, inter-jurisdictional arrangements can be expected to develop. For instance, MAAWG, the ISP organization establishes rules for communications between their subscribers. This organization makes no attempt to "harmonize" the rules that operate within each ISP beyond the degree to which they may affect inter-ISP transactions.

So, the polycentric nature of cyber law remains with the added development of second order clustering, as predicted by Vanberg and Buchanan (1990) and illustrated by various real space examples (Pospisil 1971; Benson 1991a, 1994a). While this may sound like a very complex and confusing system of rules, most of these customary communities function on a

narrow focus. Therefore, each community develops rules only according to those interactions relevant to the community function. Each community's rules are likely to be quite simple, since its purpose is to facilitate the voluntary interactions of community members and to protect those members from harm. One community's rules may differ extensively from another's but that does not put the rules at conflict as rules may arise in the context of very different kinds of interactions. In contrast, a nation state that attempts to monopolize all law will require a very large and complex set of rules, and as explained below, these rules often will have conflicting purposes. Thus, polycentric law is not necessarily any more complex and difficult to manage than monocentric law, and it may even prove much less complex.

IV. WHY NATION-STATES CANNOT AND SHOULD NOT TRY TO RULE CYBERSPACE: RELATIVE BENEFITS AND COSTS OF POLYCENTRIC CUSTOMARY LAW

In 2001, WESTLAW and LEXIS searches for computer virus prosecutions produced only one case involving a Cornell University computer science student who researched information security (Rustad 2001, p. 85).[52] While more prosecutions may have occurred since then, their numbers remain small relative to the level of violations.[53] As Rustad (2001, pp. 85-86) notes, "The poverty of cybercrime cases reflects a substantial enforcement gap between the cybercrime law on the books and the law in action. Few cybercrimes have been successfully prosecuted because of several interrelated factors, including the problem of anonymity, jurisdictional issues, and the lack of resources in the law enforcement community. Conventional law enforcement does not . . . have the resources to tackle this kind of crime." Nonetheless, there is no doubt that governments will attempt control various kinds of cyber activity by imposing rules. These government efforts will not achieve their objectives but will produce other consequences. Citizens will have substantial reason to resist such governmental efforts because they will mostly result in undesirable consequences. Consider the reasons for government's inability to control cyberspace before turning to the reasons for discouraging and even resisting government efforts to do so.

---

[52] As part of his research, Robert Morris designed a worm to test the security of a computer network, but when he performed the test by releasing the worm in a computer science laboratory at MIT, a defect in the program led to rapid replication, resulting in university, medical facility, and defense facility computer shutdowns throughout the U.S. Morris was prosecuted and sentenced to three years probation, a $10,500 fine, and 400 hours of community service.

[53] A Lexis-Nexus in March, 2005 search of several large states' case records found no prosecutions, but this was not a complete search.

IV.1. *Why Nation-States Cannot Rule Cyberspace*

While many people believe the state must step in to provide cyberspace law, they fail to realize that geographically defined nation-states are actually incapable of establishing effective cyber law. For example, the governments of the United States, and of states like New York and Minnesota have attempted to stop Internet gambling. Yet gambling sites continue to proliferate, and the number of Internet gamblers and gambling revenues continues to grow almost exponentially (Bell 1999, Strumpf 2004). Similarly, government law enforcement has failed attempts to shut down the traffic in pornography. Governments also have attempted to control spam, viruses and worms, and to prevent fraud and identity theft. These efforts result in "successes" similar to those of the antigambling and antipornography campaigns. Three primary reasons explain the nation state's inability to establish enforceable cyber law are: (1) jurisdictional issues; (2) anonymity in cyberspace; and (3) the high opportunity cost of law enforcement resources devoted to cyber detection and prosecution (Rustad 2001, pp. 85-86).

IV.1.a. Jurisdictional constraints

Jurisdictional problems represent a major impediment to nation-states' efforts to govern in cyberspace. "Cyberspace radically undermines the relationship between legally significant (online) phenomena and physical location. The rise of the global computer network is destroying the link between geographical location and: (1) the power of local governments to assert control over online behavior; (2) the effects of online behavior on individuals or things; (3) the legitimacy of local sovereign's efforts to regulate global phenomena; and (4) the ability of physical location to give notice on which set of rules apply. The Net thus radically subverts the system of rule-making based on borders between physical spaces, at least with respect to the claim that Cyberspace should naturally be governed by territorially defined rules." (Johnson and Post 1996, p. 1370).

If one government cracks down on an Internet activity, such as gambling, pornography, or involuntary externality-creating activities like spam, viruses, fraud, and identity theft, perpetrators can simply set up their operations in other geographic locations. For instance, many companies base Internet gambling sites out of places like Costa Rica, Australia, Great Britain, and Antigua and Barbuda. The owners of these gambling establishments may not be able to set foot in the United States without getting arrested (Richtel 2004a), but they can still sell their services to millions of consumer in the United States. Spammers choose similar routes to avoid detection or prosecution.

Some countries, like China, actually encourage externality generating cyber activities such as spam because they create local economic benefits (Zeller 2005, p. 2). Some corrupt governments even encourage other kinds of cyber crime "as a developing industry" (Rustad 2001, p. 86). Indeed, many of the viruses, worms, identity theft and fraud activities that attack individuals around the world originate in countries like Brazil and Turkey (Smith 2003), where governments are either unwilling or unable to stop such activities. For instance, "[t]he 20 officers working for the electronic crime division of the Sao Paulo police catch about 20 cybercrooks a month. But those criminals account for but a fraction of the 'notorious and ever increasing' number of cybercrimes in Sao Paulo" (Smith 2003, p. 2). Almost 96,000 "overt Internet attacks" were traced to Brazil between January and October of 2003. Given the large number of cyber criminals in Brazil, the small number of arrests makes virtually no difference in cyber crime activity. Brazilian law does not consider hacking and viruses illegal unless they result in some other crime (e.g., theft, fraud). As a result, these individuals have "little to fear, legally" (Smith 2003, p. 2).

Jurisdictional constraints have not stopped governments from trying to impose law in cyberspace. For instance, some governments have attempted to extend the reach of their law beyond their geographic borders using cyberspace. In December 2002, Australia's High Court allowed an Australian libel suit against Dow Jones, despite the fact that its web server was located in New Jersey. The suit concerned the contents of an article that appeared on the Dow Jones website. The article appeared in *Barron's*, a weekly financial magazine owned by Dow Jones and published in print as well as on line. Online subscriptions can be purchased anywhere in the world, including Australia. Differences between Australian and New Jersey libel law made the chances of a ruling in favor of the plaintiff much greater in Australia (Economist Global Agenda 2002, p. 2). Still, Australia will have difficulty enforcing such a ruling since the target of the suit is located outside its jurisdiction (Thierer and Crews 2002, p. 1).

In several other cases, national courts have held that cyberspace activities breached local laws could be prosecuted. For instance, a French court ruled that the sale of Nazi memorabilia on a Yahoo auction site violated a French law against displaying the Nazi insignia. Yahoo challenged the French ruling in U.S. courts but still banned all "hate paraphernalia" from its auction sites. Still, this holding did little to remove Nazi symbols or philosophy from the Internet. Several years ago, the German Interior Minister identified almost 800 neo-Nazi websites outside Germany that were accessible to Germans and violated German law (Thierer 2001). Most of these sites were in the U.S. Germany's Supreme Court held foreign distributors of neo-Nazi Internet materials liable under German law, but the ruling proves essentially unenforceable outside Germany. In another case, the government of Zimbabwe prosecuted an American reporter for "publishing a falsehood" about its government (Economist Global Agenda 2002,

p. 2). Ultimately, the court acquitted the reporter, but it did not hesitate to exercise jurisdiction in the case.

Only international companies like Yahoo are likely to respond to foreign court rulings. For example, other firms with assets in France have taken notice of the ruling since their assets could be vulnerable if the firm is sued in French courts. Intuit Inc., an online database software system, canceled its services in France. "In the uncertain legal climate still surrounding e-commerce, online businesses often choose to give up customers in some countries rather than leave themselves open to possible libel, product-liability and other kinds of lawsuits in those countries" (Newman 2003, pp. 1-2).

Government actions such as those described above naturally may offend American citizens who highly value freedom of speech and of the press. Still, governments in most other countries of the world do not share these values (Thierer 2001, p. 2). At least 59 countries have laws, such as those in France, that limit freedom of expression online (Corn-Revere 2002, p. 6). The U.S. government, constrained by the First Amendment, has not cooperated in the enforcement of such laws against its citizens. Yet, it has adopted similar tactics in its effort to stop cyber activities that do not agree with U.S. law.

In March 2004, the World Trade Organization (WTO) released its first cyberspace decision. It ruled that U.S. policy prohibiting online gambling violated international trade law. "[S]everal members of Congress said they would rather have an international trade war or withdraw from future rounds of the World Trade Organization than have American social policy dictated from abroad." (Richtel 2004b). Law enforcement agencies apparently agreed. In April 2004, U.S. marshals seized $3.2 million from Discovery Communications, as part of the continuing effort to stop Internet gambling. Tropical Paradise, the Costa Rican owner of ParadisePoker.com paid to advertise the gambling site on the Discovery Channel (Richtel 2004c, p. 1). This seizure clearly was intended to tell American companies that if they do business with offshore gambling operations, the government will seize the proceeds from that business. As a consequence of the Discovery seizure, major broadcasters including Infinity Broadcasting and Clear Channel Communications and Internet firms such as Yahoo and Google stopped running Internet gambling ads. Such advertising has since reappeared, thus illustrating the futility of U.S. government efforts.

In July 2001, the U.S. government arrested Dmitry Sklyarov, an employee of a Moscow-based computer company, Elcomsoft Co., a Moscow based company. The government charged Sklyarov with breaking U.S. copyright law because he disabled a security system used to protect electronic books. His program was legal in Russia, however, where it was produced and sold. Prosecutors made a deal with Sklyarov, agreeing to defer the charges against him if he testified against his employer. In December 2002, a jury acquitted Elcomsoft in U.S. District Court, and the charges

against Sklyarov were then dismissed (Newman 2003, p. 2). Nonetheless, the fact that prosecutors were willing to pursue a case against a foreign firm for actions taken in another jurisdiction is really no different than the French, Zimbabwe, and Australian actions against U.S. based entities.

Some countries go much further than France, Zimbabwe, Australia, and the United States in their efforts to limit online content. North Korean law completely bans Internet use, while countries like China, Singapore, Saudi Arabia, and Syria control ISPs in order to filter content that government officials consider inappropriate (Corn-Revere 2002, p. 6-8). For instance, the 30 permitted ISPs in Saudi Arabia are linked into a single web system that screens all incoming files for "offensive or sacrilegious material" before it is released to individual users (Corn-Revere 2002, p. 7). The servers are also programmed to block access to "sensitive" sites that might violate "the social, cultural, political, media, economic, and religious values of the Kingdom" (Human Rights Watch 1999, quoted in Corn-Revere 2002, p. 7). Similarly, Singapore requires ISPs to install filters that remove content the government finds objectionable. China attempts to exclude all forms of dissent on the web and prohibits publication of news not approved by the Communist Party, as well as other content. Syria has only one government-run ISP that can impose significant content blocking. Still, such efforts likely will prove unsuccessful. Indeed, illegal networks have sprung up across Asia, Africa, and Latin America in response to government efforts to restrict access (Gordon 2004). Nonetheless, these governments' "futile" efforts create considerable external costs not unlike those arising from viruses and worms (Corn-Revere 2002, p. 13).

Efforts by governments of countries like the U.S., France, Germany, Australia, China, Singapore, Syria, Saudi Arabia, and Zimbabwe create conflicts between governments. Each government claims jurisdiction over actions taken in another's jurisdiction and holds illegal actions another government legally allows. As a result, some observers believe an international organization of governments should enforce a common set of cyberspace rules. One obvious problem with this proposed solution is that governments already cannot agree on common rules. Should the limits on content imposed in Saudi Arabia and Syria apply everywhere; should China's limits rule; or should France and Germany's laws be adopted? Other countries likely would not adopt U.S. rules regarding free speech and press. Should U.S. rules regarding gambling be instituted? Indeed, what are those rules, since several states now allow gambling and even run gambling operations (e.g., lotteries), and the federal government has agreed to let Indian Reservations open Casinos where state laws do not explicitly forbid it (Johnson 2004)? Similarly, is U.S. copyright law to dominate or China's? Will spam be limited as under U.S. law or encouraged as under Chinese law?

Countries simply are not willing to give up sovereignty just because their citizens participate in global cyberspace. The World Intellectual Property Organization (WIPO) managed to negotiate terms for two interna-

tional treaties on copyright protection in 1996, and 41 countries (out of over 190) ratified the treaties by the time they finally went into effect in 2002. "[I]t's still up to the individual countries to adapt their national legislation to the treaties, and for the most part implementation has been slow. . . . There is also the problem of enforcement. While the music industry saluted the WIPO treaty related to music copyrights, it has complained that governments remain lax about prosecuting violators." (Newman 2003, p. 3). Similarly, delegates from 60 countries gathered at the Hague in 2003 for the Conference on Private International Law. The countries met in part to negotiate a treaty that would grant comity to court rulings among signing countries, but "not surprisingly, there is little political will for such a sweeping deal" (Newman 2003, p. 3). In 1999, a draft treaty proposed global standards for defamation, libel, and copyright protection, but private business protested so strongly that negotiators conceded that any accord was out of reach. Protesting businesses included: U.S. media and entertainment firms; writers and publishers who would have faced libel suits in countries with strict libel laws; and ISPs who would have faced liability for postings on websites accessible by their services,. Years of negotiation have not even resolved jurisdictional issues, so it is not surprising that they have not produced a harmonized cyber law (Newman 2003, p. 4).

IV.1.b. Anonymity

Even if governments could resolve jurisdictional issues and reach agreements regarding the appropriate rules to enforce, they would face detection and punishment problems. Individuals who want to break cyber laws work at an advantage. After all, they can hide in cyberspace. "The overlapping and truly global networks of spam-friendly merchants, e-mail list resellers, virus-writers and bulk e-mail services have made identifying targets for prosecution a daunting process. Merchants whose links actually appear in junk e-mail are often dozens of steps and numerous deals removed from the spammers . . . and proving culpability 'is just insanely difficult.'" (Zeller 2005, p. 4). Anonymity is further enhanced because spammers use viruses to gain control of PCs to use as "zombie spam transmitters." So, even if the government traces spam to a particular computer, it may find the owner unaware of the activity and completely innocent of any wrongdoing. "In contrast to a traditional crime scene, online forgers or intruders leave few digital footprints. DNA evidence, fingerprints, or other information routinely tracked in law enforcement databases are useless for investigating cybercrimes. In addition, computer records are easier to alter than paper and pencil records. Electronic robbers and forgers leave fewer clues than white-collar criminals who alter checks or intercept promissory notes." (Rustad 2001, p. 98).

IV.1.c. Opportunity costs of law enforcement resources

The high costs to pursue criminals in cyberspace means that while many governments have laws mandating prosecution of many cyber activities, most of them have no "significant law enforcement presence in cyberspace" (Rustad 2001, pp. 98-99). Even wealthy countries will not likely invest in the necessarily large increase in resources required to enforce cyber law. This investment requires governments to train and retain officers that can understand encryption; decipher digital signatures, clues, and viruses; and satisfy other technological requirements necessary to track cyber wrongdoers.

Even a very "successful" investigation involves high costs, but likely creates a negligible impact. For instance, the Justice Department filed its first criminal charges under the nation's new anti-spam law in April 2004, but in July, "the case was quietly dismissed at the government's request" (Hansell 2004a, p. 2). Similarly, in May 2004, the FBI told a Senate committee that they were developing over 50 cases against spammers and expected to announce convictions that summer. Instead, "the cases have proven more complex than expected" (Hansell 2004a, p. 2). In August 2004, the Justice Department finally announced a large number of arrests (more than 150) (Hansell 2004b). The surrounding investigation involved 37 FBI agents, 13 divisions of the Postal Inspection Service, and several other federal and local law enforcement agencies. Since the arrests, no further information about the disposition of these cases seems to have emerged. Furthermore, the U.S. government might identify a criminal who is a citizen of another government, which does not consider the activity in question illegal. Even if the alleged criminal is captured within the United States, successful prosecution is far from assured, given governments' inability to resolve jurisdictional conflicts, as suggested by the Elcomsoft case. Beyond the fact that successful prosecutions of cybercriminals are rare, society highly values alternative uses of police and prosecution resources (e.g., fighting crime on the streets), and these uses prove more likely to be successful.

IV.2. *Why Nation-States Should Not Attempt to Rule Cyberspace*

Even if states could overcome jurisdictional conflicts, the inability to agree on an appropriate set of cyberspace rules, and the lack of sufficient resources necessary to impact rule enforcement, other reasons still argue against such a system. In particular, while Coase (1960) emphasizes that one motivation for creating rules is to eliminate externalities and facilitate voluntary interaction, he also explains that rules and institutions determine the distribution of bargaining power and therefore the distribution of wealth. Coase does not focus on this issue, but these distributional conse-

quences also create incentives to make and alter rules, as emphasized in the rent-seeking literature that has evolved from Tullock's (1967) insights. In fact, the politicized "law" of nation-states almost always reflects efforts to achieve the conflicting objectives of facilitating voluntary activities and involuntary transfers (Benson 1999a). Governments enforce rules, established either by kings, dictators, "representative" parliaments, or courts, within a, coercive legal jurisdiction. Unlike voluntary joint production and exchange, which tends to increase wealth, involuntary wealth transfers tend to reduce wealth for at least six reasons (Benson 1999b, 1999c, 2000b, 2001b).

IV.2.a. Deadweight losses due to wealth transfers

When government-made law mandates wealth transfers,[54] it causes deadweight loss. "The dangers of the cyber-pork barrel should be obvious. Washington subsidy and entitlement programs typically have a never-ending lifespan . . . ." (Thierer, Crews, and Pearson 2002, p. 1). For instance, consider U.S. policy toward Internet gambling. As Bell (1999, p. 3) notes, several powerful interest groups have significant incentives to lobby for prohibition of online gambling. Offline gambling firms in Nevada, New Jersey, Mississippi, Louisiana and elsewhere make large contributions to political candidates in an effort to influence such policies. Indian-reservation gambling also has exploded over roughly the same period that Internet gaming has been developing, and Indian gambling operations direct a substantial portion of their increasing wealth to the political arena (Johnson 2004).

While gambling operators would prefer not to compete against anyone, they certainly do not want to compete with low-cost and convenient online casinos. In addition, "[s]tate and municipal authorities, having grown fond of nurturing and taxing local gambling, can easily see that Internet gambling might put their cash cows out to pasture" (Bell 1999, p. 4). Billions of dollars from gambling taxes and state lotteries flow into government treasuries in the United States. The government does not collect revenues from Internet gambling. Therefore, state and local governments support their local gambling establishments in lobbying to prohibit Internet gaming. Furthermore, "[w]hether or not Internet gambling represents a moral scourge, it certainly represents a competitive threat to church bingo games and the like," and charitable gaming brings churches several billion dollars annually (Bell 1999, p. 4).

If the government significantly restricts online gambling, established offline gambling operators (including government-run, operations) feel free

---

[54] E.g., through a tax and/or subsidy, through trade barriers and other limits on competition such as licenses and exclusive franchises, and through other similar discriminatory "legal" actions.

to reduce consumer payouts below the competitive level. Also, gamblers face additional costs because they must travel to government-designated gambling sites rather than simply going online. Similar motives underlie demands to limit sales of other goods (e.g., liquor) and services over the Internet. State and local governments are unable to collect their sales and excise taxes, and local business groups are unable to limit competition to those firms with government licenses and franchises, in order to keep their prices high.

IV.2.b. Rent-seeking costs

Many observers suggest that deadweight losses from rent-seeking induced wealth transfers are small and that institutions evolve to minimize them (e.g., Becker 1983), but Tullock (1967) emphasizes the opportunity costs to resources consumed in the competition for such transfers. Individuals and groups have incentives to invest time and resources in an effort to gain wealth through the political process. Even though government efforts to rule the Internet are relatively unsuccessfully high tech firms "are becoming more comfortable in Washington circles as they open up D.C. lobbying offices and begin spreading cash around to candidates for office in the hope of courting favor and prevailing in policy debates." (Thierer, Crews, and Pearson 2002, p. 13). Similarly, other Washington lobbyists represent a wide variety of interests with Internet concerns. These include: gambling interests; state and local governments in search of Internet commerce taxes; licensed businesses organizations, such as liquor outlets that face competition from online sellers; organizations such as the "Direct Marketing Association" which is promote "e-mail marketing" (i.e., the right to send spam); music and movie producers; and law enforcement groups in search of Internet crime budgets. These groups want the government to restrict certain Internet activities so they can capture rents.

IV.2.c. Costly protection of property rights

Prospective wealth transferors also have incentives to defend their property rights. Part of these defense costs include rent-avoidance costs from investments in political information and influence (e.g., the Direct Marketing Association mentioned above). Exit is another option and can be achieved by moving to an alternative political jurisdiction, or by hiding economic activity and wealth (e.g., moving transactions "underground" into black markets). The ability to hide anonymously in other jurisdictions makes cyberspace exits particularly easy. Again, this is one reason governments prove ineffective in imposing cyber laws. Still, exit reduces opportunities for wealth creation within jurisdictions that attempt to control Internet activities.

IV.2.d. Enforcement costs

To induce compliance with discriminatory transfer rules, governments must rely on enforcement bureaucracies to prevent exit (e.g., establish a monopoly in law) and to execute the rules. Enforcement costs represent another set of opportunity costs in the wealth transfer process. Rules that facilitate voluntary production and exchange (e.g., private property rights) also require some enforcement costs (e.g., dispute resolution, information mechanisms), but relative to costs to restrict voluntary activities(e.g., generate involuntary wealth transfers from those who may want to sell a good or service but are not suppose to because of some government-created barrier to competition), such costs prove small. The resources allocated by governments in China, Saudi Arabia, Singapore, France, the U.S. and elsewhere in what is destined to be a futile effort to control the flow of communication in cyberspace, all carry opportunity costs. For instance, the allocated resources could enforce laws in real space or produce goods and services for private consumption.

IV.2.e. Lost innovations and superfluous discoveries

Involuntary wealth-transfer rules alter the path of technological and institutional evolution. This means that discoveries which probably would have been made in the absence of such rules may never be made (Kirzner 1985, pp. 141-44). The opportunity costs of such laws include these lost discoveries (Benson 2002, Thierer, Crews, and Pearson 2002, p. 12). In addition, rules of this kind create a "wholly superfluous" discovery process based on "entirely new and not necessarily desirable opportunities for entrepreneurial discovery" (Kirzner 1985, p. 144). This process represents yet another source of opportunity costs (Benson 2002, Thierer, Crews, and Pearson 2002, p. 12).

Government policy can redirect or stifle institutional and organizational innovation. For instance, Jamal, Mairer, and Sunder (2003) compare privacy practices that have developed in the United Kingdom under codified European Union standards with those that have evolved in the United States, where much less government regulation has been imposed. They find the state regulatory regime in the U.K. stifled the development of web assurance services. Certification markets do not exist in the U.K., which exposes Internet users in the U.K. to more spam and other externality-producing cyberspace activities. Thus, while government efforts to rule the Internet generally prove futile and fail to accomplish what they intend (if one believes the political rhetoric rather than looking at the incentives of the interest groups that actually motivate most laws), they also raise costs and make the typical Internet user worse off.

IV.2.f. Uncertainty and reduced economic progress

If governments successfully impose their rules in cyberspace, then individuals face increased probabilities of involuntary transfers. This increased probability renders individual property rights to resources, wealth, and income flow insecure. It reduces incentives to invest in the maintenance of and improvements to assets and to earn income and produce new wealth., When transfers are expected to be large, frequent, and arbitrary, governments must resort to threats to motivate most wealth production (e.g., as under slavery or totalitarian socialism). This drives enforcement costs even higher. Since such threats are imperfect, they induce low production and slow wealth expansion relative to an environment with secure property rights. Government action that threatens cyberspace property rights implies slowdowns in cyber economy growth.

Perhaps some coercive authoritarian law does not produce biased wealth-transferring rules, but since the effectiveness of such rules requires strong barriers to exit for those who expect to lose wealth through transfers. Inter-jurisdictional competition can occur between legal systems that attempt to monopolize law-making and enforcement. If wealth is able to escape to another jurisdiction, it limits the potential to use the law as a transfer mechanism. This is another obvious benefit from inter-jurisdictional competition. Customary law represents another important source of competition. Institutions that do not attempt to monopolize the law can produce and support customary law, and these legal systems offer an escape alternative to jurisdictions that seek such monopolies. Customary law provides a mechanism to avoid the politicized law of nation-states.

REFERENCES

Acheson, J. M. 1988. *The Lobster Gangs of Maine.* Hanover, NH: University Press of New England.

Alchain, A. A., and Allen, W. R. 1969. *Exchange and Production, Theory in Use.* Belmont, CA: Wadsworth Publishing Co.

Alexander, K. 2003. Now you can leave the eBay selling to them. *New York Times*, December 21, Section 3, p. 8.

Alexander, R. D. 1987. *The Biology of Moral Systems.* Hawthorne NY: Aldine de Gruyter.

Anderson, S. W., J. D. Daly, and M. F. Johnson. 1999. Why firms seek ISO 9000 certification: Regulatory compliance or competitive advantage. *Production and Operations Management* 8:28-43.

Anderson, T. L., and P. J. Hill. 1975. The evolution of property rights: A study of the American west. *Journal of Law and Economics* 18:163-79.

Anderson, T. L., and P. J. Hill. 1979. An American experiment in anarcho-capitalism: The not so wild, wild west. *Journal of Libertarian Studies* 3:9-29.

Anderson, T. L., and P. J. Hill. 1990. The race for property rights. *Journal of Law and Economics* 33:177-97.

Anderson, T. L. and P. J. Hill. 2004. *The Not so Wild, Wild West: Property Rights on the Frontier.* Stanford, CA: Stanford Economics and Finance.

Anderson, T. L. and F. McChesney. 1994. Raid or trade: An economic model of Indian-White relations. *Journal of Law and Economics* 37:39-74.

Ashenfelter, O. 1987. Arbitration behavior. *American Economic Review, Papers and Proceedings.* 77:342-46.

Associated Press. 2005. Microsoft is acquiring maker of software to combat viruses. *New York Times, nytimes.com*, February 9. Available at: http://www.nytimes.com/2005/02/09/technology/09soft.html.

Axelrod, R. 1984. *The Evolution of Cooperation.* New York: Basic Books.

Bajari, P. and A. Hortacsu. 2003. Winner's curse, reserve prices and endogenous entry. *Rand Journal of Economics* 34:329-55.

Bakos, Y. 2001. The emerging landscape for retail e-commerce. *Journal of Economic Perspectives.* 15:69-80.

Barzel, Y. 1989. *Economic Analysis of Property Rights.* Cambridge, UK: Cambridge University Press.

Bayot, J. 2004. Card seem at risk? Try a stunt double. *New York Times*, May 16, Section 3, p. 10.

Becker, G. S. 1983. A theory of competition among pressure groups for political influence. *Quarterly Journal of Economics* 98:371-400.

Bell, T. W. 1999. Internet gambling: Popular, inexorable, and (eventually) legal. *Cato Institute Policy Analysis*, No. 336.

Benson, B. L. 1989. The spontaneous evolution of commercial law. *Southern Economic Journal* 55:644-61.

Benson, B. L. 1991a. An evolutionary contractarian view of primitive law: The institutions and incentives arising under customary American Indian law. *Review of Austrian Economics* 5:65-89.

Benson, B. L. 1991b. Reciprocal exchange as the basis for recognition of law: Examples from American history. *Journal of Libertarian Studies* 10:53-82.

Benson, B. L. 1994a. Are public goods really common pools: Considerations of the evolution of policing and highways in England. *Economic Inquiry* 32:249-71.

Benson, B. L. 1994b. Emerging from the Hobbesian jungle: Might takes and makes rights. *Constitutional Political Economy* 5:129-58.

Benson, Bruce L. 1998a. Arbitration in the shadow of the law. In *The New Palgrave dictionary of economics and the law*, vol. 1, edited by P. Newman. London: Macmillan Reference Limited, pp. 93-98.

Benson, Bruce L. 1998b. Evolution of commercial law. In *The New Palgrave dictionary of economics and the law*, vol. 2, edited by P. Newman. London: Macmillan Reference Limited, pp. 89-92.

Benson, Bruce L. 1998c. Law Merchant. In *The New Palgrave dictionary of economics and the law*, vol. 2, edited by P. Newman. London: Macmillan Reference Limited, pp. 500-508.

Benson, Bruce L. 1998d. *To serve and protect: privatization and community in criminal justice*. New York: New York University Press.

Benson, Bruce L. 1999a. An economic theory of the evolution of governance, and the emergence of the state. *Review of Austrian Economics* 12:131-60.

Benson, Bruce L. 1999b. To arbitrate or to litigate: That is the question. *European Journal of Law and Economics* 8:91-151.

Benson, Bruce L. 1999c. Polycentric law versus monopolized law: Implications from international trade for the potential success of emerging markets. *Journal of Private Enterprise* 15:36-66.

Benson, Bruce L. 2000a. Arbitration. *Encyclopedia of Law and Economics* 5:159-93.

Benson, Bruce L. 2000b. Interjurisdictional competition through alternative dispute resolution. In *What Price Civil Justice?*, edited by B. Main and A. Peacock. London: The Institute of Economic Affairs, pp. 69-96.

Benson, Bruce L. 2000c. Jurisdictional choice in international trade: Implications for lex cybernetoria. *Journal des Economistes et des Etudes Humaines* 10:3-31.

Benson, Bruce L. 2001a. Knowledge, trust, and recourse: Imperfect substitutes as sources of assurance in emerging economies. *Economic Affairs* 21:12-17.

Benson, Bruce L. 2001b. Law and economics. In *The Elgar Companion to Public Choice*, edited by W. Shughart II and L. Razzolini. London: Edward Elgar, pp. 547-89.

Benson, Bruce L. 2002. Regulatory disequilibrium and inefficiency: The case of interstate trucking. *Review of Austrian Economics* 15:229-55.

Benson, Bruce L. 2004. Property rights and the buffalo economy of the Great Plains. Working paper, Florida State University. Written for inclusion in *Self Determination: The Other Path for Native Americans*, edited by T. L. Anderson, B. L. Benson, and T. Flanagan. Stanford University Press, Forthcoming.

Bergerson, Stephen R. 2001. E-commerce privacy and the black hole of cyberspace. *William Mitchell Law Review* 27:1527-55.

Berman, H. J. 1983. *Law and Revolution: The Formation of Western Legal Tradition.* Cambridge, MA: Harvard University Press.

Berman, H. J. and F.J. Dasser. 1990. The 'new' law merchant and the 'old': Sources, content, and legitimacy. In *Lex Mercatoria and Arbitration: A Discussion of the New Law Merchant*, edited by T. E. Carbonneau. Dobbs Ferry, NY: Transnational Juris Publications.

Bernstein, L. 1992. Opting out of the legal system: Extralegal contractual relations in the diamond industry. *Journal of Legal Studies* 21:115-58.

Bloom, D. E., and C. L. Cavanagh. 1986. An analysis of the selection of arbitrators. *American Economic Review* 76:408-22.

Böckstiegal, K.-H. 1984. *Arbitration and State Enterprises: A Survey of the National and International State of Law and Practice.* Deventer, Netherlands: Kluwer Law and Taxation Publishers.

Cabral, L., and A. Hortacsu. 2004. The dynamics of seller reputation: Theory and evidence from eBay. National Bureau of Economic Research, Working Paper 10363.

Carter, R. B., and S. Manaster. 1990. Initial public offerings and underwriter reputation. *Journal of Finance* 45:1045-67.

Casella, A. 1992. Arbitration in international trade. Center for Economic Policy Research, Discussion Paper No. 721.

Cheung, S. N. S. 1974. A theory of price control. *Journal of Law and Economics* 17:53-72.

Coase, R. H. 1960. The problem of social cost. *Journal of Law and Economics* 3:1-44.

Commons, J. R. 1924. *Legal Foundations of Capitalism.* New York: Macmillan.

Cooter, R. 1994. Structural adjudication and the new law merchant: A model of decentralized law. *International Review of Law and Economics* 14:216.

Corn-Revere, R. 2004. Caught in the seamless web: Does the Internet's global reach justify less freedom of speech? *Cato Institute Briefing Papers*, No. 71.

Cox, Braden, and Crews, Clyde Wayne. 2005. Helping Hollywood help itself—Protecting digital property without new legislation. *Competitive Enterprise Institute C:Spin*, No. 176. Available at: http://www.cei.org/gencon/016,04412.cfm.

Crawford, S. P. 2005. Shortness of vision: Regulatory ambition in the digital age. Cardozo Legal Studies Research Paper No. 102.

Crews, Clyde Wayne. 2003. Stop this today! Unsolicited e-mail vs. unsolicited legislation. *National Review Online*, June 13. Available at: http://www.nationalreview.com/comment/comment-crews061303.asp.

Crews, Clyde Wayne. 2004. Cybersecurity and authentication: The marketplace role in rethinking anonymity—Before regulators intervene. *Competitive Enterprise Institute Issue Analysis*, 2004 No.2.

Demsetz, H. 1967. Toward a theory of property rights. *American Economic Review* 57:347-59.

Department of Commerce. 2000. *Leadership for the New Millennium: Delivering on Digital Progress and Prosperity.* Washington, DC: Department of Commerce. Available at: http://www.usembassy.it/pdf/other/ec2000.pdf.

de Soto, H. 1989. *The Other Path: The Invisible Revolution in the Third World.* New York: Perennial Library.

Dewally, Michael and Louis H. Ederington. Reputation, certification, warranties, and information as remedies for seller-buyer information asymmetries: Lessons from the online comic book market. *Journal of Business*, Forthcoming. Available at: http://ssrn.com/abstract=548383.

Diamond, D. W. 1989. Reputation acquisition in debt markets. *Journal of Political Economy* 97:828-62.

Draetta, U., R. B. Lake, and V. P. Nanda. 1992. *Breach and Adaptation of International Contracts: An Introduction to Lex Mercatoria.* Salem, NH: Butterworth Legal Publishers.

Economist Global Agenda. 2002. A jurisdictional tangle: Media companies around the world are alarmed by a high-court ruling in Australia. *Economist, Economist.com*, December 10. Available at: http://www.economist.com/agenda/PrinterFriendly.cfm?Story_ID=1489053.

Einhorn, M. A., and B. Rosenblatt. 2005. Peer-to-peer networking and digital rights management: How market tools can solve copyright problems. *Cato Institute Policy Analysis*, No. 534.

Ellickson, R. C. 1991. *Order Without Law: How Neighbors Settle Disputes.* Cambridge, MA: Harvard University Press.

Ellickson, R. C. 1993. Property in land. *Yale Law Journal* 102:1315-1400.

Friedman, D. D. 1979. Private creation and enforcement of law: A historical case. *Journal of Legal Studies* 8:399-415.

Friedman, D. D. 2001. Will strong encryption protect privacy and make government obsolete? *Independent Policy Forum.* Available at: http://www.independent.org/ events/transcript.asp?eventID=20.

Fuller, L. L. 1964. *The Morality of Law.* New Haven: Yale University Press.

Fuller, L. L. 1981. *The Principles of Social Order.* Durham, NC: Duke University Press.

Gallagher, D. F. 2004. Users find too many phish in the Internet sea. *New York Times*, September 20, p. C4.

Gibbons, L. J., R. M. Kennedy, and J. M. Gibbs. 2002. Cyber-mediation: Computer-mediated communications medium messaging the message. *New Mexico Law Review* 32:27-73.

Gordon, H. S. 1954. The economic theory of a common property resource: The fishery. *Journal of Political Economy* 62:124-42.

Gordon, J. 2004. Illegal Internet networks in the developing world. Harvard Law School Public Law Research Paper No. 94.

Gluckman, M. 1955. *The Judicial Process Among the Barotse of Northern Rhodesia.* Manchester, UK: University Press of the Rhodes-Livingston Institute.

Hadfield, G. K. 2000. Privatizing commercial law: Lessons from the middle and the digital ages. Stanford Law School, John M. Olin Program in Law and Economics, Working Paper No. 195.

Hafner, K. 2004a. Delete: Bathwater. Undelete: Baby. *New York Times*, August 5, p. G1.

Hafner, K. 2004b. With Internet fraud up sharply, eBay attracts vigilantes. *New York Times*, March 20, p. A1.

Hansell, S. 2004a. Junk e-mail and fraud are focus of crackdown. *New York Times*, August 25, p. C1.

Hansell, S. 2004b. U.S. tally in online-crime sweep: 150 charged. *New York Times*, August 27, p. C1.

Hardin, Garrett. 1968. The tragedy of the commons. *Science* 162:1243-48.

Hart, H. L. A. 1961. *The concept of law.* Clarendon Law Series. Oxford, UK: Clarendon Press.

Hayek, Friedrich A. von. 1937. Economics and knowledge. *Economica* 4:33-54.

Hayek, Friedrich A. von. 1973. *Law, legislation, and liberty*, vol. 1, *Rules and order.* Chicago: University of Chicago Press.

Houser, Daniel, and John Wooders. 2000. Reputation in auctions: Theory and evidence from eBay. Working paper, University of Arizona.

Hume, David. 1957 [1751]. *An Inquiry Concerning the Principles of Morals: With a supplement: A dialogue*, edited by Charles W. Hendel. The Library of Liberal Arts, No. 62. Indianapolis, IN: Bobbs-Merrill Company, Inc.

Ioffe, Olimpiad S. 1996. The System of civil law in the new commonwealth. In *The revival of private law in Central and Eastern Europe: Essays in honor of F. J. M. Feldbrugge*, edited by George Ginsburgs, Donald D. Barry, and Willaim B. Simons. The Hague: Martinus Nijhoff, pp. 79-97.

Ives, Nat. 2004. Yet another deal offers evidence that Internet marketing is hot again. *New York Times*, December 2, p. C2.

Jamal, K., M. Mairer, and S. Sunder. 2003. Enforcement standards versus evolution by general acceptance: A comparative study of e-commerce privacy disclosure and practice in the U. S. and the U. K. AEI-Brookings Joint Center for Regulatory Studies, Working Paper 03-8.

Jin, G. Z. and A. Kato. 2004. Blind trust online: Experimental evidence form baseball cards. Working paper, University of Maryland.

Johnsen, D. B. 1986. The formation and protection of property rights among the southern Kwakiutl Indians. *Journal of Legal Studies* 15:41-68.

Johnson, D. R. and D. Post. 1996. Law and borders—The rise of law in cyberspace. *Stanford Law Review* 48:1367-1402.

Johnson, R. N. 2004. Native American casinos: Another tragedy of the commons? Working paper, Property and Environment Research Center. Written for inclusion in *Self Determination: The Other Path for Native Americans*, edited by T. L. Anderson, B. L. Benson, and T. Flanagan. Stanford University Press, Forthcoming.

Johnson, R. N. and G. D. Libecap. 1982. Contracting problems and regulation: The case of the fishery. *American Economic Review* 72:332-47.

Katch, E., J. Rifkin, and A. Gaitenby. 2000. E-commerce, e-disputes, and e-dispute resolution: In the shadow of 'eBay law.' *Ohio State Journal on Dispute Resolution* 15:705-34.

Kesan, J. P. 2003. Private Internet governance. *Loyola University Chicago Law Journal* 35:87-137.

Khanna, T. and J. Rivkin, J. 2000. Ties that bind business groups: Evidence from an emerging economy. Working paper, Harvard Business School.

Kingson, J. A. 2004. Banks test ID devise for online security. *New York Times*, December 24, p. C2.

Kirzner, I. M. 1985. *Discovery and the Capitalist Process*. Chicago: University of Chicago Press.

Klein, B. and K. Leffler. 1981. The role of market forces in assuring contractual performance. *Journal of Political Economy* 89:615-41.

Klein, D. B. 1992. Promise keeping in the great society: A model of credit information sharing. *Economics and Politics* 4:117-136.

Krim, Jonathan. 2002. The Internet gets serious: Security, copyright problems, must be resolved as the medium matures. *Washington Post*, June 19. Available at: http://www.washingtonpost.com/ac2/wp-dyn?pagename=article&node=&contentId=A6168-2002Jun18.

Krim, Jonathan. 2004. Gates wants to give e-mail users anti-spam weapons. *Washington Post*, January 27. Available at: http://www.washingtonpost.com/ac2/wp-dyn/A50575-2004Jan26.html.

Lazarus, S., J. J. Bray, Jr., L. L. Carter, K. H. Collins, B. A. Giedt, R. V. Holton, Jr., P. D. Matthews, and G. C. Willard. 1965. *Resolving Business Disputes: The Potential for Commercial Arbitration.* New York: American Management Association.

Lew, J. 1978. *Applicable Law in International Commercial Arbitration: A Study in Commercial Arbitration Awards*, Dobbs Ferry, NY: Oceana Publications.

Libecap, G. D. 1978. Economic variables in the development of law: The case of western mineral rights. *Journal of Economic History* 38:338-62.

Libecap, G. D. 1986. Property rights in economic history: Implications for research. *Explorations in Economic History* 23:227-52.

Libecap, G. D. 1989. *Contracting for Property Rights.* Cambridge: Cambridge University Press.

Llewellyn, K. N., and E. A. Hoebel. 1961. *The Cheyenne Way.* Norman, OK: University of Oklahoma Press.

Lucking-Reilly, D., D. Bryan, N. Prasad, and D. Reeves. 2000. Pennies from eBay: The determinants of price in online auctions. Working paper, University of Arizona.

Lucking-Reiley, D., and D.F. Spulber. 2001. Business-to-business electronic commerce. *Journal of Economic Perspectives* 15:55-68.

Magid, L. 2004. Who's there? How parents can be IM watchdogs. *New York Times*, October 7, 2004, p. G6.

Markoff, J. 2005. Gates tells of Microsoft effort to fight viruses. *New York Times*, February 16, p. C15.

McDonald, Cynthia G., and V. Carlos Slawson, Jr. 2002. Reputation in an Internet auction market. *Economic Inquiry* 40:633-50.

McManus, J. 1972. An economic analysis of Indian behavior in the North American fur trade. *Journal of Economic History* 32:36-53.

McMillan, J., and C. Woodruff. 1998. Networks, trust, and search in Vietnam's emerging private sector. Working paper, Graduate School of International Relations and Pacific Studies, University of California, San Diego.

Melnik, Mikhail I., and James Alm. 2002. Does a seller's ecommerce reputation matter? Evidence from eBay auctions. *The Journal of Industrial Economics* 50:337-49.

Milgrom, Paul. R., Douglas C. North, and Barry R. Weingast. 1990. The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics & Politics* 2:1-23.

Mises, Ludwig von. 1985 [1957]. *Theory and history: an interpretation of social and economic evolution.* Auburn, AL: Ludwig von Mises Institute.

Mnookin, Robert H., and Lewis Kornhauser. 1979. Bargaining in the shadow of the law: The case of divorce. *Yale Law Journal* 88:950-97.

Murphy, Kate. 2004. EBay merchants trust their eyes, and the bubble wrap. *New York Times*, October 24, Section 3, p. 7.

Neely, Richard. 1983. *Why Courts Don't Work.* New York: McGraw-Hill.

Nelson, Phillip. 1974. Advertising as information. *Journal of Political Economy* 82:729-54.

Newman, Matthew. 2003. E-commerce (a special report): The rules—So many countries, so many laws: The Internet may not have borders; But the legal system certainly does. *Wall Street Journal*, 28 April, R8.

North, Douglass C. 1990. *Institutions, Institutional Change and Economic Performance*. Cambridge, UK: Cambridge University Press.

O'Brien, Timothy L. 2004. Gone in 60 seconds. *New York Times*, October 25, Section 3, p. 1.

O'Brien, Timothy L. and Saul Hansell. 2004. Barbarians at the digital gate, *New York Times*, September 19, Section 3, p. 1.

O'Driscoll, Gerald P. and Mario J. Rizzo. 1985. *The Economics of Time and Ignorance*. Oxford, UK: Basil Blackwell.

Pejovich, S. 1995. Privatizing the process of institutional change in eastern Europe. International Center for Economic Research, Working Paper No. 23/95.

Pejovich, S. 1997. Law, tradition, and the transition in eastern Europe. *The Independent Review* 2:243-54.

Phalon, R. 1992. Privatizing justice. *Forbes*, December 7, pp. 126-27.

Pospisil, L. 1971. *Anthropology of Law: A Comparative Theory*. New York: Harper and Row.

Pospisil, L. 1978. *The Ethnology of Law*. 2nd Edition. Menlo Park, CA: Cummings Publishing Co.

Post, D. G. 1996. Governing cyberspace. *Wayne Law Review* 43:155-71.

Powell, B. 2005. Don't regulate cyber security. *San Diego Union-Tribune*, March 2. Available at: http://www.signonsandiego.com/uniontrib/ 20050302/news_lz1e2powell.html.

Ray, L. 1992. Privatization of justice. In *Privatizing the United States Justice System: Police Adjudication, and Corrections Services form the Private Sector*, edited by G. W. Bowman, S. Hakim, and P. Seidenstat. Jefferson, NC: McFarland & Company.

Resnick, P. and R. Zeckhauser. 2002. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In *Advances in Microeconomics*, vol. 11, edited by M. R. Baye. Amsterdam: Elsevier Science.

Resnick, Paul, Richard Zeckhauser, John Swanson,and Kate Lockwood. 2003. The value of reputation on eBay: A controlled experiment. Harvard University Kennedy School of Government Faculty Research, Working Paper RWP03-007.

Richtel, M. 2004a. An industry that dares not meet in the country of its best customers. *New York Times*, May 17, p. C4.

Richtel, M. 2004b. Trade group says U.S. ban on net gambling violates global law. *New York Times*, March 26, p. C5.

Richtel, M. 2004c. U.S. steps up push against online casinos by seizing cash. *New York Times*, May 31, p. C1.

Ridley, M. 1996. *The Origins of Virtue: Human Instincts and the Evolution of Cooperation.* New York, NY: Viking Penguin.

Rustad, M. L. 2001. Private enforcement of cybercrime on the electronic frontier. *Southern California Interdisciplinary Law Journal* 11:63-116.

Rutten, A. 1997. Anarchy, order and the law: A post Hobbesian view. *Cornell Law Review* 82:1150-64.

Schwartz, John. 2004. Frontier justice; On the web, vengeance is mine (and mine). *New York Times*, March 28, Section 4, p. 1.

Shapiro, C. 1982. Consumer information, product quality, and seller reputation. *Bell Journal of Economics* 13:20-35.

Shapiro, C. 1983. Premium for high quality products as returns to reputation. *Quarterly Journal of Economics* 98:659-80.

Sherman, L. W. 1983. Patrol strategies for police. In *Crime and Public Policy*, edited by James Q. Wilson. San Francisco: Institute for Contemporary Studies.

Smith, T. 2003. Brazil becomes a cybercrime lab. *New York Times*, October 27, p. C4.

Solvason (Runolfsson), B. T. 1992. Ordered anarchy: Evolution of the decentralized legal order in the Icelandic commonwealth. *Journal des Economistes et des Etudes Humaines* 3:333-51.

Solvason (Runolfsson), B. T. 1993. Institutional evolution in the Icelandic commonwealth. *Constitutional Political Economy* 4:97-125.

Staples, B. 2004. The hard way to learn that the Internet is not Disneyland. *New York Times*, February 8, Section 4, p. 14.

Stross, R. 2005. How to stop junk e-mail: Charge for the stamp. *New York Times*, February 13, Section 3, p. 5.

Strumpf, K. 2004. Why prohibitions of Internet gambling won't work. *Cato TechKnowledge*, Issue 72. Available at: http://www.cato.org/tech/ tk/040123-tk.html.

Taylor, M. 1982. *Community, Anarchy and Liberty.* Cambridge, UK: Cambridge University Press.

Tedeschi, B. 2004a. Online auctions: Reducing risks. *New York Times*, January 4, Section 5, p. 2.

Tedeschi, B. 2004b. E-commerce report; After many have fallen by the wayside, retailers have begun to find a profit in online sales. *New York Times*, May 31, p. C5.

Thierer, A. D. 2001. Web restrictions unlikely to muzzle neo-Nazi speech. *Cato Institute, cato.org*, January 15. Available at: http://www.cato.org/cgi-bin/scripts/printtech.cgi/dailys/oi-15-01.html.

Thierer, A. D., and C. W. Crews, Jr. 2002. Internet libel ruling: Talk about a kangaroo court. *Cato Institute TechKnowledge*, Issue 46. Available at: http://www.cato.org/tech/tk/021216-tk.html.

Thierer, A. D., C. W. Crews, and T. Pearson. 2002. Birth of the digital new deal: An inventory of high-tech pork-barrel spending. *Cato Institute Policy Analysis*, No. 457.

Thompson, C. 2004. The virus underground. *New York Times*, February 8, Section 6, p. 10.

Thompson, N. 2003. More companies pay heed to their 'word of mouse' reputation. *New York Times*, June 23, p. C4.

Trakman, L. 1983. *The Law Merchant: The Evolution of Commercial Law.* Littleton, CO: Fred B. Rothman and Co.

Tullock, G. 1967. The welfare costs of tariffs, monopolies and theft. *Western Economic Journal* 5:224-32.

Tullock, G. 1985. Adam Smith and the prisoners' dilemma. *Quarterly Journal of Economics* 100:1073-81.

Umbeck, John. 1977. The California gold rush: A study of emerging property rights. *Explorations in Economic History* 14:197-226.

Umbeck, John. 1981a. *A Theory of Property Rights: With Application to the California Gold Rush.* Ames: Iowa State University Press.

Umbeck, John. 1981b. Might makes rights: A theory of the formation and initial distribution of property rights. *Economic Inquiry* 19:38-59.

Vanberg, V. J. and Buchanan, J. M. 1990. Rational choice and moral order. In *From Political Economy to Economics and Back?*, edited by J. H. Nichols Jr. and C. Wright. San Francisco: Institute for Contemporary Studies.

Vanberg, Viktor J., and Roger D. Congleton. 1992. Rationality, morality and exit. *American Political Science Review* 86:418-31.

Vaughn, K. I. 1994. *Austrian Economics in America: The Migration of a Tradition.* New York: Cambridge University Press.

Voight, S. and Kiwit, D. 1998. The role and evolution of beliefs, habits, moral norms, and institutions. In *Merits and Limits of Markets*, edited by H. Giersch. Berlin: Springer-Verlag.

Ware, Stephen J., and Sarah R. Cole. 2000. Introduction: ADR in cyberspace. *Ohio State Journal on Dispute Resolution* 15:589-95.

AOL is blocking spam. *Washington Post*, January 5, 2004, p. E02

Wesson, R. G. 1978. *State Systems: International Pluralism, Politics, and Culture.* New York: Free Press.

Williamson, O. E. 1983. Credible commitments: Using hostages to support exchange. *American Economic Review* 83:519-40.

Williamson, O. E. 1991. Economic institutions: Spontaneous and intentional governance. *Journal of Law, Economics, and Organization* 7:159-87.

Yen, A. C. 2002. Western frontier or feudal society? Metaphors and perceptions of cyberspace. *Berkeley Technical Law Journal* 17:1207-63.

Zeller, T., Jr. 2005. Law barring junk e-mail allows a flood instead. *New York Times*, February 1, p. A1.

FROM IMPERIAL CHINA TO CYBERSPACE: CONTRACTING WITHOUT THE STATE

David D. Friedman, Ph.D.[*]

I. THE PAST AS PROLOGUE

In 1895, as part of the treaty of Shimonoseki, China ceded the island of Taiwan to Japan. The Japanese government wished to maintain the existing legal system; in order to do so it had to discover what that legal system was. A scholarly commission was established, and its report provides us with a detailed picture of the legal system of at least one province of Imperial China at the end of its last dynasty (Brockman 1980, p. 130).[1]

One feature of that legal system was the combination of elaborate contractual practice with an almost total absence of contract law. Imperial China had no equivalent of our civil lawsuits. A merchant who had sold goods on credit and not been paid could, if he wished, report his debtor to the district magistrate for the crime of swindling him—but once he had done so, the case was out of the merchant's hands. The magistrate, if convinced of the justice of the claim, might compel repayment—usually only partial repayment. He might do nothing. He might even conclude that the merchant was the one at fault and sentence him to a beating. The legal system enforced by the magistrate focused almost entirely on criminal acts and criminal punishment, with only a handful of provisions dealing with matters of contract (Brockman 1980, p. 85),[2] and some, such as the statute specifying a maximum interest rate, appear to have been ignored in practice.

---

* Professor of Law, Santa Clara University. I would like to thank Bruce Benson for permitting me to read a manuscript of his which makes some of the same argument as this article from a somewhat different perspective. I have felt free to avail myself of his references where they were relevant to my argument, and have included a number of relevant articles by Benson in the list of references at the end of this piece. An earlier version of part of this chapter was published in the *Journal of Internet Law* (Friedman 2002).

[1] "The major publication in the area of customary law was *Taiwan Shiho* [the Private Law of Taiwan] (1910), a six-volume work which reprinted and analyzed documents pertaining to land law, family law, personal property and commercial law . . . with seven volumes of reference materials . . . ."

[2] "Of the 346 statutes in the Code, only eight dealt at all with what is usually called commercial law."

The Chinese empire relied heavily on non-state hierarchical structures to maintain order and settle disputes, most notably the extended family, and it supported them in doing so (Bodde and Morris 1967). That provided a possible mechanism for settling contract disputes within family, clan or guild. But merchants in Taiwan engaged in extensive large scale dealings that cut across all such categories, buying bulk agricultural products to ship across the straits to be sold in the mainland, importing mainland products to Taiwan, and much else.

The problem of settling commercial disputes without state courts was dealt with in medieval Europe, in part, by the development of private courts at the major trade fairs, run by merchants and relying heavily on reputational enforcement (Benson 1998c). No equivalent seems to have developed in China, perhaps due to Imperial hostility to any rival authority.

Nonetheless, Chinese merchants developed an elaborate set of contractual forms, including a variety of form contracts, supporting an extensive and sophisticated network of commercial relations. Part of the explanation of how they did so was presumably the existence of reputational enforcement, part the availability of state courts for dealing, when all else failed, with parties engaged in deliberate and obvious violations. But much of the explanation lies in the details of the private contract law that developed within that framework—a system of rules designed to minimize the reliance on courts and external enforcement.

One example is the rule that we call *caveat emptor*. Under any circumstances short of clear and deliberate fraud—gold bars that turned out to be gold plated lead, for example—a merchant who had accepted delivery of goods had no recourse if they turned out to be defective. Another is the linkage between possession, ownership, and responsibility; goods in my warehouse were mine, whether or not they were about to become yours, and I bore the risk of any damage that occurred to them.[3] The rules appear to have been designed, wherever practical, to let a loss lie where it fell, thus eliminating the need for legal action to shift it.

Problems arise in situations where canceling a contract and leaving everything in the possession of whoever, at the moment, has it will advantage one party, a situation that encourages opportunistic breach. One solution is to redesign the contract so that the two parties' performance is more nearly synchronized, reducing the incentive of either to breach. An alternative is to rely on reputational enforcement, structuring the contract so that the incentive to breach, if it occurs, is likely to be on the party who will suffer reputational penalties from breaching.

An example in the Chinese case is provided by contracts for future purchase of commodities at a pre-arranged price. Such contracts were not considered binding until there had been at least partial performance by one

[3] There were a few exceptions—most notably for a dye shop that would have cloth in it to be dyed.

party. Typically, that consisted of a deposit paid in advance by the purchaser. By adjusting the size of the deposit, the parties could take account of both how large the incentive of the seller to breach might become—depending on the range of likely price changes between contract formation and delivery—and how much each party was constrained to keep to the deal by reputation.

A buyer who breached forfeited his deposit—a result that required no judicial intervention, since the deposit was in the possession of the seller. That left an obvious problem—a seller who breached but kept the deposit. Presumably that was prevented by some combination of reputation and the threat that such an obviously criminal act would provide the buyer sufficient grounds for going to court.

Important elements in making the system work were the existence of a system of written forms using standard boilerplate terminology understood by the parties and others in the trade, and the use of seals—"chops"—to provide clear evidence of assent to a contract. So long as issues of fact were simple—whether a shipment of grain had been delivered and accepted, but not the precise quality or quantity—it was possible for third parties to determine, at a low cost, which party to a contract had violated its terms. Here the third party might be either another merchant interested in knowing who could be trusted or, in extreme cases, a district magistrate interested in who had committed a criminal offense and should be punished accordingly.

Whatever the mechanisms responsible—interested readers can find a more detailed account in Brockman's chapter—Chinese merchants a century ago succeeded in maintaining a sophisticated system of contracts with very nearly no use of state enforcement. It is the thesis of this paper that the past of China is our future—that parties to online transactions will, over the next few decades, face essentially the same problem and find, *mutatis mutandis*, similar solutions.

Both the Chinese past and the cyberspace future are special cases of a more general problem—contract enforcement in the absence of state enforced contract law. That problem appears in a variety of other contexts, including criminal markets and political markets. Perhaps less obviously, it appears in markets where court enforcement, although legally possible, is impractical because performance is difficult or impossible to monitor. The marriage market is an important example. For instance . . . .

Al-Tannuhki, a 10th century judge, tells the story of a vizier who gave a large sum in alms, 200 dinar, to a poor woman. Three days later he received a petition from the woman's husband, reporting that she had decided she was now too rich to be married to a poor man like him and was threatening to force him to divorce her. The husband asked the vizier to appoint some man in authority to prevent his wife from doing so. The vizier considered the problem briefly, took out paper and pen, and wrote "pay this man 200 dinar" (Margoliouth 1922).

It is possible for state courts to enforce rules permitting or restricting divorce. It is a great deal harder for them to enforce the other terms, explicit or implicit, of the marriage contract—to sanction someone for not doing a good job of being a wife or husband, even if the failure is deliberate. Given that difficulty, legal rules designed to punish one party for explicitly breaching the contract by getting a divorce may merely give that party an incentive to breach the less observable terms of the contract in order to make it in the other party's interest to agree to terminate it. That problem appears to have existed even in a medieval society whose marriage law was, on the face of it, heavily biased in favor of the husband.

Part II of the paper presents a general approach to private contract enforcement, some features of which are illustrated in the Chinese example. Part III sketches out the reasons why I expect that, for transactions in cyberspace, state enforcement of contracts will work worse and reputational enforcement better than in realspace today, including the technologies that provide an online equivalent of the seals used by Chinese merchants to establish the identity of a signatory party at a distance, in time or space. I go on to discuss how, in that environment, parties might structure their dealings, as well as the difficulties they will face due to the special nature of the cyberspace environment.

II. ENFORCING CONTRACTS WITHOUT THE STATE

Two parties wish to form a contract in a context where enforcement through a state court system is not a practical option. One simple way of doing so is the silent auction, for which we have descriptions going back to the sixth century B.C.[4] One party piles up the goods he wishes to sell, the other makes a matching pile of what he offers in exchange. If the offer is acceptable, the first party takes the second pile and leaves the first, and if not, the first party adjusts his offer. The process continues until one party accepts the other's most recent offer.

No common language is required for this simplest form of auction, but the parties still need some way of enforcing their property rights, of preventing one of them from taking both piles and departing. That might be either the threat of violence or the discipline of repeated dealings—the expectation that if one party acted that way this year, the other would not show up next year.

Difficulties arise when what the parties are contracting for is performance, by one or both, spread out over a period of time. Lloyd Cohen has discussed that problem in the context of modern marriage law, where the combination of a shift to no fault divorce and a pattern of traditional marriage within which the wife's performance of her part of the joint duties

[4] By Cosmas Indicopleustes and in the fifth century B.C. by Herodotus.

was concentrated in the early years of the marriage, the husband's more heavily weighted towards the later years, provided an incentive for opportunistic breach by the husband (Cohen 1987).

A partial solution in that case was for the wife to postpone childbearing and shift part of the cost of child rearing from household to market, thus aligning her performance more closely with the husband's. In a less intimate context, contractors building a house expect to receive payments spaced out over the time of construction and at least roughly corresponding to the spacing of costs, reducing the incentive of a contractor paid in advance to skip out with the money or a home owner promising to pay on completion to renege.

This kind of solution works reasonably well as long as the joint gains from final completion of the contract are substantial relative to the costs. Consider, however, the limiting case in the other direction—a situation where the gain to one party from breach is just equal to the loss to the other.

In such a situation, any departure from perfectly synchronized performance gives one party or the other an incentive to breach. If I have paid the contractor a little more than he has spent so far, he has an incentive to breach; if I have paid him a little less, I do. A more realistic example, and one which seems to have been a serious issue in the Chinese case, is a contract for future delivery at a pre-agreed price. If the transaction cost of arranging a replacement supplier is low, any significant drop in price provides an incentive for the buyer to breach; if the costs of finding a replacement buyer is low, any significant increase provides an incentive for the seller to breach. The parties can guard against breach by the buyer by having him pay a deposit in advance, but that increases the incentive for breach by the seller.

One solution is to create an artificial gain from completion—a cost to breach—by making it possible for the victim of breach to unilaterally impose a large cost on the other but not a correspondingly large benefit on himself. The deposit is replaced by a hostage. The threat of destroying the hostage reduces the gain to breach by the party who has given the hostage without creating a proportional increase in the gain to breach by the party holding the hostage. The logic of the situation is illustrated by Figures 1a-1c.

Figure 1a shows the situation with neither deposit nor hostage. The horizontal axis is time, starting just after the contract is negotiated. The vertical axis shows, for each party, its gain to breach—how much better off it will be if it breaches at that time than if the contract is completed. At time zero, the parties have negotiated the contract but no performance has occurred and no deposit has been made. Assuming that there was some cost to negotiation, which the parties expected to at least recover on completion of the contract, both parties should be worse off breaching at that point than carrying the contract to completion.

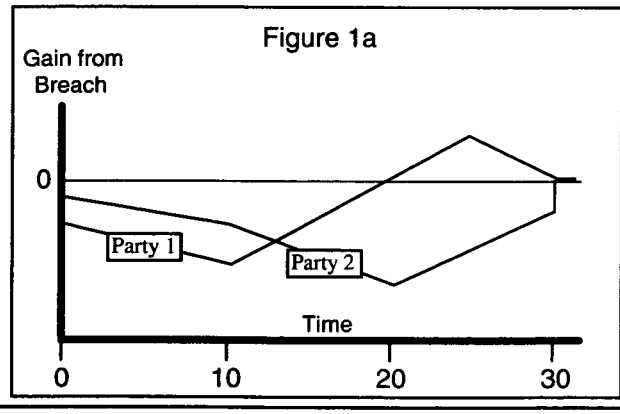
Over time, each party bears costs of performance which it expects to recover on completion, pushing the gain from breach further down, as shown between t=0 and t=10. At some point, possibly before completion, each party starts to get benefits from the partial performance that has occurred, increasing the gain (i.e. reducing the loss) from breach. In the figure, Party 1 is continuing to bear costs of performance from t=10 to t=20. Party 2 is no longer bearing costs of performance but is receiving benefits. So the gain to breach is falling for Party 1 but rising for Party 2.

If the contract is badly designed, at some point Party 1's benefit to breach rises above zero. In a world with no enforcement of contracts, legal or reputational, Party 1 breaches—perhaps immediately, perhaps with a delay to let the gain rise even higher. In the Figure, that happens at t=20. Party 1's gain is smaller than Party 2's loss, so the breach is inefficient.

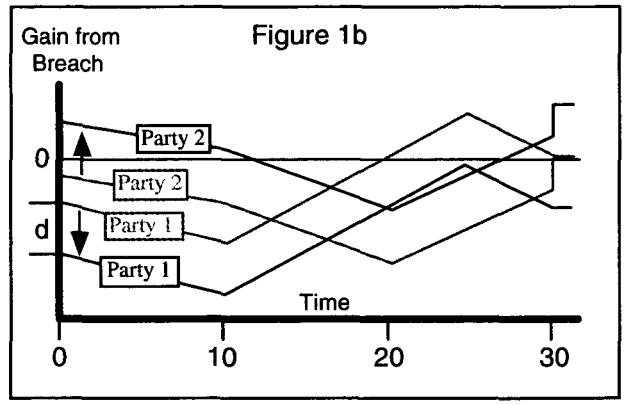
The parties can try to avoid such an outcome in the initial agreement by having Party 1 make payments to Party 2 between t=10 and t=20, shifting some of the gain from breach and keeping both parties' gain from breach negative. But in an uncertain and imperfectly observable world, this may not always work, since the parties do not know with certainty what the pattern of either performance costs or benefits will be.

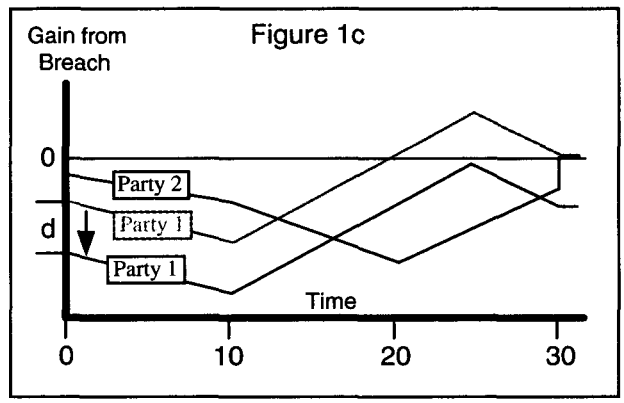
Figure 1b shows the same contract, with one change—at t=0, Party 1 pays a deposit *d* to Party 2. That shifts Party 1's gain from breach down, since if the contract does not go to completion the deposit will remain with Party 2; it is now no longer in Party 1's interest to breach the contract. Unfortunately, it also shifts Party 2's gain from breach up, since Party 2 can breach and keep the deposit. The result is that it is now in the interest of Party 2 to breach the contract—in this example, immediately after signing it and receiving the deposit.[5]

Figure 1c shows the same contract again, this time with Party 1 giving Party 2 a hostage rather than a deposit. The result is to shift Party 1's gain from breach down without shifting Party 2's gain from breach up, so neither party has an incentive to breach the contract.



[5] The situation without the deposit or hostage is shown by the grey lines on figures 1b and 1c.





One example of this approach is the literal hostage, offered by one party in a conflict as a guarantee that he will abide by terms agreed to. For a less obvious equivalent, consider the role of the state court system in Taiwan. If a merchant failed to either deliver the goods contracted for or return the buyer's deposit, the buyer could report him to the district magistrate as a swindler. The resulting legal case might or might not provide any benefit to the buyer but was likely to impose large costs on the seller. From the standpoint of the strategic situation of the two parties, the threat to make use of the court served the same function as the threat to execute a hostage.

A similar situation arises in a world without courts, but in which parties are concerned about their reputation. When you cheat me, I gain very little by making the fact public. But, assuming my report is credible, you lose a great deal. That fact, combined with a credible commitment on my part to report breach, increases the net cost of breach and so reduces the risk

that breaching will be in the interest of one of the parties.[6] Both this case and the previous have an additional attractive feature: the existence of the hostage depends at least in part on the existence of breach, since the accusation of breach, whether to the court or to the general public, is more damaging if true.

For a final example, consider the enforcement of contractual agreements in modern criminal markets. Antonio pays Ricco $100,000 for a large container of what turns out to be talcum powder. Antonio pays a hit man an additional $10,000 to kill Ricco—after first prudently convincing the local *capo* of the justice of his case in order not to get a reputation as a dangerous man to do business with but only a dangerous man to cheat. Here Ricco is, in effect, hostage for his performance—and Antonio for his.

It may occur to readers familiar with Coase that there is a problem with the use of hostages as a solution to the problem of opportunistic breach. The son I give as a hostage may be of no value to you, but he is of considerable value to me. If you decide to breach our agreement, you also inform me that you will kill him unless I buy him back from you for a suitable price. Similarly, if we are merchants in Taiwan, you breach the contract and then offer to buy my silence. You thus convert the hostage back into a deposit, eliminating the wedge between the terms on which it pays me to breach and the terms on which it pays you to.

Nonetheless we observe the use of hostages in contexts where a deposit would be of little use. One explanation is that parties who wish to give hostages are able to commit themselves not to accept such offers. Another is that the situation sets up a bilateral monopoly bargain with a large bargaining range and such bargains are likely to generate substantial transaction costs. That said, the issue deserves further thought.[7]

The discussion of hostages brings us to the last and, for this purposes of this paper, most interesting mechanism for enforcing contracts without the state—reputation.

Reputational Enforcement

For a simple example of reputational enforcement, consider a department store that guarantees to refund your money if you are not satisfied. If, when you discover that the jacket you bought is the wrong size and your

---

[6] This point was suggested to me by R.C. Friedman.

[7] This problem parallels a similar issue in the use of inefficient punishments, such as imprisonment, in criminal law as a way of reducing the risk of setting off a rent seeking struggle as some people attempt to use control of the criminal law to expropriate others. Thus we sometimes observe an inefficient punishment converted into a less inefficient punishment when police or prosecutors let one criminal off in exchange for testimony (true or false) that will allow them to convict another (Friedman 1999).

wife points out that purple is not really your color, the store refuses to give you a refund, you are unlikely to sue them—the amount at stake is not enough to make it worth the time and trouble. Nonetheless, almost all stores in that situation will, at least in my experience, take the product back—because they want the reputation, with you and with other people you may discuss the incident with, of living up to their promises.

For a more elaborate example, consider the New York diamond industry, as described in a classic article by Lisa Bernstein (Bernstein 1992). At one point, somewhat before the time she studied it, the industry had been mostly in the hands of orthodox Jews, forbidden by their religious beliefs from suing each other. They settled disputes instead by a system of trusted arbitrators and reputational sanctions. If one party to a dispute refused to accept the arbitrator's verdict, the information would be rapidly spread through the community, with the result that he would no longer be able to function in that industry. The system of reputational enforcement survived even after membership in the industry became more diverse, with organizations such as the New York Diamond Dealer's Club providing both trusted arbitration and information spreading.

The reason the department store, or the dishonest diamond merchant, is concerned about his reputation is not fear of being disliked but of losing business. The reason your friend will shop at another store if you tell him that this one refused to take your jacket back is not that he wishes to punish the store for cheating you but that he does not himself want to be cheated. Reputational enforcement works by spreading true information about bad behavior, information that makes it in the interest of some who receive it to modify their actions in a way which imposes costs on the person who has behaved badly.

How well that works depends on two things. One is the degree to which reputation matters; if I am a confidence man who plans to cheat you out of a million dollars and then retire, my future reputation is not very important. I don't care if anyone trusts me again. But most firms are in business for more than one transaction. Hence, for most firms, a reputation for cheating those they do business with is a costly liability.

The other critical variable is the cost to third parties of obtaining reliable information about what happened. In most disputes, both parties claim that they are in the right and the other in the wrong. When I tell my friend how badly the department store treated me he, hopefully, knows me well enough to decide whether or not to believe my story. But when I read a post on Usenet, a very large collection of online conversations, I do not have that sort of information about the author. I have to form my opinion based on internal evidence—does the poster sound reasonable—and consistency with other sources of information, such as other people posting in response.

If I claim that you cheated me, and you claim that I cheated you, a third party who cannot easily find out which of us is telling the truth is

likely to attribute some probability to both stories—and avoid doing business with either of us in the future. It follows that if you have cheated me, but I cannot easily demonstrate the fact to an interested third party, I may be better off saying nothing, since complaining will lower my reputation as well as yours. This raises a serious problem of incentive compatibility for a system which depends on action taken, not by government employees hired to enforce the law, but by private individuals acting in their own self interest.[8]

One way of lowering information costs to third parties is to have a legal system where the obligations of the parties depend on easily observed facts—loosely speaking, a system of bright line rules rather than standards. That appears to describe some features of the Chinese system already discussed.

For controversies with substantial amounts at stake, arbitration[9] provides a second mechanism for lowering information costs to interested third parties. A New York diamond merchant does not have to know the details of a controversy—merely the verdict of the arbitrator as to who was at fault and whether or not the party at fault provided suitable compensation to the injured party. That system works because, even if the interested third party does not know the details of the controversy, he does know that the arbitrator is competent and honest. As we will see, computer technology provides an equivalent that requires considerably less information and functions at even lower cost.

Cheating on the Reputational Bond

There is another problem, however, which is likely to be more serious for online commerce than for the traditional realspace version. My current reputation functions as a bond to guarantee performance—if I cheat on a contract, I lose (or reduce) the reputation, which is costly. It follows that I will not cheat unless the gain from doing so is more than the value of the forfeited bond. If my reputation is worth a million dollars to me, you should be safe in trusting me up to that sum—in, for example, lending me $600,000.

I borrow $600,000 from you. I also borrow $400,000 from another lender and $500,000 each from two more. I then default on all the loans, forfeiting my million dollar reputation—in exchange for $2,000,000.

---

[8] Friedman (2002), on which this article is in part based, includes a simple model of reputational enforcement showing the link between cost to third parties and the amount of cheating.

[9] Some readers may associate arbitration primarily with institutions for settling disputes that are selected only after the dispute arises. In this article, my primary interest is in arbitrators chosen in advance—by parties when they sign a contract that might lead to future disputes.

In order for reputational enforcement to work, the party who relies on it must have some way of knowing not only what opportunities the other party has to cheat him but what opportunities he has to simultaneously cheat other people. That may not be too much of a problem in the sort of ordinary market where most of the players know each other—before agreeing to lend me $600,000 you first discuss the situation with other potential lenders. But it could be a very serious problem for anonymous dealings online in a worldwide marketplace. In order for reputational enforcement to work in that setting, we need either an environment where the sort of opportunities made possible by my particular reputation are scarce enough so that I am unlikely to have a chance to take many at once, or procedures sufficiently transparent so that someone who relies on my reputation can know how many other people are currently doing so—how far I am stretching my reputational bond.

Whose Reputation?

A contract involves at least two parties, but they do not both have to have reputations. One is normally enough, since the parties can usually structure the contract to put the risk of breach on whichever can best bear it. If you are performing a service for me and I trust you but you do not trust me, because you are a repeat player with a reputation and I am not, I pay in advance. Reverse the situation and I pay on completion. In ordinary commerce, individual purchasers pay for goods when they get them, in the expectation that if the computer in the box turns out to have no innards, the store will take it back. The seller, in almost any field, is a repeat player with a reputation—the buyer often is not.

What about the situation where neither player has a reputation? In that case, they can solve the problem by bringing in a third party who does. An escrow agency provides a familiar example. I agree to pay you $50,000 for a sixteenth century painting by a known artist which you are offering on eBay. You deliver the painting to an escrow agency. I inspect it. If it fits the description I send you the money and claim the painting, if it does not the agency sends it back to you. Neither of us has to trust each other—only the agency. The mechanism does not depend on the existence of state courts to enforce the agreement, only on a third party with an adequate reputation.

III. PROBLEMS WITH PUBLIC ENFORCEMENT IN CYBERSPACE

Commercial activity in cyberspace, mostly on the World Wide Web, is increasing rapidly. Such commerce poses two rather different problems for conventional mechanisms of public contract enforcement. One, already important, is that cyberspace has no geographical boundaries. Purchasing

goods or services from the other side of the world is as easy as purchasing them from your next door neighbor. Delivery of physical goods is more costly from the other side of the world—but some cyberspace commerce is in information goods and services, and they can be delivered online just as they can be purchased online. It follows that an increasing fraction of commercial transactions, especially of transactions by private individuals, will be between parties in different countries.

Public enforcement of contracts between parties in different countries is more costly and uncertain than public enforcement within a single jurisdiction. Furthermore, in a world where geographical lines are invisible, parties to publicly enforced contracts will frequently not know what law those contracts are likely to fall under. Hence public enforcement, while still possible for future online contracts, will be less workable than for the realspace contracts of the past.

A second and perhaps more serious problem may arise in the future as a result of technological developments that already exist and are now going into common use. These technologies, largely based on public key encryption, make possible an online world where many people do business anonymously, with reputations attached to their cyberspace, not their realspace, identities (Friedman 1996).

There are a variety of reasons why people may in the future wish to avail themselves of such technologies. One is privacy; many people do not want others to know what they are reading, buying, or saying online.[10] A second is to evade taxes; it is hard for a government to collect taxes on activities it cannot see. A third is to evade regulations, whether commercial regulations in the U.S. or religious regulations in a country controlled by Muslim fundamentalists. Anonymity is likely to be particularly attractive to people living in parts of the world where property rights are insecure, making secrecy a valuable form of protection (Friedman 2004). If, for these or other reasons, a significant amount of commerce becomes anonymous, public enforcement of contracts will become increasingly irrelevant. It is hard to sue someone when you do not know who he is or what continent he lives on.

Private Enforcement of Contracts

What about the private alternative? At first glance, one might think that the same changes that made public enforcement of contracts more difficult in cyberspace would make private enforcement not only difficult but impossible. My local department store keeps its promises in part because if I am dissatisfied with their behavior, the people I talk to are likely to also be

[10] For a discussion both of the puzzle of why people favor more privacy, for others as well as themselves, and of the relation between privacy and technology, see Friedman (2000).

their customers; in a future without geography, where everyone is shopping everywhere, that is far less likely. And it is not obvious how you can injure someone's reputation without knowing his name.

Both of these problems are soluble; in each case, online commerce provides not merely substitutes for the reputational mechanisms with which we are already familiar, but superior substitutes.

Consider first the problem of getting information from one customer to another. Considered as a mechanism for spreading information, local gossip is very much inferior to a well designed search engine. If, today, I am considering dealing with an online merchant and want to know whether other customers have had problems with him, I do not bother to ask either friends or the Better Business Bureau. A one minute search with Google will tell me whether anyone on Usenet News has mentioned that firm any time in the past year, and show me what was said.

Online commerce is already institutionalizing such mechanisms. Consider eBay. Their software permits anyone who has won an auction to post comments on the seller—whether the goods lived up to their description, were delivered promptly, or whatever else he wants to say. The comments are available, both in summary form and in text, to anyone bidding in an auction with that seller.

So far I have been considering informal reputational enforcement, the online equivalent of the reputational mechanism that keeps your local department store honest. What about formal enforcement, along the lines of the diamond industry, as described by Bernstein (Bernstein 1992)? Here too, cyberspace has significant advantages over realspace.

Keys and Signatures: A Brief Digression

To explain how the cyberspace equivalents of arbitration by the Diamond Dealer Club of New York and verification by the use of seals in 19th century China work, I must first briefly sketch some relevant technology; readers already familiar with public key encryption and digital signatures may want to skip this section.

Public key encryption is a mathematical process for scrambling and unscrambling messages. It uses two keys, numbers containing information about a particular way of scrambling a message. The special feature of public key encryption is that if one of the two related numbers is used in the scrambling process, the other must be used in the unscrambling process. If I have one of the two keys I can encrypt my messages with that key, but someone who wishes to decrypt messages that have been encrypted with that key needs to use the other one. While the pair of keys is generated together, there is no easy way of calculating one of the two keys from the other.

To make use of public key encryption, one generates such a pair of keys. One, called your public key, you make available to anyone you might

be corresponding with. The other, called your private key, you keep entirely secret.

Someone who wants to send you a message encrypts it using your public key; since only you have the matching private key, only you can decrypt it. Someone who wants to digitally sign a message encrypts it using his private key[11] and attaches unencrypted information identifying himself. The recipient obtains the sender's public key and uses it to decrypt the message. The fact that what he gets is a message and not gibberish demonstrates that it was encrypted with the matching private key; since only the sender possesses that particular private key, the digital signature authenticates the message. Thus the digital signature in cyberspace serves the same function as the physical seal used to authenticate contracts in China a century ago—to prove authorship and responsibility at a distance in space or time.

Not only does a digital signature prove who sent the signed message, it also proves that the message has not been altered, and it proves both in a form that the sender cannot deny. If the sender tries to deny the message, the recipient can point out that he has a version of it encrypted with the sender's private key, something that only the sender could have produced.

Convincing Interested Third Parties

Imagine that you and I are signing a contract online, specifying our mutual rights and obligations for some substantial transaction. We include in the contract the name and public key of the arbitrator who we agree will settle disputes between us. We then both digitally sign the contract. Each of us gets a copy.

A dispute arises; I accuse you of violating the terms of the contract. We put the question to the arbitrator. He rules in my favor and instructs you to pay me $5000 in damages. You refuse. He writes up his account of what happened (he ruled in my favor and you refused to abide by his ruling), digitally signs it, and gives me a copy.

I now make up a package consisting of the original contract (digitally signed by both of us, and including the arbitrator's public key) and the arbitrator's account (digitally signed by him). I send the package to any third party who I think might want to know whether or not you are trustworthy—and post it on a web page with your name all over it, to be found by anyone

[11] The process used for digital signatures in the real world is somewhat more elaborate than this, but the differences are not important for the purposes of this article. A digital signature is produced by using a hash function to generate a message digest—a string of numbers much shorter than the message it is derived from—and then encrypting the message digest with the sender's private key. The process is much faster than encrypting the entire message and almost as secure. It also means that it is possible to read the message without bothering to check the signature.

searching for information about you. The third party (more precisely, his computer) checks the digital signatures on the contract and on the account, using the public key included in the contract to check that the account is by the arbitrator we agreed to. The third party now knows that you agreed to accept the ruling of that arbitrator and reneged on that agreement—and finding that out has taken him essentially no time at all.

Digital signatures provide a way of drastically reducing the cost to interested third parties of discovering whether someone is trustworthy.[12] By doing so, they increase the cost to individuals or firms engaged in repeat transactions of reneging on their contractual agreements.

Private enforcement of contracts along these lines solves the problems raised by the fact that cyberspace spans many geographical jurisdictions. The relevant law is defined not by the jurisdiction but by the private arbitrator chosen by the parties. Over time, we would expect one or more bodies of legal rules with regard to contract to develop, with many different arbitrators or arbitration firms adopting the same or similar legal rules.[13] Contracting parties could then choose arbitrators on the basis of reputation.

For small scale transactions, you simply provide your browser with a list of acceptable arbitration firms; when you contract with another party, the software picks an arbitrator from the intersection of the two lists. If there exists no arbitrator acceptable to both parties, the software notifies both of you of the problem and you take it from there.

Private enforcement also solves the problem of enforcing contracts when at least one of the parties is, and wishes to remain, anonymous. Digital signatures make it possible to combine anonymity with reputation. A computer programmer living in Russia or Iraq and selling his services online has an online identity defined by his public key; any message signed by that public key is from him. That identity has a reputation, developed through past online transactions; the more times the programmer has demonstrated himself to be honest and competent, the more willing people who want programming done will be to employ him. The reputation is valuable, so the programmer has an incentive to maintain it—by keeping his contracts.[14]

---

[12] Strictly speaking, what the third party learns is that the accused either is not trustworthy or has agreed to use a dishonest or incompetent arbitrator. The latter alternative implies that while the accused may not be dishonest, save in the very limited sense of refusing to be bound by his own mistake, he is incompetent.

[13] As Bruce Benson has pointed out, this development is closely analogous to the development of the Lex Mercantoria in the early Middle Ages. That too was a system of private law enforced by reputational penalties, in an environment where state law was inadequate for contract enforcement, due in part to legal diversity across jurisdictions (Benson 1998b,c).

[14] The first discussion of privacy through anonymity online of which I am aware of was in a work of fiction by a computer science professor, Verner Vinge's novelette "True Names." A good recent description of the combination of anonymity with online reputation occurs early in Marc Siegler's novel *Earthweb*.

Cheating in a Reputational System[15]

There are, unfortunately, ways in which the online world I have been describing makes contract enforcement harder than in the real world. One is that, in the real world, my identity is tied to a particular physical body, identifiable by face, finger prints, and the like. I do not have the option, after destroying my realspace reputation for honesty, of spinning off a new me, complete with new face, new fingerprints, and an unblemished reputation.

Online, I do have that option. As long as other people are willing to deal with cyberspace personae not linked to realspace identities, I always have the option of rolling up a new public key/private key pair and going online with a new identity and a clean reputation.

The implication is not that reputational enforcement will not work but that it will only work for people who have reputations—sufficient reputational capital so that abandoning the current online persona and its reputation is costly enough to outweigh the gain from a single act of cheating. Someone who wants to deal anonymously in a trust intensive industry may have to start small, building up his reputation to the point where its value is sufficient to make it rational to trust him with larger transactions. Presumably the same thing happens in the diamond industry today.[16]

The problem of spinning off new identities is not limited to cyberspace. Real persons in realspace have fingerprints but legal persons may not. The realspace equivalent of rolling up a new pair of keys is filing a new set of incorporation papers. There is a well developed literature on the result, explaining marble facing for bank buildings and expensive advertising campaigns as ways of posting a reputational bond that makes it in a corporation's interest to remain in business and hence gives others a reason to trust it to act in a way that will preserve its reputation (Nelson 1974; Williamson 1983; Klein and Leffler 1981). Cyberspace personae do not have the option of marble, at least if they want to remain anonymous, but they do have the option of investing in a long series of transactions, or advertising,

---

15 A firm that breaches a contract but pays damages according to the terms specified in the contract has not cheated in the sense in which I am using the terms. To cheat, it must both breach the contract and fail to pay any damages agreed on in advance or awarded by a pre-agreed upon arbitrator.

16 *Earthweb* contains an entertaining illustration of this point. A central character has maintained two online personae, one for legal transactions, with a good reputation, and one for quasi-legal transactions, such as purchases of stolen property, with a deliberately shady reputation. At one point in the plot, his good persona is most of the way through a profitable honest transaction when it occurs to him that it would be even more profitable if, having collected payment for his work, he failed, at the last minute, to deliver. He rejects that option on the grounds that having a persona with a good reputation has just given him the opportunity for a profitable transaction; if he destroys that reputation it will be quite a while before he is able to get other such opportunities.

or some other publicly visible expenditure, in order to bond future performance.

What if only one of the parties to an online contract is a repeat dealer with a reputation? The solution, as in realspace, is to structure the contract so that it is not in the other party's interest to breach it. The simplest example is the purchase of goods or services. The party who does not have a reputation performs first—pays in advance if he is the buyer, delivers in advance of payment if he is the seller.

We are left with an obvious problem—how can a pair of entities neither of which is engaged in long term dealings guarantee contractual performance in this world? One solution has already been mentioned—piggyback on the reputation of another entity that is engaged in such dealings.

I am, again, an anonymous online persona forming a contract which may provide me an opportunity to benefit by defaulting on my contractual obligations. This time, however, I have no reputation and no time in which to build one. Instead I offer to post a performance bond with the arbitrator—in anonymous digital currency,[17] assuming that I am seriously interested in protecting my own anonymity. The arbitrator is free to allocate all or part of the bond to the other party as damages for breach.

This approach still depends on reputational enforcement, but this time the reputation belongs to the arbitrator. If he steals bonds posted with him, he is unlikely to stay in business very long. If I am worried about such possibilities, I can require the arbitrator to sign a contract specifying a second and independent arbitrator to deal with any conflicts between me and the first arbitrator. My signature to that agreement is worth very little, since it is backed by no reputation—but the signature of the first arbitrator to a contract binding him to accept the judgment of the second arbitrator is backed by the first arbitrator's reputation. For a less extreme example of the same approach, consider the current use of escrow agencies for transactions on eBay.

As that final example suggests, it is possible to combine realspace and cyberspace institutions, state and private enforcement mechanisms. If court enforcement in realspace turns out to provide a more reliable mechanism than reputational enforcement online, anonymous online parties can use identifiable real space third parties as escrow agencies, arbitrators, and in other contexts in which a trusted third party eliminates the need for trust between the other parties to a transaction. If, on the other hand, courts prove less reliable, realspace parties can make use of online reputational mechanisms instead—as they now do.

As long as parties are identifiable in realspace, the state has the option of imposing its own terms on them—an option some parties may wish to

---

[17] For a discussion of how such currency would work, see Friedman and Macintosh (2001, 2003).

avoid. But anonymous parties in cyberspace who wish to make use of a trusted third party in realspace can choose which third party, and hence which state, they wish to deal with. States will thus be constrained by competition in their dealing with online personae, just as U.S. states are currently constrained in dealing with corporations.

One way of succeeding in that competition is to make it possible for online parties to take advantage of realspace enforcement without revealing their realspace identities. A possible approach would be for a state to recognize transfers of claims from cyberspace to realspace persons, validated by the former's digital signature. So if anonymous X has a valid claim against realspace Y, X sells the claim to realspace Z who prosecutes it—without either Z or the court having to know X's realspace identity.

One problem with reputational enforcement online is that a party can roll up additional identities. A second problem is that a party can conduct multiple transactions, each invisible to those party to the others. As discussed earlier, that means that a party with a million dollar reputation might put together a collection of transactions, each of which was not worth cheating on (and forfeiting the reputation) but which together were.

One solution is to have a million dollar reputation and engage in thousand dollar transactions in a context where one is unlikely to be able to run as many as a thousand of them at once. In realspace that is often practical. It may work less well in cyberspace, where the identity of the party behind a reputation, including how many actual persons that party consists of, may be unknown.

An alternative is for a party to deliberately create transparency in order that everyone who contracts with him will be aware of the existence (but not necessarily the identity) of everyone else currently contracting with him.

I wish to create an online identity, post a reputational bond, and be trusted. My identity consists not only of a public key but also of a transactional protocol—a set of rules associated with that identity and its reputation, specifying how people are to deal with me. The protocol is designed to enforce transparency.

For a simple example, let the protocol specify that all transactions become binding only when posted to a particular web page, publicly accessible. That way, anyone transacting with me can see how many other transactions I am engaged in and whatever relevant features of the transaction—the size of a loan, say—are specified in the protocol.

Reputation: Version Two

In the discussion so far, "reputation" meant "reputation for fulfilling your contracts." But there is another sort of reputation that is important in realspace—a reputation for competence in the activity you are performing for pay. When you hire a lawyer or a heart surgeon, it isn't enough to know

that he is honest. That sort of reputation can also be established in cyberspace—and there too, the special circumstances of cyberspace raise problems, but problems that have their parallel in realspace.

Suppose I claim to be an expert in predicting real world events that potential customers wish predicted—the weather, the outcome of a particular legal case, the performance of a stock. Just as in realspace, I can establish a track record by making a series of correct predictions. There is, however, a problem.

To see it, imagine that my claim is to be able to predict, with certainty, the outcome of coin tosses—which many potential customers want predicted. Some people have that ability but, unfortunately, I do not. I proceed as follows:

1. I obtain a list of 10,000 potential customers.

2. I create 128 identities, each of which claims to be an expert predictor of coin flips, and divide the potential customers among them.

3. The first time a flip is to be predicted, half my identities predict heads, half predict tails. The coin is flipped and comes up heads. I scrap all of the identities that predicted tails and remove their customers from my current list—retaining their names and email addresses for future iterations of my business plan.

4. I repeat the previous step six more times.

I now have one surviving identity with about forty customers. Each of them has seen that identity predict a coin flip correctly seven times in a row, an event that could happen by chance less than one time in a hundred. Predicting coin flips is valuable, so each should be willing to pay a sizable sum for the next prediction.

I have just described the cyberspace equivalent of the market for investment newsletters or mutual funds. The chief difference—leaving aside the simplification of my coin flipping model—is that in my version the multiple identities all belong to the same person, making the fraud a deliberate one. In the realspace case, the publishers of each investment newsletter or the administrators of each mutual fund may actually believe that they know what the market will do next—and, each time, about half of them are right.

How might someone who really did know how to predict coin flips distinguish himself from those who did not but who might attempt to simulate that ability as described? The obvious answer is again some form of bond. When I first go into business making (public) predictions of coin flips, I also donate $100 in e-cash to some popular charity that is willing to

testify to the receipt of the money from my online identity. I then make seven consecutive correct public predictions.

A hundred dollars is not very much money. But in order to follow the business plan described in steps 1-4 above, I needed 128 identities—which, at $100 per identity, gets expensive. Furthermore, in addition to selling my prediction of flip number eight to paying customers, I also post it on my web page—after the customers have gotten their bets down but five minutes before the coin is flipped. After another ten correct calls, a potential customer can calculate that either I know something, or I am fantastically lucky, or I am the sole survivor of a collection of identities that cost somewhat over twelve million dollars to create. The generalization to someone selling investment advice, legal advice or medical advice online is left as an exercise for the reader.

CONCLUSION

If the arguments I have offered are correct, we can expect to see a substantial shift in the direction of reliance on private enforcement via reputational mechanisms online, with an associated development of private law. To some degree, the same development can be expected in realspace as well. Digital signatures lower information costs to interested third parties whether the transactions being contracted over are occurring online or not. And the existence of a body of trusted online arbitrators will make contracting in advance for private arbitration more familiar and reliance on private arbitration easier for realspace transactions as well as for cyberspace transactions.

REFERENCES

Benson, B. L. (1998a). "Economic Freedom and the Evolution of Law." *Cato Journal.* 18, 209-232.

Benson, B. L. (1998b). "Evolution of Commercial Law." In P. Newman, (ed.). *The New Palgrave Dictionary of Economics and the Law*, London: Macmillan Press.

Benson, B. L. (1998c). "Law Merchant," In P. Newman, (ed.). *The New Palgrave Dictionary of Economics and the Law*, London: Macmillan Press.

Benson, B. L. (1998d). "How to SECEDE in Business Without Really Leaving: Evidence of the Substitution of Arbitration for Litigation." In D. Gordon. (ed.). *Secession, State, and Liberty.* New Brunswick, NJ: Transaction.

Benson, B. L. (1999). "To Arbitrate or to Litigate: That is the Question," *European Journal of Law and Economics.* 8, 91-151.

Benson, B. L. (2000). "Arbitration." In B. Bouckaert and G. De Geest. (eds.). *The Encyclopedia of Law & Economics*. London: Edward Elgar.

Bernstein, Lisa, "Opting Out of the Legal System: Extralegal Contractual Relations in the Diamond Industry," 21 *Journal of Legal Studies*, 1992, pp.115-157.

Bodde, Derk and Morris, Clarence, *Law in Imperial China*, Harvard University Press 1967.

Brockman, Rosser H., "Commercial Contract Law in Late Nineteenth-Century Taiwan," in Jerome Alan Cohen, R. Randle Edwards and Fu-mei Chang Chen, editors, *Essays on China's Legal Tradition*, Princeton University Press 1980, pp. 76-136.

Choi, Stephen J., "Gatekeepers and the Internet: Rethinking the Regulation of Small Business Capital Formation," *The Journal of Small and Emerging Business Law*, Volume 2 (Summer 1998) Number 1. http://www.lclark.edu/~lawac/LC/jsebl/summer98.htm.

Cohen, Lloyd, "Marriage, Divorce, and Quasi Rents; Or 'I Gave Him the Best Years of My Life.'" 16 *Journal of Legal Studies* 267-304 (1987).

Friedman, David, "A World of Strong Privacy: Promises and Perils of Encryption," *Social Philosophy and Policy* (1996), pp. 212-228. http://www.best.com/~ddfr/Academic/Strong_Privacy/Strong_Privacy.html.

Friedman, David, "Why Not Hang Them All: The Virtues of Inefficient Punishment," *Journal of Political Economy*, vol. 107, no. 6 1999 pp. S259-269.

Friedman, David, "Privacy and Technology," *Social Philosophy and Policy*. 17:2 (Summer 2000).

David D. Friedman and Kerry L. Macintosh, "The Cash of the Twenty-First Century," *Santa Clara Computer and High Technology Law Journal,* 17 Santa Clara Computer & High Tech. L.J. 273, 2001(273).

Friedman, David D. and Macintosh, Kerry L. "Technology and the Case for Free Banking," chapter in *The Half-life of Policy Rationales: How New Technology Affects Old Policy Issues*. Klein, D. and Foldvary, F., editors. New York University Press (2003).

Friedman, David, "The Case For Privacy" in *Contemporary Debates in Applied Ethics*, Blackwell (2004).

Friedman David, "Contracts in Cyberspace," 6 *Journal of Internet Law* 12 (Dec. 2002).

Klein, B. and Leffler, K. (1981). "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*. 89, 615-641.

Margoliouth, D.S. (tr), *The Table-Talk of a Mesopotamian Judge,* by al-Muhassin ibn Ali al-Tanukhi, Royal Asiatic Society (1922).

Nelson, P. (1974) "Advertising as Information," *Journal of Political Economy* 76, 729-754.

Siegler, Marc, *Earthweb*, (Baen Books: 1999).

Vinge, Verner, "True Names," included in *True Names and Other Dangers.* Baen (1987).

Williamson, O. E. (1983) "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review.* 83, 519-540.

THE CAPABILITY OF GOVERNMENT IN PROVIDING PROTECTION AGAINST ONLINE FRAUD: ARE CLASSICAL LIBERALS GUILTY OF THE NIRVANA FALLACY?

Edward Stringham, Ph.D.[*]

ABSTRACT

Online merchants are exposed to serious threats of fraud, which has the potential to cripple electronic commerce. Classical liberals such as Epstein and North believe that markets require prohibitions against fraud and that government can solve the problem. Although the classical-liberal solution seems clear, how it will be implemented is less clear. For government to prohibit online fraud a number of conditions must be met. By compiling evidence from government testimonies and interviews in Silicon Valley, this article studies the extent to which government can provide protections against online fraud. It finds a number of obstacles that inhibit government from enforcing laws against online fraud. Technology moves at a rapid pace and government often lacks the capability to identify those who commit fraud. In addition, questions remain about how domestic law enforcement can enforce laws against fraud around the globe. Even if domestic law enforcement had the ability to identify fraudsters, it would need to rely on law enforcement agencies from around the globe to help enforce the laws. Under these conditions the ability for government to prohibit fraud is extremely limited. Classical liberals appear to be guilty of the Nirvana Fallacy.

1. INTRODUCTION

Electronic commerce poses many potential dilemmas for consumers and businesses alike. In non-face-to-face transactions, consumers need to rely on merchants delivering the product and merchants need to rely on consumers delivering the payment. Although much attention has been paid

[*] Department of Economics, San Jose State University. Email: Edward.Stringham@sjsu.edu. The author thanks Peter Boettke, Dan Klein, Peter Leeson, Benjamin Powell, and participants at the Critical Infrastructure Project at George Mason University and at the Association of Private Enterprise Education meetings for helpful comments and suggestions. Research funding from the working group on Law, Economics and Technology of Private Enforcement on the Internet is greatly appreciated. The usual disclaimer applies.

to traditional consumer fraud,[1] merchants are perhaps in an even more difficult situation. Customers at least have the ability to look into the reputation of sellers,[2] whereas merchants have no such luxury. Merchants can check that a bank account has funds, but the order still might be placed with a stolen bank account.[3] Fraud often goes undetected until the cardholder notices his bill, well after the goods have shipped. When a transaction goes sour, the merchant usually has to foot the bill.[4] Even though commerce gives businesses access to many additional customers, it also exposes them to many perpetrators of fraud. In today's world, up to 40 percent of online international orders that merchants receive (but do not necessarily accept) are fraudulent, which has the potential to cripple electronic commerce.[5] If merchants have no recourse when fraud occurs and cannot easily distinguish between good and bad orders, they will end up acting cautiously and turning down a number of legitimate orders. Some merchants may even eschew electronic commerce altogether, and the market will not reach its full potential.

The problem of fraud is real, but what is the solution? Most lawyers and economists are influenced by classical liberal theory and look to government to step in. After all, prohibition against fraud is one of the core functions of government. For example, Microsoft General Counsel Bradford Smith stated, "So long as people use the Internet to perpetrate frauds, steal property, and defame and assault one another, governments will be justified in seeking to prevent such behavior through law."[6] The only people who would deny government such a role are anarchist libertarians who reject government altogether. Chicago Law Professor Richard Epstein provides a representative summary of the limited-government or classical-liberal view: "Under its classical liberal formulation, the great social contract sacrifices liberty, but only to the extent that it is necessary to gain security against force and fraud. Perhaps we might go further, but surely we

---

[1] Karen Alboukrek, *Adapting to a New World of E-Commerce: The Need for Uniform Consumer Protection in the International Electronic Marketplace*, 35 Geo. WASH. INT'L L. REV. 425 (2003); Miriam R. Albert, *E-Buyer Beware: Why Online Auction Fraud Should be Regulated*, 39 Am. BUS. L.J. 575 (2002).

[2] Boettke and Steckbeck document how online merchants can build up their reputation, which can be conveyed with review websites or rating systems such as on eBay. Peter Boettke & Mark Steckbeck, *Akerlof Problems and Hayek Solutions: Local Knowledge and Self-governance in E-Commerce*, *in* AUSTRIAN PERSPECTIVES ON THE NEW ECONOMY (Jack Birner ed., 2003), in press.

[3] Other ways consumers commit fraud against merchants is by disputing a bill, denying they made a transaction or by saying the goods arrived damaged.

[4] Cliff Ennico, Get Yourself Paid: Try these two techniques for dealing with deadbeat customers, Entreprenuer.com, July 07, 2003, http://www.entrepreneur.com/article/0,4621,309711,00.html.

[5] Jeff King, Seminar on Accepting International Orders in Real Time, (Cybersource, Inc.) (File author downloaded 2004).

[6] Bradford L. Smith, *The Third Industrial Revolution: Policymaking for the Internet*, 3 COLUM. SCI. & TECH. L. REV. 1 (2002).

go this far."[7] To Epstein, the government must perform certain roles such as providing law against fraud; otherwise, markets would be unable to function. In contrast to the anarchist libertarians, Epstein argues that one would be a "naïve visionary" to "believe that markets could operate of their own volition without any kind of support from the state."[8] He writes, "It is at this juncture that the rule of law becomes critical to offer a secure framework for these voluntary transactions to take place."[9] Similarly, Nobel Laureate Douglas North states that "realizing the economic potential of the gains from trade in a high technology world of enormous specialization and division of labor characterized by impersonal exchange is extremely rare, because one does not necessarily have repeated dealings, nor know the other party, nor deal with a small number of other people."[10] He concludes, "A coercive third party is essential."[11]

The idea that government is needed to enforce laws against fraud is held not only by classical liberals, but also by the vast majority of lawyers and economists as well.[12] Yet the idea is more of an assumption in economic and legal analysis, rather than a hypothesis which is subjected to investigation. The vast majority of lawyers and economists simply assume that government should prohibit fraud and do not give the issue another thought. Although the classical-liberal solution seems clear, how it will be implemented is less clear. Just because something is *de jure* illegal does not mean that an action is effectively prohibited. Passing a law pronouncing something illegal is easy but effective prohibition requires more than just official proclamations. Princeton economist Avinash Dixit states, "the problem is that [conventional economic theory] takes the existence of a

---

[7] Richard A. Epstein, *Hayekian Socialism*, 58 MD. L. REV. 271 (1999).

[8] *Id.* at 285.

[9] *Id.*

[10] DOUGLAS C. NORTH, INSTITUTIONS, INSTITUTIONAL CHANGE AND ECONOMIC PERFORMANCE 12 (1990).

[11] *Id.* at 35.

[12] Avinash Dixit writes, "Even the most libertarian economists, who deny the government any useful role in most aspects of the economy, allow that making and enforcing laws that give clear definitions of property rights, and ensuring adherence to voluntary private contracts, are legitimate and indeed essential functions of government." AVINASH DIXIT, LAWLESSNESS AND ECONOMICS 2 (2004). Similarly, South Carolina Law Professor Henry Mather maintains that even the most "extreme libertarian theories" still give government the "nightwatchman's task of protecting individual liberty against force and fraud" (332). Henry Mather, *Natural Law and Liberalism*, 52 S.C. L. REV. 331 (2001).

well-functioning institution of state law for granted."[13] In many cases, real world difficulties may make enforcing laws against fraud more difficult than economists and lawyers assume.

Pointing out the problem of fraud is simple but the real question is whether government is capable of solving the problem. One can believe that government has the ability to solve the problem, but that does not mean that the belief is true. In this sense, lawyers and economists might be falling into the trap of what Harold Demsetz called the Nirvana fallacy.[14] Many theorists highlight a problem in the world and then conclude that government can solve it.[15] But rather than jumping to the conclusion that the government has the ability to solve the problem, we must look to see if it really does.

Online merchants sold over $100 billion worth of goods in 2003,[16] and although numerous federal, state, and local agencies have computer divisions that aim to "stop perpetrators of fraud and deception,"[17] the extent to which the law actually helps merchants is unestablished.[18] Since it was passed in 1984, the Computer Fraud and Abuse Act (18 U.S.C. § 1030) has been criticized for being "overly vague and too narrow in scope,"[19] and "largely symbolic."[20] As late as 1996 there were only 174 convictions for computer fraud, which includes hacking, copyright infringement, and gambling fraud. The US Department of Justice writes, "experts have long admitted that there are no centralized computer crime statistics, not even within the law enforcement community."[21] We have to investigate, but we might have a case where the laws are on the books but are not really being enforced.

For the government to be able to stop online fraud, a number of conditions must be met. Former Attorney General Janet Reno highlighted some of the problems in a 2000 "five-year strategy" to develop enforcement ca-

---

13 DIXIT, *supra* note 13, at 3.

14 Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & ECON. 1 (1969).

15 An example of this is John Rothchild, *Protecting the Digital Consumer: The Limits of Cyberspace Utopianism*, 74 IND. L.J. 893 (1999).

16 Keith Regan, Report: Online Sales Top $100 Billion, E-Commerce Times, June 16, 2004.

17 Mozelle Thompson, *The Challenges of Law in Cyberspace—Fostering the Growth and Safety of E-Commerce Commissioner*, 6 B.U. J. SCI. & TECH. L. 1, (2000) Par. 9.

18 FTC Commissioner Mozelle Thompson stated, "it's not the "Wild, Wild West" out there. Fraud and deception for example in consumer protection, it does not matter whether it occurs on the telephone or on the Internet, it is still illegal." Mozelle W. Thompson, *The Federal Trade Commission and Regulating E-Commerce*, 16 ST. JOHN'S J. LEGAL COMMENT. 609 (2002).

19 Reid Skibell, Cybercrimes *& Misdemeanors: A Reevaluation of the Computer Fraud and Abuse*, 18 BERKELEY TECH. L.J. 909, 912 (2003).

20 Brent Wible, *A Site Where Hackers Are Welcome*, 112 YALE L.J. 1577, 1581 (2003).

21 National White Collar Crime Center and Federal Bureau of Investigations, Internet Fraud Complaint Center 2002 Internet Fraud Report (National White Collar Crime Center) (2003), 16.

pability against cybercrimes.[22] The plan noted that effective enforcement against cybercrime includes the following four requirements:[23]

I) A round-the-clock network of federal, state, and local law enforcement officials with expertise in, and responsibility for, investigating and prosecuting cybercrime;
II) Computer forensic capabilities, which are so essential in computer crime investigations;
III) Adequate legal tools to locate, identify, and prosecute cybercriminals, and procedural tools to allow state authorities to more easily gather evidence located outside their jurisdictions;
IV) Effective partnerships with other nations to encourage them to enact laws that adequately address cybercrime and to provide assistance in cybercrime investigations.[24]

Other requirements exist, but these four requirements touch on some of the most important issues for law enforcement today.[25] Law enforcement requires financial resources, trained personnel, advanced equipment, an understanding of technology, and a capability to identify and track down those who commit fraud. In addition, law enforcement needs legal authority and the ability to enforce those laws. If the government is deficient in any of these ways, its ability to enforce laws against fraud will be diminished. If the probability of capture were to approach zero, government would need to respond by increasing penalties infinitely high to maintain deterrence. Although in theory this would make the law just as effective, whether the government could actually do this has yet to be determined.[26]

This article looks into the extent to which governments have the capability to prohibit online fraud. The focus is fraud against merchants, but much of the analysis might apply to traditional consumer fraud or other types of computer crimes. The article goes through the four requirements outlined by Reno and documents whether government appears likely to be

---

22 Janet Reno, Statement of Janet Reno Attorney General of the United State Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime" February 16, 2000.

23 The plan contained ten points but this article focuses on four of the more important ones.

24 Janet Reno, Statement of Janet Reno Attorney General of the United State Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime" February 16, 2000.

25 Thomas Kubic of the FBI comes up with a near identical list: "The Internet presents new and significant investigatory challenges for law enforcement at all levels . . . . These challenges include: the need to track down sophisticated users who commit unlawful acts on the Internet while hiding their identities; the need for close coordination among law enforcement agencies; and the need for trained and well-equipped personnel to gather evidence, investigate, and prosecute these cases." Thomas T. Kubic, Statement for the Record, House Committee on the Judiciary, Subcommittee on Crime, June 12, 2001.

26 Wible, *supra* note 21, at 1622.

able to solve the problem of online fraud. Most of what I have learned in the research comes from interviews and conversations with technology workers in Silicon Valley. In this sense, the paper will shed little light on the situation to those who work in the industry. Instead, analysis of the industry may shed light on the extent to which classical-liberal theories apply to markets. Much of the evidence in this paper comes from interviews, which are admittedly anecdotal and have the potential to be biased. Whenever possible, I attempt to supplement information from interviews with quotes from government testimonies or other printed publications. The government testimonies may also be biased, but the direction will unlikely portray them as less capable than they truly are. The readers will be left to interpret whether they think the conditions under which government can prohibit online fraud are met.

Although different interpretations of the evidence may be possible, in my opinion the situation is quite clear. Many obstacles make enforcing laws against online fraud very difficult, if not impossible. Although government does enforce prohibitions in a select few transactions, in the vast majority of transactions, government does not appear to provide any redress, leaving merchants virtually helpless against online fraud. I find that the government is not able to solve the problem as the classical liberals would assume. Interestingly, the market does not break down as classical liberal theory would predict. It appears that classical liberals have a number of incorrect assumptions about markets. Perhaps the theories of Epstein and North are just theories with little applicability to the way the economy works.

2. DOES GOVERNMENT HAVE THE CAPABILITY OF PREVENTING ONLINE FRAUD?

Requirement I: A round-the-clock network of federal, state, and local law enforcement officials with expertise in, and responsibility for, investigating and prosecuting cybercrime.

If government is to enforce laws against fraud, it needs resources, computers, and enough personnel who are up to date in the latest technology. This condition seems as if it should be straightforward, but real-world practicalities get in the way. Despite some economic models that assume law enforcement to be costless,[27] law enforcement agencies have limited budgets and must decide where to allocate their scarce resources. The more government devotes to an endeavor such as online fraud, the less it can de-

---

[27] Karen Clay, *Trade, Institutions, and Credit*, 34 EXPLORATIONS IN ECONOMIC HISTORY, 503 (1997).

vote to other areas of law. Numerous cases of online fraud exist, and to expect government to deal with a significant portion of them may be unrealistic.[28] Bruce Townsend of the U.S. Secret Service stated, "Law enforcement does not have the financial or technological resources to cope with all these cases."[29] Although the U.S. government has been devoting more resources to online fraud in recent years, for much of the history of the Internet, a night watchman was not present.[30] Hiring around-the-clock law enforcement agents devoted to computer crime may be costly, but is at least possible.

Expecting law enforcement to have enough expertise in the latest technologies, on the other hand, is more problematic. Markets and technology are evolving at such a rapid rate that keeping up with all of the latest technologies is extremely difficult. Many agencies do have a number of extremely knowledgeable agents. That does not mean, however, that the agencies can keep up with all occurrences of fraud. With millions of potential incidents of fraud, any individual agent can only do so much. Government would need to hire numerous agents who are up to date with technology, and this may not be possible. One of the main obstacles is labor costs, because government must compete with the private sector for talent. If talented security experts can make more money in the private sector, the government may have a difficult time retaining enough workers who are knowledgeable about the technology.[31] If agencies do not have enough people with a sufficient understanding of the technology, they will be unable to enforce the laws against fraud.

Evidence of this problem was explained by one corporate executive whose company was a victim of a considerable online fraud. Not only were

---

[28] Consider the possible objectives for law enforcement. A public interest view would model them preventing crime and a public choice view would model them as taking actions to maximize budget or advance other government interests. Whatever we assume about their goals, they still might not devote resources to preventing online fraud. Agencies understandably might devote resources to where they get the most bang for the buck. If solving computer fraud does not bring the headlines or advance government interests as much as another endeavor, they might not devote as much resources as would be needed.

[29] Quoted in Jon Swartz, *Is the Future of E-mail under Cyberattack?*, USA TODAY, June 14, 2004.

[30] Rustad writes, "Most states have computer crime statutes, but do not have significant law enforcement presence in cyberspace." Michael Rustad, *Private Enforcement of Cybercrime on the Electronic Frontier*, 11 S. CAL. INTERDISC. L.J. 63, (2001) 98-99. See also CLIFFORD STOLL, THE CUCKOO'S EGG: TRACKING A SPY THROUGH THE MAZE OF COMPUTER ESPIONAGE (1989). Stoll discovered someone stealing time on his computer system and spent months tracking the hacker. After making numerous phone calls to various authorities, he was basically told that they were uninterested because his organization had not sustained losses over $1 million. In the end government became involved because Stoll gave them evidence that the hacker was also breaking into military systems.

[31] Rustad writes "Local law enforcement lacks the resources to recruit, train, and retain law enforcement officers with good computer skills. Low salaries and a high turnover of experts in cybercrime curtail the effectiveness of law enforcement at both the state and federal level." Rustad, *supra* note 31, at 99.

the legal authorities unknowledgeable of cutting-edge technologies; they were unknowledgeable about even the simplest technology. The company's own investigation had determined that a man named Mr. Yagolnitser was defrauding the company of money. After the company did the difficult work of identifying the culprit and reporting him to the authorities, was law enforcement any help? The executive said:

> The positive place where [government] failed was in providing security. The natural thinking was that when people are defrauding you, you can go to the police. Maybe Mr. Yagolnitser is not going to go to the police, but maybe we can go to the police and report Mr. Yagolnitser. We proceeded to do that. The FBI showed up at his home and concluded he was totally innocent. We'd given them Web pages. They were asking us, 'What's a banner ad?' [32]

For government to investigate whether someone is guilty of fraud, it needs to be up on current technology. The unawareness of basic aspects of the technology seems to indicate that it was years behind. In an interview, another employee from a Silicon Valley security firm told me, "In my view, government is ten years behind what's going on."[33]

One possible solution would be to devote more resources to government law enforcement,[34] but how much this would solve the problem is uncertain. One has to consider how much government would need to know to enforce all the laws. Whereas private companies spend significant resources mastering technologies that they know they will use, government would have to spend significant resources mastering all technologies that people may or may not use. To be able investigate any particular case, government would need a working knowledge of the systems employed by each company. Does government have this capability? Michael Vatis, Director of the FBI's National Infrastructure Protection Center, was quite frank that the answer is no. Vatis said, "It would be impossible for us to retain experts in every possible operating system or network configuration."[35] Given the limited resources of government and the numerous technologies in existence, law enforcement agencies are understandably unable to keep track of all of them. Under these circumstances, wrongdoers have the ability to move their efforts to technologies with which governments are

---

32 Presentation, Independent Institute, San Francisco, CA. April 21, 2004.

33 Personal interview, San Jose, CA. June 30, 2004.

34 Most government agencies believe that the solution lies with more money. For example, Janet Reno stated, "Resource issues are also critical. We must ensure that law enforcement has an adequate number of prosecutors and agents . . . trained in the necessary skills and properly equipped to effectively fight cybercrime." Statement of Janet Reno Attorney General of the United State Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime" February 16, 2000.

35 Michael Vatis, Statement of Michael A. Vatis on Cybercrime Before the Senate Judiciary Committee, Criminal Justice Oversight Subcommittee and House Judiciary Committee, Crime Subcommittee, February 29, 2000.

less familiar.[36] Without a knowledge of the various systems, government agencies may be unable to investigate.

One can dream up a world where government knows all technologies inside out and where government knows as much about the future course of technology as private companies. This may be possible, but there is little evidence that this is likely. In countries that rely on such a model, the track record of government guiding technology has not been positive. Government agencies appear to be at least one step behind everyone else.[37] Without enough people with an understanding of the latest technology, government will be unable to enforce laws against fraud.[38] This brings into question whether government is capable of enforcing laws that classical liberals say government needs to enforce.

Requirement II: Computer forensic capabilities, which are so essential in computer crime investigations.

Despite the poor track record of law enforcement agencies in recent years, one can imagine a world in which they are able to keep up with technological change. Even if this were the case, government still may be unable to enforce laws against online fraud. The next requirement for effective law enforcement is the ability to locate and identify those who commit fraud. But difficulties collecting evidence make enforcement of laws against online fraud quite difficult. The first reason investigating fraud can

---

[36] The problem of shifting activities to avoid prohibitions also surfaces with the law as well. Rustad describes what he calls a Cyberlaw Enforcement Lag: "By the time a statute is enacted to counter an Internet-related threat, the creative cybercriminal finds new technologies to bypass an essential element of the prohibited act or offense." Rustad, *supra* note 31, at 96.

[37] Brent Wible, *A Site Where Hackers Are Welcome*, 112 YALE L.J. 1577, 1581 (2003) ("Enforcement remains difficult, especially given the near impossibility of prosecuting attempts under 18 U.S.C. 1030(b), and the need for a great investment of time, resources, and skill—even assuming that local law enforcement agents have the requisite training.").

[38] Karen Alboukrek, *Adapting to a New World of E-Commerce: The Need for Uniform Consumer Protection in the International Electronic Marketplace*, 35 GEO. WASH. INT'L L. REV. 425, 440 (2003) ("Until law enforcement catches up with computer technology, [market participants] will be virtually unprotected from crime in the electronic marketplace.").

[39] Rustad, *supra* note 31, at 98 (2001) ("Internet crimes are seldom detected or prosecuted largely because there is no traditional crime scene.").

be difficult is the high degree of anonymity in non-face-to-face transactions. Although some types of fraud involve shipping goods to an actual address, other types of fraud involve no physical goods, so the fraudster need not ever reveal his real address. Where traditional law enforcement entailed sending investigators to the scene of the crime, online fraud has far fewer clues.[39]

With no witnesses to interview and no footprints to follow, law enforcement may simply be unable to figure out who is committing the fraud. A Report of the President's Working Group on Unlawful Conduct on the Internet (hereinafter President's Working Group) explains, "Another thorny issue stems from the lack of identification mechanisms on global networks .... Simply stated, given the current state of technology, it can be difficult to accurately identify an individual."[40] Even if they know that a law has been broken they may not know who the lawbreaker is. The government may be unable to identify the perpetrator or may not even know where to begin looking. A digital trail, if one even exists, can span around the globe.[41] As the President's Working Group explains:

> The communication may also pass through carriers in a number of different countries, each in different time zones and subject to different legal systems. Indeed, each of these complications may exist within a single transmission. This phenomenon makes it more difficult (and sometimes impossible) to track criminals who are technologically savvy enough to hide their location and identity.[42]

With each communication, the fraudster can use a different path, so figuring out the location and identity of the fraudster is often impossible.

Matters become even more problematic when fraudsters take active steps to hide their identity.[43] People can forge identities, forge IP addresses, use stolen accounts, and employ anonymity tools that make identification less likely.[44] Janet Reno admits, "Criminals can use a variety of methods to hide their tracks, allowing them to operate anonymously or through masked

---

[40] President's Working Group on Unlawful Conduct on the Internet, The Electronic Frontier: The Challenge of Unlawful Conduct Involving the Use of the Internet, March 2000.

[41] "The communications of a hacker or other criminal may pass through as many as a dozen (or more) different types of carriers, each with different technologies (e.g., local telephone companies, long-distance carriers, Internet service providers ("ISPs"), and wireless and satellite networks)." *Id.*

[42] President's Working Group, *supra* note 40.

[43] Wible, *supra* note 37, at 1581.

[44] "Sophisticated criminals can alter data concerning the source and destination of their communications, or they may use the Internet account of another." *Frontier*, *supra* note 40.

[45] Janet Reno, Statement of Janet Reno Attorney General of the United State Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime" February 16, 2000.

identities. This makes it difficult—and sometimes impossible—to hold the perpetrator criminally accountable."[45] The President's Working Group writes, "Encryption now presents and will continue to present a challenge to law enforcement confronting Internet-related crime. Robust encryption products make it difficult or impossible for law enforcement to collect usable evidence using traditional methods."[46] All of this "can plainly frustrate legitimate law enforcement efforts."[47] Matters become even more difficult if fraudsters are also hackers and have the ability to modify data that could be used as evidence.[48] Even if the data existed at one point in time, if the information can be deleted or altered, it can confuse an investigation.[49]

Computer forensic capabilities are also complicated by the fact that computer data are often not stored. Internet providers and networks have numerous users, and unless they track and report all user activities to law enforcement agencies, the activities of a fraudster may not be traced. Reno stated:

> Even if criminals do not hide identities online, we still might be unable to find them. The design of the Internet and practices relating to retention of information means that it is often difficult to obtain traffic data critical to an investigation. Without information showing which computer was logged onto a network at a particular point in time, the opportunity to determine who was responsible may be lost.[50]

Some communications may be recorded but not saved for any length of time, while other communications may go unrecorded.[51] If government lacks the necessary evidence to investigate a fraud, the fraud will go unsolved.

These technical difficulties pose obstacles for identifying perpetrators of online fraud. Although accessing, recovering, and decrypting data necessary for an investigation may be technically feasible, expecting that government will have the resources to do it in more than just a few cases may be unrealistic. In a few high-profile cases, the government has indeed caught perpetrators of online fraud, but the vast majority of cases go unre-

---

46 President's Working Group, *supra* note 40.

47 *Id.*

48 Albert writes, "Because of the ephemeral nature of information on the Internet, online fraud cases differ from traditional fraud cases, as data can be purged or reworked in such a way as to hinder investigation into suspected Internet fraud." Albert, *supra* note 1, at 592.

49 President's Working Group on Unlawful Conduct on the Internet, The Electronic Frontier: The Challenge of Unlawful Conduct Involving the Use of the Internet, March 2000, p.23.

50 Janet Reno, Statement of Janet Reno Attorney General of the United State Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime" February 16, 2000.

51 President's Working Group, *supra* note 40, at 30, 32.

ported, uninvestigated, or unsolved.[52] Without being able to identify the perpetrators of online fraud, the de facto situation is that government is unable to enforce the laws. Douglas North argues that anarchic markets can function when trading is face to face, but argues that markets cannot function when trading is relatively anonymous.[53] Perhaps one should apply his logic to law enforcement. As markets become more anonymous, how will government have the capability of enforcing the law?[54]

Requirement III: Adequate legal tools to locate, identify, and prosecute cybercriminals, and procedural tools to allow state authorities to more easily gather evidence located outside their jurisdictions.

Even if government could keep up with technology and locate and identify fraudsters, government still may lack the legal authority to enforce laws against fraud. Because online fraud can be committed from anywhere on the globe, a number of jurisdictional issues arise. The lack of geographical boundaries on the Internet gives companies access to many potential customers,[55] but it also exposes them to many potential fraudsters.[56] A

---

[52] Alex Kim, et al., *Fraud Over the Internet: The Same Old Story, Different Medium*, LEGAL COLUMN ARCHIVES (Ford Marrin, Esposito, Witmeyer & Gleser, LLP, New York, NY), Jan. 1999, http://www.fmew.com/archive/fraud; *see also* Wible, *supra* note 20, at 1577.

[53] NORTH, *supra* note 10, 34-35.

[54] Santa Clara Law Professor David Friedman predicts that government will become less able to enforce the law over time in his unpublished book manuscript Future Imperfect. David Friedman, Future Imperfect (Feb. 10, 2003) (unpublished manuscript), http://patrifriedman.com/prose-others/fi/commented/Future_Imperfect.html.

[55] Karen Alboukrek, *Adapting to a New World of E-Commerce: The Need for Uniform Consumer Protection in the International Electronic Marketplace*, 35 Geo. Wash. Int'l L. Rev. 425, 429 (2003).

[56] Reno states, "The Internet is a global medium that does not recognize physical and jurisdictional boundaries. A hacker—armed with no more than a computer and modem—can access computers anywhere around the globe. They need no passports and pass no checkpoints as they commit their crimes." *Cybercrime: Hearing Before the Subcomm. on Commerce, Justice, and State of the S. Committee on Appropriations*, 106th Cong. (2000) (Statement of Janet Reno, Attorney General of the United States).

[57] The President's Working Group writes, "In short, cybercriminals are no longer hampered by the existence of national or international boundaries, because information and property can be easily transmitted through communications and data networks. As a result, a criminal no longer needs to be at the actual scene of the crime (or within 1,000 miles, for that matter) to prey on his or her victims." President's Working Group, *supra* note 40.

fraudster might reside in one country, use computers in a second country, and commit fraud against a company in a third country.[57] What laws apply? And what law enforcement agency has jurisdiction in such a case? The fact that fraud takes place across geographical boundaries poses a number of problems.

The first problem stems from the fact that laws and legal procedures between countries differ. For example, if one government outlaws an action but another does not, the first government may be unable to apply the laws to the citizens of the second country.[58] Similar problems arise if one government treats fraud as a criminal matter and another treats it as a civil matter. The United States government has signed a number of extradition treaties with other countries, but unless both countries criminalize the act, the U.S. may be unable to pursue a case originating in the other country.[59] The President's Working Group recognizes, "When one country's laws criminalize high-tech and computer-related crime and other country's laws do not, cooperation to solve a crime, as well as the possibility of extraditing the criminal to stand trial, may not be possible."[60] Laws often differ greatly between countries and even differ within the same country through time; for example, the Computer Fraud and Abuse Act was adopted in 1984 and was amended in 1986, 1994, and 1996.[61] Even if a country adopted the exact same laws as those in the United States, unless they continue updating them over time, the two sets of laws might become incompatible.

When a case involves residents from other nations, a number of problems surface. Can law enforcement in the first country issue subpoenas, interview witnesses, and seize equipment for residents in the second nation if the action is not prohibited in that nation?[62] Each country has different ways of dealing with suspects, so how countries should deal with suspects

---

[58] David R. Johnson & David Post, *Law and Borders—The Rise of Law in Cyberspace*, 48 STAN. L. REV. 1367 (1996).

[59] The President's Working Group wrote, "The issue of dual criminality is not an academic or theoretical matter. In 1992, for example, hackers from Switzerland attacked the San Diego Supercomputer Center. The U.S. sought help from the Swiss, but the investigation was stymied due to lack of dual criminality (i.e., the two nations did not have similar laws banning the conduct), which in turn impeded official cooperation. Before long, the hacking stopped, the trail went cold, and the case had to be closed." President's Working Group, *supra* note 40.

[60] *Id.*

[61] *Id.*

[62] *Id.*

[63] Rustad, *supra* note 31, at 94.

in other countries is unsettled.[63] The President's Working Group highlights the difficulties associated with international investigations.[64] Consider what happens when a U.S. law enforcement agency has a search warrant from U.S. courts. Law enforcement may be authorized to search computers within the U.S., but does the warrant enable it to search computers in other countries? Even if a search has been authorized by the US government, another country may not consider the search legitimate. Problems arise with computer investigations because the location of computers is often unknown. Do governments have the authority to search computers around the globe just because a government says they can? The President's Working Group states: "ignorance of physical location may not excuse a trans-border search; consider how we would react to a foreign country's 'search' of our defense-related computer systems based upon a warrant from that country's courts."[65] A U.S. search warrant will be of little use when a different country does not wish to cooperate.[66] If law enforcement agencies need to get warrants from all other courts to begin an investigation, enforcing laws against fraud is that much more difficult.

One can dream up a world where all the laws and legal procedures were the same, but such circumstances are quite different than those in the world today. The President's Working Group explains the problem succinctly:

> The solution to the problems stemming from inadequate laws is simple to state, but not as easy to implement: countries need to reach a consensus as to which computer and technology-related activities should be criminalized, and then commit to taking appropriate domestic actions. Unfortunately, a true international 'consensus' concerning the activities that universally should be criminalized is likely to take time to develop. Even after a consensus is reached, individual countries that lack appropriate legislation will each have to pass new laws, an often time-consuming and iterative process.[67]

Although it may be possible for all countries to coordinate their laws and legal procedures, the likelihood of this happening in the near future is low.

The second problem arising from international fraud is that the question of what agency has jurisdiction is ill-defined.[68] Even if the laws and legal procedures are the same, what government will investigate and deal

---

64 President's Working Group, *supra* note 40.

65 *Id.*

66 Rothchild states, "Unduly aggressive enforcement action by government agencies in the context of cross-border online fraud risks giving rise to this sort of conflict, with detrimental effects on the efficacy of cross-border enforcement actions." Adding, that "the result can be conflict between two sovereigns." Rothchild, *supra* note 15, at 923.

67 President's Working Group, *supra* note 40.

68 *Internet and Federal Courts: Issues and Obstacles: Oversight Hearing Before the Subcommittee on Courts and Intellectual Property of the House Comm. on the Judiciary*, 106th Cong. (2000) (Statement of D. Jean Veta, Deputy Associate Attorney General, Department of Justice), *available at* http://commdocs.house.gov/committees/judiciary/hju66042.000/hju66042_0.htm.

with the fraud is an open question.[69] A merchant might be located in one country, a fraudster might be located in another country, and their computers might be located in yet another country. When fraud occurs, which government has jurisdiction? One might assume that a US company can simply turn to his local authorities, who coordinate with state and federal agencies, who in turn coordinate with authorities in the other nation. Despite the apparent simplicity, the situation is much more complicated.

Some examples can illustrate this problem. I listened to one former Silicon Valley executive describe his situation when his company was the victim of fraud originating in another country. When he attempted to follow standard procedures and contact officials, he soon realized that government would be of little help. He said, "There was a jurisdictional dispute between the FBI office in San Jose and San Francisco over which of them had jurisdiction over Kazakhstan, and which could handle it. So there were some very serious sorts of problems."[70] In the end the government did nothing to rectify his situation, and his company sustained tremendous losses due to fraud. Although the law against fraud is on the books, whether the government can do anything about it is uncertain.[71]

The classical-liberal conception of law enforcement is that all parties need to be subject to a monopolist arbiter of law.[72] Yet the ability for anyone with an Internet connection to transact with numerous parties around the globe brings into question where there can be a monopolist enforcer of law. Unlike spatially-based interaction, electronic commerce enables parties to interact without knowledge of their counterpart's location.[73] As more people interact with those outside their jurisdiction, it creates problems for government's geographically-based system of law. Incidentally, most of the classical-liberal arguments against private law enforcement apply to the situation at hand. How can parties interact when they are not both subject

---

69 Karen Alboukrek, *Adapting to a New World of E-Commerce: The Need for Uniform Consumer Protection in the International Electronic Marketplace*, 35 GEO. WASH. INT'L L. REV. 425, 434 (2003). Edward Stringham, *Market Chosen Law*, 14:1 J. LIBERTARIAN STUD. 53 (1999).

70 Presentation, Independent Institute, San Francisco, CA (April 21, 2004).

71 Wible, *supra* note 20, at 1581 (concluding that "With jurisdictional uncertainties looming in cases that are expensive to investigate and that require sophisticated tracking capabilities, state prosecution is almost impossible").

72 Gordon Tullock (ed.) Explorations in the Theory of Anarchy, (Center for the Study of Public Choice) (1972); Tyler Cowen, *Law as a Public Good: The Economics of Anarchy*, 8 ECONOMICS AND PHILOSOPHY 249 (1992); ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA (1974).

73 The President's Working Group writes, "In the physical world, one cannot visit a place without some sense of its geographic location. Whether a particular street address or an area of the world, human travel is spatially based. By contrast, because one can access a computer remotely without knowing where, in physical space, that computer is located, many people have come to think of the collection of worldwide computer linkages as 'cyberspace.'" President's Working Group, *supra* note 40.

to the same enforcer of law? One potential solution would be world government, but the desirability of that is questionable. Whereas North argues that we need government enforcement as trade moves outside small groups, he does not have a theory about how government enforcement can function as the groups become so big as to encompass people from many different nations.

One potential solution advocated by some lawyers is to give governments the authority to enforce laws on people outside their jurisdiction.[74] That would ensure that a merchant and a fraudster could be subject to a government regardless of the parties' locations. Johnson and Post point out a number of problems with this position.[75] Do we really want to give all governments on earth the authority to enforce laws on any citizen? Should American citizens be subject to Singaporean law enforcement if the Singaporean police are conducting an investigation or a prosecution?[76] If any government could subject residents of any other country to its procedures, the few legal protections against search and seizure might vanish, and the result could be a race to the bottom of legal rights.[77] Whether the citizens around the world would want to be subjected to all other countries' laws is unclear. That means that a government model of international law enforcement would require some type of coordination between countries, which is the final requirement.

Requirement IV: Effective partnerships with other nations to encourage them to enact laws that adequately address cybercrime and to provide assistance in cybercrime investigations.

Following Reno's sentiment, Deputy Assistant Attorney General Bruce Swartz states that international enforcement of law requires the "establishment of strong mechanisms for international cooperation, since computer-related crimes are often committed via transmissions routed through numerous countries."[78] For example, if a U.S. agency identifies a fraudster residing in a different country, the U.S. agency has to work with the authorities in the second country if it wishes to enforce the law. Even assuming that the laws are the same and the jurisdictional issues are sorted out,

---

74 Michael Geist, *Cyberlaw 2.0*, 44 B.C L. REV. 323 (2003), 345-7; Rothchild, *supra* note 15, at 986.

75 Johnson & Post, *supra* note 58.

76 To lawyers such as Geist, the answer is actually yes. Michael Geist, *Cyberlaw 2.0*, 44 B.C. L. Rev. 323, 345-47 (2003).

77 Laura W. Murphy, *ACLU Letter to the Senate Foreign Relations Committee on the Council of Europe Convention on Cybercrime*, (ACLU), June 16, 2004.

78 Multilateral Law Enforcement Treaties, June 17, 2004 (See statement of Bruce Swartz, Deputy Assistant Attorney General, Criminal Divison, Senate Foreign Relations Committee).

the extent to which different countries can coordinate their efforts is unclear.

One can imagine a world where all law enforcement agencies work in concert at little cost, but clearly the world is quite different. Given that even intranational coordination between agencies is often difficult, international coordination will likely remain more difficult. Contacting other law enforcement agencies and getting them involved in a case is usually time consuming and costly. The President's Working Group explains the problem: "law enforcement agencies are burdened with cumbersome mechanisms for international cooperation, mechanisms that often derail or slow investigations."[79] If an investigation is time sensitive, delays between agencies can stifle a would-be investigation.[80] Unless the U.S. government can rely on governments around the globe to assist and enforce its laws, then people will be able to commit fraud in other countries and remain outside the law.[81] Yet prohibition of fraud hinges on the law being enforced regardless of where the fraudster resides. The President's Working Group recognizes this very real problem: "With scores of Internet-connected countries around the world, the coordination challenges facing law enforcement are tremendous."[82] The result is that even though international fraud might attract attention from multiple law enforcement agencies, it possibly might attract the attention of none.[83]

The only real way to solve the problem would be to have tremendous coordination between law enforcement agencies around the globe. The President's Working Group brings up the many difficult requirements.

> Because the gathering of information in other jurisdictions and internationally will be crucial to investigating and prosecuting cybercrimes, all levels of government will need to develop concrete and reliable mechanisms for cooperating with each other. The very nature of the Internet—its potential for anonymity and its vast scope—may cause one law enforcement agency to investigate, inadvertently, the activities of another agency that is conducting an undercover operation. Likewise, the law enforcement agency of one state may require the assistance of another for capturing and extraditing a criminal to its state for prosecution. In other words, crimes that were once planned and executed in a single jurisdiction are now

---

79 President's Working Group, *supra* note 40.

80 *Id.*

81 Alboukrek, *supra* note 69, at 440.

82 President's Working Group, *supra* note 40.

83 Rothchild writes, "A technique commonly employed by professional perpetrators of consumer fraud is to set up operations in one country, but to target only residents of other countries. They hope that by doing so they will slip under the radar of law enforcement authorities, as authorities in the country in which they are located will perceive little interest in expending resources to protect foreign consumers, and authorities in the country where the victims are located will face practical difficulties in taking action against a seller located outside the country. In some cases, the laws are inadequate to respond to this problem. n110" Rothchild, *supra* note 15, at 921.

> planned in one jurisdiction and executed in another, with victims throughout the United States and the world. [84]

As wrongs can be planned and committed across borders, government enforcement would require the law enforcement in all countries to coordinate. The government would either need bilateral agreements with every country or a multilateral agreement with all countries. The Council of Europe has spent the past fifteen years debating and drafting a Cybercrime Convention, which to date has yet to be ratified.[85] Perhaps not surprisingly, the Cybercrime Convention has little to do with protecting online merchants and more to do with regulating business and creating laws against hate speech. International politics does not operate in a classical-liberal vacuum, so the treaty contains numerous aspects which are opposed by groups ranging from the US Chamber of Commerce to the American Civil Liberties Union.[86] Although matters may change, the likelihood of a worldwide multilateral agreement (or numerous bilateral agreements) to help online merchants does not seem high.[87]

Critics of private self-governance argue that without uniform government standards, competition will lead to a race to the bottom, where the weakest level of self-regulation will prevail.[88] One can debate the validity of this argument against self-regulation,[89] but it seems to apply to the current problem with multiple governments. If one country has lax laws or inferior enforcement ability, fraudsters can set up operations in that country knowing that the likelihood of getting caught is less. The President's Working Group writes, "Inadequate regimes for international legal assis-

---

84 President's Working Group, *supra* note 40.

85 UNITED STATES DEPARTMENT OF JUSTICE, FREQUENTLY ASKED QUESTIONS AND ANSWERS ABOUT THE COUNCIL OF EUROPE CONVENTION ON CYBERCRIME (2001); LAURA W. MURPHY, AM. CIVIL LIBERTIES UNION, ACLU LETTER TO THE SENATE FOREIGN RELATIONS COMM. ON THE COUNCIL OF EUROPE CONVENTION ON CYBERCRIME (2004).

86 Press Release, Linda S. Rozett, Media Relations Dir., U.S. Chamber of Commerce, U.S. Chamber Opposes European Cyber Crime Treaty (Dec. 8, 2000); LAURA W. MURPHY, AM. CIVIL LIBERTIES UNION, ACLU LETTER TO THE SENATE FOREIGN RELATIONS COMM. ON THE COUNCIL OF EUROPE CONVENTION ON CYBERCRIME (2004).

87 Even the people who believe that all countries have the same goals still do not put a lot of faith in governments ability to coordinate. Breslin writes, "These governments, international organizations, and businesses also agree on general policy issues concerning electronic commerce. For example, these institutions want to foster consumer trust and security, protect privacy, and permit continued technological innovation. When it comes time to act, however, general policy agreement does not necessarily translate into a consistent global regulatory scenario or perhaps even the likelihood of one in the future." Adrienne J. Breslin, Electronic Commerce: Will It Ever Truly Realize Its Global Potential? 20 PENN ST. INT'L L. REV. 275, 299 (2001).

88 Joel Trachtman, Regulatory Competition and Regulatory Jurisdiction, 3 J. INT'L ECON. L. 331 (2000).

89 Roberta Romano, Empowering Investors: A Market Approach to Securities Regulation, 107 YALE L.J. 2359 (1998).

tance and extradition can therefore, in effect, shield criminals from law enforcement: criminals can go unpunished in one country, while they thwart the efforts of other countries to protect their citizens."[90] Although classical liberals such as Richard Epstein argue that government is created to eliminate externalities,[91] unless all the externalities in the globe can be internalized, externalities between nations will still exist.

One of the classical-liberal arguments against private enforcement is that prohibitions against wrongdoing create spillover benefits to all people in society. Even if private parties could solve the problem, private parties would bear all the costs and not gain all of the benefits, so a free-rider problem would be present. Thus, according to the classical liberal, the government steps in to eliminate the externalities. But the arguments of why private law enforcement cannot function are just as easily applied to the current situation. Any effort by one country to prevent online fraud would be costly and would provide benefits to all other nations. The costs are local and the benefits are not, so the same free-rider problem may rear its ugly head. If the U.S. government devotes resources preventing wrongs in other countries, American taxpayers foot the bill and see little results. Public-goods theory notwithstanding, there is little evidence that law enforcement agencies act to maximize the social-welfare function of the entire world. Law enforcement agencies have objectives and limited budgets just like anyone else, so to assume that they only act to serve the global public good might be unrealistic.

Even if we assume that all law enforcement agencies act to reduce fraud, they may have different incentives to do so. For example, each agency will likely want to devote resources to solving fraud against its own citizens, because the agency will appear better to voters than if it spent its resources helping residents abroad. Even if the culprits reside in the agency's country, victims in other countries give law enforcement agencies no political support, so governments have less incentive to help them. Reno brings up this important problem:

> While we are working with our counterparts in other countries to develop an international response, we must recognize that not all countries are as concerned about computer threats as we are. Indeed, some countries have weak laws, or no laws, against computer crimes, creating a major obstacle to solving and to prosecuting computer crimes. I am quite concerned that one or more nations will become "safe havens" for cybercriminals.[92]

---

90 President's Working Group, *supra* note 40.

91 Richard Epstein, *Skepticism and Freedom: The Intellectual Foundations of Our Constitutional Order*, 6 U. PA. J. CONST. L. 657 (2004).

92 Janet Reno, Statement of Janet Reno Attorney General of the United States Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime" (February 16, 2000) (transcript available at: http://www.cdt.org/security/dos/000216senate/reno.html).

Classical liberals must recognize that not all governments act according to public-goods theory. Even if governments had the ability to do so, it seems unlikely that all countries will take the same interest in going after cyber fraud. If we introduce the possibility that certain governments simply do not care about American merchants, the likelihood that foreign governments will devote resources to eliminating fraud becomes lower.[93] Given that numerous governments have little concern for business in general, it seems unrealistic to think that they will help prevent fraud against foreign businesses.

Problems are exacerbated by the fact that governments in other countries may be even less knowledgeable about computer technology than is U.S. law enforcement. Yet international coordination of law enforcement hinges upon law enforcement agencies in every nation being up to date in the latest technology. Assistant Attorney General Michael Chertoff said, "When we deal with a transborder cybercrime, we *need* foreign law enforcement counterparts who not only have the necessary technical expertise, but who are accessible and responsive, and who have the necessary legal authority to cooperate with us and assist us in our investigations and prosecutions" (emphasis added).[94] "Technical expertise," "accessible," and "responsive" are not words that usually come to mind when thinking of governments around the world. To expect law enforcement agencies in less developed countries to solve a problem that U.S. agencies are incapable of solving might be a bit questionable. Can anyone honestly expect the government of Zimbabwe to help enforce laws against online fraud?

3. CONCLUSION

One of the main justifications of government is the idea that markets require government prohibitions against fraud. Yet we must recognize that law enforcement is not a perfect agent that can enforce laws without cost. Even if problems exist, the government may not have the ability to solve them. Wishful thinking notwithstanding, in the current world few of the conditions that government needs to prohibit fraud are met. Government has been unable to keep up with technology, lacks the necessary resources, and has difficulties collecting evidence and locating perpetrators of fraud. Government also faces organizational and jurisdictional uncertainties be-

---

[93] Rustad writes "It is difficult to discover the identity of cybercriminals, who often operate in countries with corrupt governments that encourage Internet crime as a developing industry. Crimes on the Internet cross national borders, creating the need for international cooperation in law enforcement." Rustad, *supra* note 31, at 86, 98-99.

[94] *Fighting Cybercrime: Efforts by Federal Law Enforcement: Hearing Before the Subcomm. on Crime of the H. Comm. on the Judiciary*, 107th Cong. 106 (2001) (Statement of Michael Chertoff, Assistant Att'y Gen., Criminal Division, U.S. Department of Justice).

cause fraud can take place across national boundaries where laws and legal procedures differ. Effective prohibition against fraud would require coordination between all law enforcement agencies, a situation that appears unlikely. Under these conditions the ability for government to prohibit online fraud is extremely limited.

To date, governments do not appear close to solving the problem.[95] Describing all types of computer fraud, attorneys Kim, Pinter, and Witmeyer estimate that "no more than 10% of the crimes involving computers get reported to authorities; further, less than 2% result in convictions."[96] Private companies know they cannot rely on government to rectify the situation, so in many cases they avoid reporting incidents. Even if government had a 100-percent recovery rate, companies would be reluctant to involve law enforcement because the cost of the legal process may exceed the cost of the stolen goods.[97] As the probability of recovery approaches zero, it is no wonder why companies would not turn to the law. Attorney General John Ashcroft recognized this issue saying, "victims are often reluctant to refer their cases to law enforcement," and adding, "we hope to convince the high tech community that when they report incidents of cybercrime, they are not just doing the right thing for their community—they are also doing the right thing for their business."[98] To state the issue is to admit that government does little to help merchants victimized by fraud. If involving government was really in the interest of firms, they would not need persuasion from officials.

Although the evidence presented in this paper does not prove that law enforcement agencies are inherently incapable of prohibiting online fraud, it does show that they have been ineffective to date. The classical liberal might respond that all law enforcement needs is more resources and more laws.[99] The important fact remains that merchants have been unable to rely on prohibitions against fraud for virtually the entire history of electronic commerce. If past performance is any indicator of future success, we should not expect government to have the ability to solve the problem any-

---

95 Wible writes, "The first cases of computer crime were heralded as an unprecedented phenomenon that law was not equipped to handle. Scholars and policymakers have since proposed a number of deterrence strategies, from criminal sanctions to tort law and the architecture of the web itself, but none of these methods has proved successful." Wible, *supra* note 21, at 1581.

96 Alex Kim, Edward Pinter, and John Witmeyer, *Fraud Over the Internet: The Same Old Story, Different Medium*, E-Zine of Ford Marrin Esposito & Gleser, LLP, January 2000, http://www.fmew.com/archive/fraud/index.html.

97 Albert, *supra* note 1, at 588.

98 John Ashcroft, U.S. Attorney Gen., Attorney General Ashcroft's Speech Announcing Expansion of CHIP Program and Establishment of Nine New CHIP units (July 20, 2001) (U.S. Department of Justice's transcript *available at* http://www.usdoj.gov/criminal/cybercrime/chipagsp.htm).

99 President's Working Group, *supra* note 40.

time soon. As Janet Reno stated, "these challenges are daunting."[100] Another, perhaps more realistic, way of looking at the problem is to recognize that government is not close to being able to solve the problem.

The situation and the proposed government solution are not as simple as the classical liberals assume. After looking at the evidence, we come to the exact opposite conclusion as Douglas North. In contrast to North, who argued that government must provide external enforcement as markets move outside of small circles, we have found that government enforcement becomes less possible in these circumstances. Relatively anonymous markets such as electronic commerce may pose problems for trade, but they pose even more problems that perplex government. Just because a problem exists does not mean that government has the ability to provide the solution. Whether the market breaks down as classical-liberal theory would assume is left to future research. Preliminary observation, however, suggests that electronic commerce is alive and well despite the fact that merchants are unable to rely on the law. This seems to indicate that markets are more robust than classical liberals assume. Indeed, Klein, Benson, Rothbard, Friedman, Caplan, and Stringham argue precisely that.[101]

One of the great contributions of economists is to point out that public policy requires more than wishful thinking.[102] Coming up with a theory of how markets are imperfect and how government can solve the problem is not enough. But lawyers and economists such as Epstein and North are guilty of exactly this. Classical liberals have theorized how markets require government prohibitions against fraud and how government can solve the problem. Yet in reality, the situation is quite different. George Mason economist Alex Tabarrok warns against what he calls theoretical empiricism.[103] People come up with a theory and then assume that the world conforms to their theory. But just because one assumes that the government can solve the problem does not mean that it actually can. It seems that classical liberals are indeed guilty of the Nirvana fallacy.

---

[100] Janet Reno, Statement of Janet Reno Attorney General of the United State Before the United States Senate Committee on Appropriations, Subcommittee on Commerce, Justice, and State, "Cybercrime," February 16, 2000.

[101] BRUCE L. BENSON, THE ENTERPRISE OF LAW: JUSTICE WITHOUT THE STATE (1990); MURRAY N. ROTHBARD, FOR A NEW LIBERTY: THE LIBERTARIAN MANIFESTO (Fox and Wilkes 1989) (1973); DAVID D. FRIEDMAN, THE MACHINERY OF FREEDOM: GUIDE TO RADICAL CAPITALISM (2d. ed. Open Court 1989) (1971); Bryan Caplan and Edward Stringham, *Networks, Law, and the Paradox of Cooperation*, 16 REV. OF AUSTRIAN ECON. 309 (2003).

[102] Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & ECON. 1 (1969); LUDWIG VON MISES, ECONOMIC CALCULATION IN THE SOCIALIST COMMONWEALTH (S. Adler trans., Ludwig von Mises Institute, 1990) (1920); George J. Stigler, *Public Regulation of the Securities Markets*, 37 J. OF BUS. 117 (1964).

[103] Alex Taborrok, *Market Challenges and Government Failure: Lessons from the Voluntary City*, In the Voluntary City: Choice, Community, and Civil Society 405-28 (David T. Beito, et al. eds., University of Michigan Press 2002).

PRIVATE DISPUTE RESOLUTION IN THE CARD CONTEXT: STRUCTURE, REPUTATION, AND INCENTIVES

Andrew P. Morriss, Ph.D. & Jason Korosec, J.D.***

ABSTRACT

Explosive growth in credit, debit, and other card payment systems in recent years has produced a parallel growth in private dispute resolution systems based on the web of contracts entered into by merchants, merchant acquirers, consumers, card issuers, card associations, and transaction processors. These contracts have produced legal systems based on contract and the enforcement of which rests primarily on reputational constraints. To cost-effectively resolve disputes, these private legal systems have evolved innovative procedures using resources at the lowest-possible level, including incentive-payments for producing information and rigid deadlines for parties' actions. This paper describes and analyzes these legal systems and their procedures as a potential model for resolving other categories of disputes.

---

* Galen J. Roush Professor of Business Law and Regulation & Director, Center for Business Law and Regulation, Case Western Reserve University and Senior Associate, Property & Environment Research Center, Bozeman, Montana; A.B. 1981, Princeton; J.D., M. Pub. Aff. 1984, University of Texas at Austin; Ph.D. (Economics) 1994, Massachusetts Institute of Technology. We would like to thank the following: Giancarlo Ibarguen S., Rector, and the Facultad de Derecho of Universidad Francisco Marroquin, Guatemala City, Guatemala, where the hospitality shown us during our visit in the summer of 2004 greatly facilitated work on this paper; Peter Boettke for organizing the project of which this paper is a part and inviting us to participate; seminar participants at a Critical Infrastructure Project seminar at George Mason University, Jonathan Adler, Olufunmilayo Arewa and Roger Meiners for comments; and Dean Gerald Korngold for research support. The views expressed in this paper are our own and do not reflect the views of any of our employers.

** Vice President and Director, Citishare Corporation; Chief Executive Officer, EagleCheck, Ltd.; and Adjunct Associate Professor of Law, Case Western Reserve University School of Law. B.A. 1990, University of Rochester; J.D., M.B.A. 1997 Case Western Reserve University.

Credit and other payment cards are revolutionizing many aspects of our economy, a revolution that "is arguably more profound than the introduction of paper money."[1] Use of payment cards of various types (charge, credit, and debit) has exploded, with the share of purchases using payment cards growing from 6 percent in 1984 to 32 percent in 2002.[2] In the year 2000, VISA[3] alone handled more than $1.7 trillion in global transactions.[4]

---

[1] David Evans & Richard Schmalensee, Paying With Plastic: The Digital Revolution in Buying and Borrowing 25 (1999).

[2] Paying with Plastic.org, *Stats and Facts, Payment Instruments*, http://www.payingwithplastic.org/index.cfm?gesture=statsDetailPrinter&aid=1312 (last visited February 7, 2005).

[3] Note that "VISA" is a used as a short-hand reference to the entire VISA network of organizations (e.g. VISA, USA, VISA Europe, etc.) Where our points depend on reference to a specific VISA entity, we give the precise name.

[4] PAUL CHUTKOW, VISA: THE POWER OF AN IDEA 81 (2001).

In this paper we explore how card-based payment systems[5] (e.g. VISA, MasterCard, Discover, American Express) have evolved in response to market and regulatory pressures to include dispute resolution systems, which have largely automated resolution of disputes. These card-based payment systems' dispute resolution procedures use reputational and financial incentives to induce the parties to reveal the information necessary to resolve the disputes. They systematically push dispute resolution procedures to the lowest possible level, and to use procedures that do not require significant investments of time or human capital to resolve the most frequent types of disputes. They also make use of the information generated by disputes to impose constraints which alter behavior to prevent similar disputes in the future involving different parties. These dispute resolution systems solve many of the problems of the public legal system (e.g. high costs, lack of speed). We believe that the card-based payment systems' dispute resolution procedures accomplish these ends without sacrificing many of the important values protected by the public legal system, including due process and fairness, and therefore, can serve as a model for rethinking dispute resolution more generally.

In Part I we sketch the structure of disputes and briefly compare public legal systems to card-based payment systems' dispute resolution processes. In Part II, we describe card-based payment systems' technology in more detail as part of an examination of their use of reputation and incentives. In Part III, we discuss the dispute resolution procedures in detail. In Part IV we examine the role of the state in shaping those systems. Part V concludes with our assessment of the viability of modeling public legal system reforms on card-based payment systems.

I. RESOLVING DISPUTES

By proposing card-based systems' dispute resolution systems as models for reforming the courts, we are suggesting a radical rethinking of dispute resolution.[6] To evaluate this alternative, we examine several core rea-

---

[5] We use the somewhat awkward term "card-based payment system" because there are many types of such systems, including "general purpose and limited-purpose credit cards, automated teller machine (ATM) cards, debit cards, smart cards, and check guarantee cards." MICHAEL AURIEMMA ET AL., THE BANK CREDIT CARD BUSINESS 1, 2 (2d ed. 1996). All of these types of cards offer at least the potential for the dispute resolution systems described here.

[6] The limited legal literature that discusses card based payment systems' dispute resolution processes has, thus far, largely rejected them as a model. At times the reasons for this rejection are not clear. *See, e.g.*, William Krause, *Do You Want to Step Outside? An Overview of Online Alternative Dispute Resolution*, 19 JOHN MARSHALL J. COMPUTER & INFO. L. 457, 472 (2001) (contending that card based systems' dispute resolution systems have "limited applicability" to dispute resolution generally because a "few disgruntled, vocal consumers can convince many others that there is an unacceptable risk. Even small anecdotal evidence of online misfortune is potential poison in the water."). Prof. Krause does not

sons that disputes require resolution by outsiders, which we describe to allow us to benchmark the alternative against the features of the public legal system.[7]

When parties have a dispute, they have a choice between "litigating" the dispute (i.e. using a dispute resolution mechanism that does not require agreement on the ultimate resolution, although it may require agreement on the process) and "settling" the dispute (i.e. agreeing on the ultimate resolution without an outside decision-maker reaching the final decision).[8] In this section we briefly sketch the structure of disputes and compare the approaches of the public legal system to the card-based payment system dispute resolution processes.[9]

A. *Approaches to Disputes*

We assume that the decision to litigate disputes between two parties occurs because one (or more) of the following conditions exists: (1) disagreement between the parties as to the facts; (2) disagreement between the parties as to the appropriate rule governing the dispute; (3) tactical advantages that reward a party for delaying resolution of the dispute; (4) the transactions costs of settlement exceed those of litigation; (5) one or more of the parties may be using the dispute to send a signal about its future behavior (or for other strategic reasons)[10] and the signal's value may exceed the savings from resolving the dispute without litigation; and (6) attempts to "roll the dice" and win because of the presence of random elements[11] in the

---

elaborate but this argument strikes us as implausible, particularly since the online services he suggests will succeed have not displaced the card-based systems' procedures. *Id.*

[7] We use the term "public" to refer to government courts, law, and so forth.

[8] Thus our definition includes mediation within settlement and arbitration within litigation. We recognize that this definition imposes a binary settle/litigate framework on a process that is a continuum between pre-filing settlement through settlement before the final decision is rendered on appeal to a final judgment upheld on appeal, but the key distinction appears to us to be whether an outsider renders the final decision or whether a resolution is voluntarily agreed upon by the parties.

[9] We provide a more detailed description of the card-based systems' process in Section II below.

[10] In IP cases, for example, the owner of an IP right may have strategic reasons for litigating unrelated to the costs and benefits from the particular suit. *See, e.g.*, Olufunmilayo Arewa, *Blocking, Tackling and Holding: Boundaries, Marking and Strategic Business Uses of Intangibles* 3 (Case W. Reserve Sch. of L., Working Paper, Case Legal Studies Research Paper No. 04-13, 2004), *available at* http://ssrn.com/abstract=586483 (last visited May 31, 2005) (arguing that "broad grants of intellectual property rights combined with intangibles paradigm business practices permit and even encourage the holders of such rights to use them as strategic weapons in a manner that may actually be a disincentive to future innovation.").

[11] *See* Gillian K. Hadfield, *The Price of Law: How the Market for Lawyers Distorts the Justice System*, 98 MICH. L. REV. 953, 970-71 (2000) (noting existence of "Professional platitudes such as, 'litigation is always a crap shoot' or 'give a problem to 10 different lawyers and you'll get 11 different answers.'").

dispute resolution process.[12] If none of these are present, no dispute or reason to litigate exists. We will briefly consider each of these possible causes of a decision to litigate.[13]

Where the parties disagree as to the facts, the primary function of the dispute resolution process is to prompt the exchange of information which brings the parties' divergent view of the facts closer together, making resolution possible.[14] Thus, for example, where the parties disagree about whether a debt is owed, the creditor's provision of documentation showing that the debtor incurred the debt would resolve the disputed factual issue and could cause the debtor to reevaluate the facts and recognize the debt. Resolving factual disputes is a "socially useful" function[15] of a dispute resolution system.

In the public legal system, such disputes are handled by a demand by the creditor for payment. If the debtor does not pay in response to the demand, a formal legal complaint must be filed.[16] The parties exchange legal

---

[12] We thus expand the traditional categories of differences in information and differences in optimism. *See, e.g.*, Robert H. Gertner, *Asymmetric Information, Uncertainty, and Selection Bias in Litigation*, 1993 U. CHI. L. SCH. ROUNDTABLE 75, 79 (1993) ("There are numerous possible explanations for why inefficient litigation may occur despite the cost savings from settlement. Most can be classified as either differences of opinion between litigants or differences of information between litigants.").

[13] Of course, more than one reason may be present in any particular instance.

[14] "Information is the lifeblood of [litigation]. Litigants battle to learn information, to conceal information, and to spin information so that it might better persuade judges, juries, and opponents to accept their view of the facts and law . . . it is probably no exaggeration to claim that litigation is all about the process of learning information, the cost of learning information, and the optimal response to information." Joseph A. Grundfest & Peter H. Huang, *The Unexpected Value of Litigation* 1 (Stanford L. Sch., John M. Olin Program in L. & Econ. Working Paper No. 292, 2004) *available at* http://ssrn.com/abstract=585803.

[15] We define "socially useful" to mean instances in which an institution increases the participants' total utility, not simply allocating a share of wealth to any particular party ("zero sum game"). Voluntary trades are thus the paradigmatic case, since they must increase the joint value in order to occur. This differs from the neoclassical definition of efficiency, since a Kaldor-Hicks efficient transaction would not meet our definition and yet would meet the neoclassical definition of efficiency. In the case of a dispute, of course, the loser's utility is reduced by the loss of the dispute if the parties' utilities are measured *ex post*. If they are measured *ex ante*, however, we contend that the expected utility of a dispute resolution institution is greatest for those institutions which provide the greatest possibility of factually-correct outcomes. This limitation is more rigorous than the test of neoclassical efficiency and is, we think, more consistent with a Hayekian approach to law. *See* FRIEDRICH A. HAYEK, RULES AND ORDER 96-97 (1973) (question to be decided by judge "will not be whether the parties have obeyed anybody's will, but whether their actions have conformed to expectations which the other parties had reasonably formed because they corresponded to the practices on which the everyday conduct of the members of the group was based."); *See also* Andrew P. Morriss, *Hayek and Cowboys: Customary Law in the American West*, 1 N.Y.U. J. L. & LIBERTY 35, 41-42 (2005) (describing Hayekian dispute resolution).

[16] Of course, summary collection procedures exist in the courts to reduce the transactions costs of handling more routine debt collections. Even these, however, involve more elaborate procedures than card-based dispute resolutions systems routinely use.

papers, conduct discovery by serving each other with written requests for production, requests for admission, and interrogatories, conducting formal depositions, and so forth. Disputes during this process may result in the need for intermediate decisions by a court. If the parties cannot agree to settle the dispute, a trial is conducted and a decision reached by either a judge or jury (depending on the circumstances). Appeals to higher courts may follow. At all stages, both parties are likely to be represented by attorneys; the court will employ expensive decision makers (legally trained judges, staff, and multiple lay decision makers on juries); the process takes months and, often, years. As an example, consider that under a fixed rate plan designed to *reduce* legal costs, the evaluation of a case and an initial complaint costs $6,000 if the amount in dispute is under $150,000.[17]

Within the payment systems context, this fact pattern arises when a cardholder disputes a charge and alleges that he did not use the card to incur the charge. In this situation, the cardholder complains first to the financial institution that issued the card. In the process that we outline in more detail below, the merchant attempts to retrieve a copy of the receipt and, if it is found, a copy is provided to the consumer. In many cases, this process ends the dispute, as the merchant either can or cannot prove the receipt exists. When it does not, the dispute is resolved after the financial institutions involved (on both the merchant's and the cardholder's side) have exchanged information through a highly structured process that specifies what information is to be exchanged and when the exchange is to occur. This process provides both positive financial incentives and negative substantive incentives for prompt compliance with the information exchange rules, and leads to a decision without the involvement of legally-trained personnel. Such disputes are required by the card system's rules (which are described in detail below) to be resolved within a relatively short, fixed period of time (typically within weeks rather than months).

When the parties disagree about the law, but not necessarily the facts, the primary function of the dispute resolution process is to clarify the legal rules which are applicable to the dispute. For example, where one party believes the case will turn on legal rules concerning how to interpret an employment agreement while the other believes it turns on the rules governing the fiduciary obligations to minority shareholders in a close corporation,[18] the dispute resolution process classifies the legal character of the dispute and resolves the uncertainty. As with resolving factual disputes, resolving uncertainty about the applicable law is a socially useful function of the dispute resolution system.

---

[17] Hadfield, *supra* note 11, at 958.

[18] *See, e.g.*, Jordan v. Duff and Phelps, Inc., 815 F.2d 429 (7th Cir. 1987) (dispute between majority opinion by Judge Easterbrook, holding case was a question of rights of a shareholder of close corporation, and dissent by Judge Posner, arguing case was a question of the rights of an at-will employee).

In the public legal system, formal legal proceedings must be initiated to resolve such disputes, as described above. The proceedings are comparatively unstructured, parties generally have legal counsel at all stages, and decision makers are expensive. In the payment system context, an example of a dispute over which rules govern is a consumer's warranty-related claims to the financial institution that issued the card used to purchase the underlying good or service. The consumer contends that the product is defective in a quality dimension that is covered by the warranty; the merchant denies that the relevant dimension is covered. As we describe below, both the consumer and merchant must follow predefined rules in resolving the dispute. Briefly, the consumer must first make a good faith attempt to resolve the dispute with the merchant before raising the issue with the consumer's financial institution. If the merchant and consumer fail to resolve the dispute, the consumer raises the disputed item with the issuer. That dispute will be resolved using the structured, technology-based dispute resolution process described below. The process is administered by non-lawyers, usually by telephone. The decision is reached through either the procedure or the substance of the applicable rule, typically within about a month or less. Or, if the dispute continues to through the maximum number of phases including a formal arbitration procedure (with a non-lawyer arbitrator), the maximum elapsed time from dispute to decision is about five months.

Where there is no fundamental disagreement over either the facts or the law, one party may still prefer to submit a dispute to resolution through a dispute resolution process to contracting for a settlement because the party believes it will benefit from the process itself. Thus, for example, where a party recognizes that it is liable but nonetheless believes it can earn a return on the amount it must ultimately pay in excess of any prejudgment interest obligations or that the opposing party may settle for a lower amount because of a pressing need for cash, there is an incentive to litigate to delay the inevitable payment.[19] Where one party is gaming the system there is no socially beneficial function of the litigation since the gain of one party is at

---

[19] For example, Prof. Elihu Inselbuch argues that

> [t]he real source of delay in the tort system . . . arise[s] from the economics of the tort system and the insurance industry, which combine to create an impetus for defendants to withhold realistic settlement offers. Insurance companies earn their profits from the investment of premiums that they collect from their insured. The longer the insurers can delay payments to plaintiffs, the greater the return they will realize on the funds withheld. The insurers' incentive to exploit the time value of money is compounded by a tort system that imposes no costs on them or their insured clients for delay in the payment of claims. If an insurer can settle a case on the eve of trial for the same amount it would have cost to settle the claim years earlier when the plaintiff first sued, then no incentive exists to move the insurer to settle and pay the claim earlier. Indeed, given that the insurer is given a free float of the amounts owed the tort victim, the system gives insurers and self-insured defendants a huge incentive not to settle early because an early settlement would forfeit the time value of the money.

Elihu Inselbuch, *Contingent Fees and Tort Reform: A Reassessment and Reality Check*, 64 LAW & CONTEMP. PROBS. 175, at 183 (2001).

the expense of the other party rather than as a result of increasing the size of the joint surplus.

In the public legal system, gaming the system is difficult to police, as parties have a great deal of freedom to structure their legal pleadings and actions. Even after repeated, clear cases of abuse, for example, the public legal system is rarely able to restrict future abuses by plaintiffs.[20] The problem for the public legal system is two-fold. First, there are no mechanisms that identify gaming behavior because there is no institutional remedy which can identify such behavior based on a comparison of particular parties' behavior to other transactions. Second, even where such information is found on an ad hoc basis, the public legal system is often reluctant to sanction parties who game the system because doing so forecloses access to the courts.

In the payment system context, there are also opportunities for gaming behavior. For example, if a consumer complains about a charge, during the dispute period the amount in dispute is temporarily debited from the merchant's account and credited back to the consumer. This provides the consumer with additional credit, since charges are not applied to the account during the dispute. (Once the dispute is resolved, the temporary debits and credits are either reversed or made permanent.) Consumers who repeatedly game the system, however, self-identify themselves to their card-issuer. Since the issuer bears some of the costs from consumer complaints, these consumers' poor reputation for honesty can be a basis for the issuer to cancel the consumers' cards. The distinctive feature of card-based payment systems is their ability to make use of the parties' reputations in controlling attempts to game the system.[21]

In some situations, the parties may simply be unable to resolve their dispute because of particularly high transactions costs involved in settlement compared to the transactions costs of the dispute resolution system

---

[20] Prisoners who abuse *in forma pauperis* filings are the only area where the courts regularly impose such sanctions. *See, e.g.*, Slicher v. Thomas, 111 F.3d 777, 780-782 (10th Cir. 1997) (sanctioning a pro se prisoner litigant who had filed 33 matters with the 10th Circuit from 1989 to 1997, most of which were summarily dismissed, and concluding that court had "determined to call a halt to Mr. Schlicher's wasteful abuse of judicial resources" by barring him from future filings *in forma pauperis* except in cases alleging an imminent danger of personal injury, forbidding him to file pro se pleadings, and requiring production of specific information on his litigation to accompany any filings). Even in these cases, however, the sanctions are imposed only after repeated abuses, and are limited in their ability to forestall future abuses. Further, no systemic learning about how to deter others from similar abuses occurs.

[21] Reputation can serve two important functions. First, reputation may be a means of dealing with non-verifiable information about a customer. Second, reputations aggregate information about discrete events, with the aggregation providing more information than the sum of the information of the individual events. For example, knowledge is gained about a customer with multiple disputes with merchants in which the facts are unclear by virtue of there being multiple disputes beyond the ambiguous information contained in each dispute.

itself. For example, it may be too costly to educate a corporate defendant's decision making authority about the facts and law concerning a small dispute once the opportunity costs of the decision makers' time is considered, and so litigating may be less expensive than settlement. Here the invocation of the dispute resolution process is socially useful only because of the relative transactions costs of litigation and settlement. In the public legal system, there are high transactions costs of understanding the legal system's substantive and procedural rules, as is demonstrated by, and perhaps caused by, the extensive involvement of lawyers in all stages and types of public legal system disputes. In the payment system context, however, these transactions costs are reduced significantly by the use of a comprehensive set of rules categorizing disputes and standardizing procedures, allowing an almost lawyer-free dispute resolution process. Moreover, standardization itself reduces the cost of participation by expensive participants. Many financial institutions, for example, regularly have summary reports reviewed by high level employees. These reports identify outliers and anomalous transactions, which can thus attract high level input, while routine cases do not waste resources. Further, the shifting of costs to losing parties provides an additional motivation for participants to prefer settlement to litigation.[22]

Finally, to the extent a random element exists in a dispute resolution process, parties who know that an accurate process would find them liable may be willing to "roll the dice" through litigation, in effect seeing the additional costs of litigation over settlement as the price of a lottery ticket whose prize is elimination of the obligation. Thus, for example, where both parties to an oral contract know that the contract was properly made, one party may be unwilling to perform because of the positive probability that a jury will not believe the plaintiff's truthful testimony about the contract's formation, believing instead the defendant's false testimony. In this case the dispute resolution process does not serve a socially useful function and the decision to litigate is motivated solely by the existence of a substantial enough random element.

The public legal system's reliance (in the United States) on lay juries includes a relatively large random element, since it reduces the predictability of trial outcomes, perhaps best captured by the oft-repeated comment

---

[22] In the payment systems context, certain merchants and consumers may be less organized and responsive than others. Since one aspect of the payment system's dispute resolution process, as described below, is grounded on responsiveness, a non-responsive party is likely to lose the dispute permanently. Since neither party knows how responsive the other party will be in a dispute, there is always a chance that the other party will be a slouch and the likelihood of winning increases substantially. This possibility is not a problem to the dispute resolution system within the payment system because all disputes are tracked. Cardholders who appear to be gaming the system may have their accounts canceled or not renewed.

that "you never know what a jury will do."[23] Card-based payment systems' dispute resolution systems, by contrast, rely more heavily on structured analyses, which provide predictability. Moreover, since the financial institutions and networks involved in card-based payment systems are repeat players, they profit from reducing randomness and invest in data analysis to do so.

Dispute resolution processes thus serve socially useful functions when they resolve factual and legal disputes and privately useful, but socially costly functions when dispute resolution processes have characteristics that motivate parties to litigate when no factual or legal dispute exists. In evaluating alternatives to the public legal system, the relative performance of alternatives in addressing each of these scenarios is one means of evaluating their value. In particular, the ability of a legal system to focus its resources on socially useful categories of dispute resolution is important indicia of success.

Our preliminary account of card-based payment systems points toward several important differences between card-based payment systems' dispute resolution processes and those of the public legal system. First, card-based payment systems are able to make use of lower cost inputs because of the more structured nature of their processes. Second, card-based payment systems make extensive use of positive and negative incentives and reputation in securing compliance with procedures, reducing enforcement costs. Third, card-based payment systems collect information that allows them to eliminate future socially costly disputes by imposing constraints on participants.

B. *Dispute Resolution as a Technology*

Let us consider a somewhat stylized version of dispute resolution, to identify the civil justice system's features that can serve as a benchmark for the card-based systems' dispute resolution processes by examining the technology the public courts use to resolve disputes with respect to three key attributes: (1) the provision of factual information to the decision maker; (2) the means of deciding questions of both fact and law when the information provided is insufficient to compel a particular resolution clearly

---

[23] *See, e.g.*, ROBERT A. KAGAN, ADVERSARIAL LEGALISM: THE AMERICAN WAY OF LAW 127 (2001) (listing unpredictable "lawyer-driven, jury-centered methods of adjudication" as a distinguishing feature of U.S. legal system); Gary T. Sachs & Neal W. Settergren, *Juries Should Not Be Trusted to Decide Maritime Cases*, 34 J. MAR. L. & COM. 163, 170 (2003) ("A better instrument could scarcely be imagined for achieving uncertainty, capriciousness, lack of uniformity, disregard of former decisions—utter unpredictability.").

and conclusively;[24] and (3) the methods of correcting decisions or ensuring correct, unbiased initial outcomes.

The rules of civil procedure and evidence used in civil trials in the various common law jurisdictions provide one approach to accomplishing this objective. The quite different rules used in civil law jurisdictions provide another.[25] How a legal system (public or private) handles each of these functions depends on the incentives and constraints created by the system's rules and its technology. For example, Rule 55 of the Federal Rules of Civil Procedure encourages participation in the dispute resolution procedure of the federal courts by allowing parties whose opponents do not respond in a timely way to obtain default judgments, and Rule 11 creates disincentives for parties to make misrepresentations by penalizing the lawyers who sign pleadings containing the misrepresentations.

In general, the public legal systems rely primarily on notice pleading, a system of rules in which "the factual allegations contained in the pleadings are judged by lax standards, because it is understood that more detailed knowledge of the facts must await further development through the elaborate discovery process provided for in the rules."[26] These rules place few limits on the participants' ability to conduct expensive and time consuming searches for evidence that may fit the broad set of claims of which the pleadings give "notice."

The rules governing disputes in the public courts are largely built around encouraging voluntary compliance with the rules through a combination of appeals to professional norms and sanctions for violations. The rules rarely include positive incentives, such as cash payments, to encourage compliance. The problem with this approach is that professional obligations require attorneys to put compliance with rules ahead of their clients' and, sometimes, their own personal financial interests.[27] Furthermore,

---

[24] [I]f a controversy should arise between two men concerning the ownership of property, and there be no statute upon the subject, the unwritten law must, nevertheless, decide it. No matter how novel the question, it must be determined. It would not be endurable that one man should hold unchallenged possession of property to which another honestly laid claim, for the reason that the case was so novel as to render it difficult to determine to whom it justly belonged. Society may leave a criminal unpunished; private citizens do not feel an additional burden on this ground; but it cannot leave private controversies undecided, or to be decided by force.

JAMES C. CARTER, THE PROPOSED CODIFICATION OF OUR COMMON LAW 34-35 (1884).

[25] The universe of possible approaches is considerably broader than the set of approaches used by various official legal systems.

[26] Martin H. Redish, *Electronic Discovery and the Litigation Matrix*, 51 DUKE L. J. 561, 606 (2001).

[27] *See, e.g.*, Hadfield, *supra* note 11, at 955 ("the very concept of professionalism requires that a disregard of economic incentives be a moral duty for the professional."); *see also* Philip J. Havers, *Take the Money and Run: Inherent Ethical Problems of the Contingency Fee and Loser Pays Systems*, 14 NOTRE DAME J. L. ETHICS & PUB. POL'Y 621, 625 (2000) ("Because of this large personal financial stake [due to contingent fees], the attorney can no longer look upon his practice of law as one devoted primarily to justice. Besides calling into question this basis of our professional rules that he is now

negative sanctions are discounted for the probability that one will not be caught.[28] The public courts thus are handicapped by their inability to make systematic use of the reputations of either the lawyers or the parties and the lack of certainty that those violating the rules will be identified and punished.

With respect to the technology of dispute resolution, the public courts depend primarily on live testimony before the decision maker, with cross-examination by the opposing counsel and observation by the decision maker to test the veracity and accuracy of the testimony. The rules' approach may vary with the nature of the dispute or information. For example, in fraud cases, where the nature of the claim raises concerns about the factual basis of the claim, Federal Rule 9(b) requires more specific pleading than in a negligence case.[29] Similarly, the federal rules of evidence exclude much hearsay evidence on the grounds that it is inherently unreliable.[30] While an improvement over many earlier technologies (e.g. trial by ordeal),[31] this technology remains largely unchanged since the early twentieth century introduction of modern civil procedure,[32] and has significant imperfections.

---

more likely to ignore or, at the least, will play with at the margins, the negative aspects of the contingency system work their way into the sacred relationship between the attorney and client.").

[28] RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 220 (6th ed. 2003) (noting "growing empirical literature on crime" that shows people respond to changes in "opportunity costs, in the probability of apprehension, in the severity of punishment, and in other relevant variables . . . .").

[29] F.R.CIV. P. 9(b) ("In all averments of fraud or mistake, the circumstances constituting fraud or mistake shall be stated with particularity. Malice, intent, knowledge, and other condition of mind of a person may be averred generally."). Rule 9(b) has been criticized heavily. *See, e.g.*, Christopher M. Fairman, *An Invitation to the Rulemakers—Strike Rule 9(b)*, 38 U.C. DAVIS L. REV. 281, 282 (2004) ("At best, Rule 9(b) is an anachronism—harkening back to the abandoned pleading practices of the past that spawned the modern Federal Rules.")

[30] *See* Paul F. Kirgis, *Meaning, Intention, and the Hearsay Rule*, 43 WM. & MARY L. REV. 275, 301-06 (2001) for a discussion of the justifications for hearsay exclusions.

[31] *See, e.g.*, J.H. BAKER, AN INTRODUCTION TO ENGLISH LEGAL HISTORY 5 (3rd ed. 1990):

> The procedure [in early English law] in contentious matters was calculated to avoid reasoned decision-making [which some times included] the physical test of an ordeal. . . . Ordeals involved an appeal to God to reveal the truth in human disputes, and they required priestly participation to achieve this rapport with the Deity. . . . [I]n England, they usually took the form of fire or water. In the former, a piece of iron was put into a fire and then in the party's hand; the hand was bound, and inspected a few days later: if the burn had festered, God was taken to have decided against the party. The ordeal of cold water required the party to be trussed and lowered into a pond; if he sank, the water was deemed to have 'received him' with God's blessing, and so he was quickly fished out.

The parody of such procedures in the film *Monty Python and the Holy Grail*, in which Sir Bedevere examines a purported witch and concludes that if she weighs the same as a duck then she is a witch, is humorous precisely because of the seeming ridiculousness of early English trial methods. *See* http://www.rit.edu/~smo4215/monty.htm#Scene%205 (for a link to the script segment on the trial listed under Scene 5) (Last visited February 7, 2005).

[32] *See generally* Stephen N. Subrin, *How Equity Conquered Common Law: The Federal Rules of Civil Procedure in Historical Perspective*, 135 U. PA. L. REV. 909, 1001-1002 (1987) (describing evolution of modern civil procedure and tracing contemporary problems to structural design of rules).

To correct errors and biases, courts rely on appeals, which focus primarily on legal questions rather than factual issues.[33] To prevent errors and biases, the structure of compensation and working conditions for judges aim at eliminating conflicts of interest.[34]

In cases where the outcome depends upon resolution of factual uncertainty, the lack of technological progress is unsurprising since many such cases ultimately depend on choosing between two competing versions of the truth as set out by witnesses, physical evidence, and documents. Particularly in the case of single episode interactions between strangers (e.g. a transaction by a traveler away from home or an automobile accident), the involvement of the public legal system's dispute resolution process only after the dispute exists prevents any other approach, since the events leading to the dispute are over before the legal system is involved. Therefore, only a retrospective accounting of the facts is possible.

The lack of technological progress is also not surprising given the substantial monopoly power of the legal profession in disputes in the public legal system.[35] Monopolies generally tend to under-produce innovations.[36] Lawyers' role as one of the major costs of the public legal system suggests that cost-reducing innovations would likely reduce their profits. Even the few cost-reducing innovations which have been introduced, such as the use of paralegals to do work previously done by lawyers, are limited by the legal profession's control over the practice of law.[37]

The formal legal system's methods of resolving disputes thus include (at least) four problematic features. First, there are a substantial number of single event litigations, where one or more parties is not a repeat player. The lack of repeat interactions reduces the usefulness of parties' reputations in creating incentives for honesty and rule compliance. Of course, lawyers

---

33 POSNER, *supra* note 27, at 601.

34 *See* Richard A. Epstein, *The Independence of Judges: The Uses and Limitation of Public Choice Theory*, 1990 BYU L. REV. 827 (1990); Richard A. Posner, *What Do Judges Maximize (The Same Thing Everybody Else Does)*, 3 S. CT. ECON. REV. 1 (1993).

35 *See, e.g.*, Hadfield, *supra* note 11, at 999:

> The market for lawyers is fundamentally noncompetitive. As a consequence of the complexity of legal reasoning and procedure, the profession's derived monopoly on the legitimate use of coercion, and the unification of the profession to serve the diverse needs for access to law, the price of law that emerges from the free market for lawyers is too high.

36 3 PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATION ¶ 631 (1996) ("[W]e worry about monopoly because of its generally evil result or potentialities: reduced output and higher prices, diminished incentives for innovation, and fewer alternatives for suppliers and customers.").

37 *See* Carl M. Selinger, *The Retention of Limitations on the Out-of-Court Practice of Law by Independent Paralegals*, 9 GEO. J. LEGAL ETHICS 879 (1996) (discussing means of limiting the impact of independent paralegal practice on lawyers through ethical rules); Benjamin H. Barton, *An Institutional Analysis of Lawyer Regulation: Who Should Control Lawyer Regulation—Courts, Legislatures or the Market?*, 37 GA. L. REV. 1167, 1189 (2003) ("Raising entry barriers has been the sine qua non of the formation of modern bar associations and lawyer lobbying.").

are by definition repeat players,[38] but the size of the legal profession in many larger communities and the limited control measures available to courts means that reputation effects are often diluted even for attorneys.

Moreover, the public legal system usually makes only limited use of information about parties' prior conduct.[39] Legal rules limit the circumstances in which even previous formal disputes can be considered.[40] Although the card-based payment systems do not use prior complaints to determine the outcome of a particular dispute, the financial institution may cancel or not renew the account of a consumer who frequently initiates formal disputes because the costs of servicing that consumer are larger. Similarly, merchants that receive a significant number of formal disputes will be required by the financial institutions handling their accounts to pay a higher discount rate[41] to gain the ability to accept card based payment systems, or in egregious cases, abusive merchants may be expelled from the system altogether.[42]

Second, there is a significant random, or at least, a non-merits related element to dispute resolution, largely due to the role of juries[43], but also

---

[38] *See, e.g.*, W. Bradley Wendel, *Informal Methods of Enhancing the Accountability of Lawyers*, 54 S.C. L. REV. 967, 970 (2003) ("A repeat-player lawyer with a contrary reputation faces numerous costly obstacles, such as the refusal by other lawyers to agree to reasonable schedule changes, the need to memorialize every agreement in writing, and difficulty making credible commitments."). Wendel uses accounts of lawyers in Charleston, S.C. and the Chicago municipal courts to argue that reputational constraints are powerful constraints on the practice of law. He notes, however, problems with the form of these constraints, including the possibility that the relevant reputational constraint is the lawyer's allegiance to his legal community rather than to the client, encouraging a tradeoff of the client's interest for the lawyer's, and the possibility that the relevant legal community norms themselves may be problematic (e.g. to exclude minorities from lucrative areas of law practice).

[39] We recognize that we are speaking at a high level of generality—there are a variety of public legal systems (federal, state, small claims, bankruptcy, etc.) and each has its own rules.

[40] Collateral estoppel and res judicata principles, of course, allow some use.

[41] The discount rate is the fee that the merchant pays to the financial institution processing these transactions on their behalf. The fee is typically a percentage of the transaction. *See* MASTERCARD DICTIONARY (December 2003), at 34.

[42] *See* Henry H. Perritt, Jr., *Dispute Resolution in Cyberspace: Demand for New Forms of ADR*, 15 OHIO ST. J. ON DISP. RESOL. 675, 691-92 (2000). Higher fees and expulsion serve as important limits on fraudulent uses of the card systems. *See, e.g.*, Barry Cutler, Statement of the Federal Trade Commission before the Select Committee on Aging, Subcommittee on Health and Long-Term Care, Committee on Small Business, Subcommittee on Regulation, Business Opportunities, and Energy, U.S. House of Representatives (June 21, 1991), *in* PRACTISING LAW INSTITUTE, RECENT TRENDS IN TELEMARKETING FRAUD, Nov. 1991, (759 PLI/Corp 479) at 505 (describing how use of 900 numbers substituted for credit card billing in fraud schemes, and how "this option solves several problems for the fraudulent telemarketer. First, the company need not meet the criteria that major credit card companies impose for obtaining a merchant account with a bank. Second, the company is able to use a payment system that lacks the dispute resolution procedures and other safeguards for credit card transactions found in the Fair Credit Billing Act.").

[43] *See, e.g.*, W. Kip Viscusi, *Punitive Damages: How Jurors Fail to Promote Efficiency*, 39 HARV. J. ON LEGIS. 139 (2002) (describing experiments demonstrating jurors' failure to follow instruc-

attributable to quality issues in the judiciary.[44] Because we rarely observe parties outside the formal legal system investing in either randomly chosen lay panels or individuals chosen through the political process to decide disputes, we can also infer that such panels' and judges' primary advantages do not lie in accuracy or reduced transactions costs. This random element, of course, increases the number of times the process is invoked in a socially costly way.

Third, litigation in public legal systems is costly.[45] The costs are due to both the length of proceedings and the involvement of large numbers of highly trained and highly compensated individuals, including both judges and lawyers. The success of privately provided alternative dispute resolution systems which offer reduced costs[46] suggests that the public legal system's costs are higher than necessary to resolve at least some disputes.

Fourth, the public legal system is involved in many disputes only *ex post* (although parties may change their behavior *ex ante* in anticipation of litigation). The public legal system cannot, therefore, dictate parties' behavior in advance of disputes. Thus, for example, the negligence rule in tort law may produce efficient levels of care by potential tortfeasors in some situations,[47] but it does not influence decisions on activity levels leading to inefficiently high activity levels.[48] Intervention in transactions or other interactions before a dispute arises may be a less costly means of handling a matter than *ex post* dispute resolution.[49]

---

tions in assessing punitive damages and resulting penalties for firms that engage in risk analysis). On juries more generally, *see generally* Dan Simon, *A Third View of the Black Box: Cognitive Coherence in Legal Decision Making*, 71 U. CHI. L. REV. 511, 550-58 (2004) (discussing cognitive biases introduced into jury decision making by the structure of trials).

44 *See, e.g.*, Geoffrey P. Miller, *Bad Judges*, 83 TEX. L. REV. 431, 431 (2004), who notes that "In jurisdictions across the country, complaints are heard about judges and magistrates who are incompetent, self-indulgent, abusive, or corrupt. These bad judges terrorize courtrooms, impair the functioning of the legal system, and undermine public confidence in the law. They should not be allowed in office, yet many retain prestigious positions even after their shortcomings are brought to light. The situation, moreover, does not appear to be under control."

45 *See* Graham C. Lilly, *The Decline of the American Jury*, 72 U. COLO. L. REV. 53, 57-58 (2001) (describing costs of trials).

46 *See, e.g.*, Barak D. Richman, *Firms, Courts, and Reputation Mechanisms: Towards a Positive Theory of Private Ordering*, 104 COLUM. L. REV. 2328, 2341-42 (2004) (describing categories of efficiencies available in private dispute resolution).

47 *See, e.g.*, STEVEN SHAVELL, ECONOMIC ANALYSIS OF ACCIDENT LAW 16 (1987) (describing circumstances under which model predicts negligence rule will produce optimal outcome).

48 *Id.* at 23-24 (noting that injurers do not "have a reason to consider the effect that engaging in their activity has on accident losses. Consequently, injurers will be led to choose excessive activity levels.").

49 *See, e.g.*, Michigan Manufacturers Association, Insurance & Benefits, http://www.mma-net.org/insurance/workers_comp.asp (last visited Sept. 12, 2005) ("Amerisure combines the expertise of experienced professionals and sophisticated programs to find solutions to plant safety issues, prevent losses and, ultimately, reduce insurance costs.").

Card-based payment systems' dispute resolution systems offer a different approach to solving each of these problems. Card-based payment systems convert all interactions between participants into repeat player transactions.[50] Rather than a single episode merchant-customer interaction, the use of a card-based payment system to make a purchase creates a series of related, repeat player transactions: merchant-bank, bank-payment mechanism provider,[51] payment mechanism provider-bank, and bank-customer. Card-based payment systems harness reputations to prevent and resolve disputes because these related repeat transactions are counted and analyzed. The value of the relationship itself is defined by the analysis of the transaction counts, and conclusions drawn about the revenue, cost, and profit of the overall relationship. These systems also structure the underlying transactions, dictating features of the cards, recordkeeping, and transaction processing which decrease the frequency of disputes and the non-objectively verifiable aspects of those disputes which do occur, reducing the scope for randomness in decisions. Card-based payment systems' dispute resolution procedures also drive the processing of claims down to relatively low level employees and contain incentives for constant cost reductions, thereby lowering transaction costs. In the following section we examine how card-based payment systems accomplish these results.

II. PAYMENT SYSTEMS AS A TECHNOLOGY

In order to understand how incentives and features in the card-based payment dispute resolution systems lower costs, we must first understand how the payment system itself works. Some features will be difficult to translate to other dispute resolution systems; other aspects are more readily applicable to dispute resolution in general. Distinguishing among these features and incentives requires an examination of the technology itself.

---

[50] This point is sometimes missed in the literature on e-commerce, which overlaps to some extent with the card-based payment systems literature. For example, Prof. Llewellyn Joseph Gibbons calls for government intervention into e-commerce consumer contracts to "encourage the development of consumer institutions to counter the market advantages enjoyed by repeat players (such as merchants) . . . ." Llewellyn Joseph Gibbons, *Creating a Market for Justice; a Market Incentive Solution to Regulating the Playing Field: Judicial Deference, Judicial Review, Due Process, and Fair Play in Online Consumer Arbitration*, 23 NW. J. INT'L L. & BUS. 1, 6-7 (2002).

[51] We use the term "payment mechanism provider" to cover associations, such as VISA and MasterCard; historically closed networks, such as American Express and Diners' Club; and the new, third party networks, such as FirstData, which link the participants in card-based transactions.

A. *Payment Systems*

We define a payment system as the combination of law, contracts, and physical technology that enables the movement of value from one party to another to meet the objectives of the parties involved.[52] Money is itself a payment system. So are a VISA credit card, a MasterCard debit card, a traveler's check, a bank check, a money order, a stored value card, and a certified check, to name but a few. Different payment systems have varied features and attributes, enabling the users of that payment system to accomplish particular objectives in different ways. Where different payment systems are available, users select among them based on the combination of costs and benefits each payment system offers.

Consider one of the simplest payment systems available: U.S. dollars.[53] In a typical cash transaction conducted in dollars, a consumer offers a merchant dollars for goods or services. If acceptable,[54] a merchant transfers the goods or services to the consumer in exchange for physical Federal Reserve notes.[55] The notes the merchant receives are fungible; that is, they are functionally identical to other, similarly valued notes in circulation. The merchant can then exchange these notes for goods and services from others.

To allow a comparison between dollars and other payment systems, consider the features the dollar offers. There are technical features of Federal Reserve notes which facilitate transactions. The notes include antifraud technology, such as complex printing techniques, watermarks and serial numbers.[56] Inexpensive technology is available to the note-receiver

---

[52] For an overview of some of the characteristics of payment systems, *see generally* Jane Kaufman Winn, *Clash of the Titans: Regulating the Competition Between Established and Emerging Electronic Payment Systems*, 14 BERKELEY TECH. L.J. 675 (1999).

[53] *See* Henry H. Perritt, Jr., *Legal and Technological Infrastructures for Electronic Payment Systems*, 22 RUTGERS COMPUTER & TECH. L.J. 1, 5 (1996) ("The definitive payment system is money.").

[54] In the United States, prior to 1862, a merchant could decide which currency to accept, and whether to accept currency at all. *See generally* Lewis D. Solomon, *Local Currency: A Legal and Policy Analysis*, 5 KAN. J.L. & PUB. POL'Y 59 (1996). Once the Civil War era currency laws were adopted, merchants no longer had a choice. Outside of the United States, merchants make this determination every day. Will they accept their own local currency, or some other global currency (i.e., dollars, Euros, Yen, etc.)? Two parties freely elect to accept that currency—with all of the risks inherent in accepting that currency and cash in general.

[55] For a summary of legal issues surrounding the merchant acceptance of paper and electronic currency, *see* James S. Rogers, *The New Old Law of Electronic Money*, 12-15, 35-52 (Boston College Law School Research Paper No. 62, 2005), *available at* http://ssrn.com/abstract=680803.

[56] *See* Bureau of Engraving and Printing, Anticounterfeiting: Security Features, http://www.moneyfactory.com/section.cfm/7/35 (last visited February 7, 2005); Homer Brickey, *Credit Card Firms Battle Crooks with Technology*, THE PATRIOT LEDGER, June 24, 1995, at 27, *available at* 1995 WL 8199543 (describing antifraud technology).

to evaluate the notes' genuineness.[57] Even here, reputation plays a role: the notes' continued value depends on the reputation and credibility of the issuer.[58] Cash also makes anonymous transactions possible and enables merchants to under-report income to tax and regulatory authorities.[59] Further, the issuer profits from the circulation of the notes (known as seigniorage).[60] The use of currency has implications for loss allocation, if the currency turns out to be counterfeit and the consumer who used it cannot be located, the merchant bears the cost.

Federal Reserve notes do not provide any dispute resolution characteristics. That is, people cannot invoke the jurisdiction of U.S. government courts merely by denominating a transaction in dollars. Thus, a transaction conducted entirely in dollars among Guatemalans in Guatemala does not allow any of the parties to bring an action in the U.S. courts, although the Guatemalans may make use of the antifraud technology built into the physical notes (i.e. use the detector pens to verify the bill's genuineness) and the U.S. government profits from the seigniorage produced by dollars circulating outside the United States. Transactions originating outside the U.S. can also take advantage of the reputation-based value retention of the currency by using dollars as the basis for transactions physically located outside the United States.[61]

---

[57] *See, e.g.*, http://www.centercoin.com/coin_supplies/counterfeit_detector.htm (last visited February 8, 2005) (selling for $4.95 each, "[d]etector pens are an inexpensive and reliable screening device to be used in conjunction with other counterfeit detection methods"); Wikipedia, Counterfeiting, http://en.wikipedia.org/wiki/Counterfeit (last visited February 8, 2005) (describing historical methods to defeat counterfeiting).

[58] *See* Melvyn King, *The Institutions of Monetary Policy*, *available at* http://www.bankofengland.co.uk/publications/speeches/2004/speech208.pdf) (last visited February 8, 2005) (speech by Bank of England Governor describing importance of reputation of monetary institutions).

[59] *See* Richard R. Holley III, *One-Card 101: Wachovia Hits the Pit and Becomes the Partner Bank of the UNC One Card*, 4 N.C. BANKING INST. 371, 381 (2000) (noting that experience with stored value cards in 1996 Olympics revealed that merchants viewed anonymity as an advantage of cash over card-based systems).

[60] *See* Federal Reserve Bank of Cleveland, Governments and Money, http://www.clevelandfed.org/annual/essay.htm#compcur (last visited February 8, 2005) ("The authority to print money allows the government to raise revenue because the cost of producing the money itself is far less than the government's command over the purchase of goods and services."). *See also* Paul D. Glenn, *Electronic Banking*, Glasser LegalWorks, *available at* ELCEC GLASS-CEC 85, 105) (1998) ("The government has a lot to lose from the benefits of seigniorage if its role is somehow usurped . . . Most of the $23 billion that the [Federal Reserve Board] returned to the Treasury last year comes from the seigniorage attributable to the approximately $400 billion of cash outstanding at any one time.").

[61] The recent decline of the dollar, which has created uncertainty over the future value and stability of the dollar, against many currencies has raised concern over whether the dollar can maintain its position as a desirable currency for non-U.S. transactions, potentially costing the U.S. some of its seigniorage revenue.

Using cash as a payment system has transactions costs, however.[62] Physically transferring cash for large payments can be expensive.[63] Cash's anonymity makes recovering stolen cash difficult. And in single-instance cash transactions between strangers, the parties lack reputational incentives to deal fairly with one another.[64] By adding additional features that reduce transactions costs relative to those offered by cash, alternative payment systems may provide buyers and sellers with sufficient additional value to allow an entrepreneur to make a profit from charging for the use of the payment system and still offer the user a transaction cost below that of cash. For example, cash transactions over large distances are vulnerable to the theft of the currency while it is in transit.

An early solution to the problem of transporting cash over large distances was the development of the money order.[65] Money orders provided a substitute for cash that offered the holder protection against theft or loss. However, money orders also had the potential for fraud, since a purchaser could alter a money order after purchase and, if successful, cash it for more than he or she had paid for it. To prevent such fraud, the U.S. post office required that money orders issued by it be cashed only at designated post offices. By sending the cashing post office a separate communication from the issuing post office confirming the value, the potential for fraud was reduced. These restrictions reduced the usefulness of the postal money order, however, because they raised the cost to the recipient in cashing the money order. Competing with the post office for the money order business, American Express developed a secure money order in 1881 whose design

---

62 On the costs of cash, *see* Walter A. Effross, *Putting the Cards Before the Purse?: Distinctions, Differences, and Dilemmas in the Regulation of Stored Value Card Systems*, 65 UMKC L. REV. 319, 325-26 (1997). The development of uniform units of currency reduced the transactions costs of cash. After describing the variety of currency circulating in British colonies in the Caribbean in 18th and early 19th centuries, for example, one author notes that "[w]ith all this complicated variety of currency it must have been extremely difficult undertaking a simple transaction like buying a drink, and the temptation to swindlers was great." CYRIL HAMSHERE, THE BRITISH IN THE CARIBBEAN 156-57 (1972).

63 *See* Michael K. Salemi, *Hyperinflation, in* THE CONCISE ENCYCLOPEDIA OF ECONOMICS, *available at* http://www.econlib.org/library/Enc/Hyperinflation.html (last visited Feb. 8, 2005); ADAM SMITH [GEORGE J.W. GOODMAN], PAPER MONEY 57 (1982) ("In 1923, at the most fevered moment of the German hyperinflation, the exchange rate between the dollar and the Mark was one trillion Marks to one dollar, and a wheelbarrow full of money would not even buy a newspaper."). In general, cash is too expensive to use for large payments even in the absence of hyperinflation. *See* David A. Balto, *Can the Promise of Debit Cards Be Fulfilled?*, 53 BUS. LAW. 1093 (May 1998) ("To oversimplify grossly, cash cannot be used for large transactions.").

64 Cash has other transactions costs, including the possibility of the transmission of disease, or at least the fear of transmission. *See* John Dorschner, *Flu Scare Leaves Everyone Waiting*, FORT COLLINS COLORADOAN, Oct. 18, 2004, at 1E. Whether or not there is a scientific basis for the concerns about how disease may have been transmitted through money, some have a fear of this factor, raising the cost (to those market participants) of using cash.

65 PETER Z. GROSSMAN, AMERICAN EXPRESS: THE UNOFFICIAL HISTORY OF THE PEOPLE WHO BUILT THE GREAT FINANCIAL EMPIRE 83 (1987).

reduced fraudulent use by incorporating the amount physically into the money order, eliminating the incentive for forgery because tampering would only reduce the value of the money order while eliminating the need to visit the post office to cash the money order.[66]

Similarly, travelers have long used letters of credit to secure cash outside their home banking areas. With a letter of credit, after depositing money with a U.S. bank before leaving the country, a traveler received a letter stating the amount available. By presenting this letter to a foreign bank which had an agreement with the U.S. bank, the traveler could receive cash. When a foreign bank was presented with the letter, it compared the signature of the issuing bank official to a file signature to determine the letter's validity, then noted the amount withdrawn on the letter. Unfortunately this required considerable time at each bank, as everyone "from the charwoman up" would be asked to verify the signature.[67] Again, American Express innovated, creating the traveler's check as a substitute for the letter of credit.[68] A crucial feature was American Express's agreement to guarantee the recipient of traveler's checks against fraud and currency risk.[69]

We can compare the payment systems' technologies in these two examples to illustrate how a payment system offers a mix of characteristics which affect its relative advantages in the marketplace. In the case of money orders, the post office had a secure technology (the two-part transmission of information on the value of the money order) which depended in part on the limitation of locations where the money order could be cashed.

---

66 *Id.* at 83 ("On the left side of the money order . . . [the company] placed nine columns of figures [which] . . . depicted all 5-cent denominations from $1 to $10, the maximum amount of the first express money orders. When a customer purchased an order, the express clerk wrote the name of the payee and the amount on two stubs, and gave one to the buyer and kept the other for company records. But instead of writing an amount on the [money order] itself, he cut the protective margin to the designated sum. The customer could no longer raise the value of the order because the figures simply were no longer there.").

67 *Id.* at 88.

68 *Id.* at 91; Dean Perritt suggests that the development of the traveler's check was tied to the creation of a national currency. PERRITT, *supra* note 53, at 10 ("Nationalizing the currency had the effect of encouraging the invention of a variety of quasi-currencies, most significantly, personal checks and traveler's checks. . . . Historically, the traveler's check was a consolidation of a traveler's letter of credit and separate drafts drawn upon the authority created by the letter."). We see no reason in principle why a traveler's check requires a *national* currency; we read Perritt's argument to be that the key was the development of a national banking system and the legal disadvantages imposed on private notes as part of the effort to create the national currency.

69 GROSSMAN, *supra* note 65, at 91 ("By offering such a sweeping guarantee . . . Amexco made the [traveler's check] not just another money order, but rather a kind of universal currency."). When World War I broke out, American Express successfully honored its promise to pay on traveler's checks in Europe by building up reserves in the months before the war began. *Id.* at 123. As a result, European merchants began to prefer to hold traveler's checks to holding, for example, French currency. *Id.* at 124. The company also later added an inspector's office to track down fraud involving traveler's checks. *Id.* at 212.

This approach offered significant advantages over the use of cash for long distance transfers by reducing the risk of loss to the sender, at least partially transforming the sender's risk of physical loss or robbery into a risk of fraud borne by the post office. The post office's technology then reduced the risk of fraud to a manageable level. The technology did so at the cost to the consumer of restricting the locations at which the money order could be cashed. American Express's innovation was to develop a technology which enabled it to sufficiently lower the risk of forgeries to permit it to broaden the network of locations at which the money order could be cashed without incurring losses. It did so by reducing the risk of fraud through characteristics of the document itself, eliminating the need for the second transmission of information which had made the postal money order secure.

Similarly, the development of the traveler's check reduced the chance of fraud in a letter of credit by eliminating the need for the bank officials' signatures to be on file in the receiving bank, by putting the authorization signature on the traveler's check itself via the requirement of two signatures by the customer. Moreover, by requiring customers to prepay for the traveler's checks (whose face value would be difficult to alter because of the use of preprinted values), American Express captured the "float" for itself. Finally, American Express transformed the potential dispute over a letter of credit into a smaller set of potential disputes. The only issues with respect to a travelers' check were the validity of the customers' signatures, something the technical specifications of the check itself made simple to verify, and the validity of the travelers' check itself, something that the technical specifications of the document also made easy to verify through counterfeit-defeating printing techniques. Letters of credit, on the other hand, were open to dispute as to the amount authorized, as well as the disputes possible with travelers' checks. Moreover, by transforming the travelers' check transaction from consumer-domestic bank-foreign bank-consumer into one of consumer-domestic bank-American Express-foreign bank-consumer, American Express inserted itself as a repeat player into every transaction involving a traveler's check, giving it leverage to control the terms of the banks' behavior and using its reputation to guarantee payment.[70] Because individual domestic banks' pair-wise transactions with foreign banks were limited in number, the banks had not been able to exert such leverage. This enabled American Express to insist on conditions that minimized the opportunities for fraud.

More complex payment systems are ubiquitous, although often invisible to the average consumer who benefits from their use. Figure 1 summarizes some of the payment systems used in different contexts in the U.S. economy today.

---

[70] *See* Perritt, *supra* note 53, at 19 ("American Express traveler's checks are accepted because the market trusts that American Express will remain solvent.").

FIGURE ONE

Business	Factoring, LOC, Asset Lending	**b2c:** Private Label, Fuel	**b2b:** Fuel, EDI "terms"
Consumer	LOC, Home Mortgage, etc.	**c2c:** PayPal, IOU, HSBC	**c2b:** Terms, IOU, Private Label
Financial Institution	ACH, FedWire, Swift	VISA, MC, Amex, JCB, etc.	Purchasing Card, etc.
	Financial Institution	**Consumer**	**Business**

Many payment systems described in the business-to-business and business-to-financial institution segments of Figure 1 also have a relatively long history. Letters of credit, for example, date to at least the fifteenth century[71], and individually negotiated contracts ("terms") between businesses with repeated interactions have long existed.[72] Payment systems may also develop as part of financial services. For example, businesses have for centuries monetized inventory and accounts receivables through transactions with financial institutions.[73] Legal rules for such practices reduce transactions costs, making possible transactions that would otherwise not occur, but also introduce new potential for fraud by expanding the set of possible

---

[71] *See* EDWIN S. HUNT & JAMES M. MURRAY, HISTORY OF BUSINESS IN MEDIEVAL EUROPE, 1200-1550 at 66 (1999).

[72] For example, a trucking company might arrange an account system with a truck stop chain to allow its trucks to refuel without the necessity of the truck drivers making a payment, with the truck stop chain doing consolidated monthly billing for all fuel purchases.

[73] These transactions are not typically thought of as a payment system because such transactions are limited to the business-financial institution relationship. As one of the earliest examples of payment system development, however, we think they deserve mention and fall within our definition. For example, a business with a capital need will have a financial institution purchase the business's accounts on a nonrecourse basis and provide "specific credits for the business and function as servicer on the accounts. Thus, traditional factoring is a nearly complete outsourcing of all of the credit functions of the originator. The outsourcing itself provides the source of value in a factoring transaction. Given the factor's expertise in credit determinations and the economies of scale produced by servicing accounts from a number of originators, the factor places a higher value on the accounts than does the originator." Christopher W. Frost, *Asset Securitization and Corporate Risk Allocation*, 72 TUL. L. REV. 101, 146 (1997).

transactions.[74] The demand for the financial service of monetizing inventories produced the legal rules, including the development of financial instruments that facilitated such practices.

The structure of payment systems can have important implications beyond particular transactions. For example, the introduction of credit cards dramatically changed consumer credit. Before credit cards, "most consumer loans were made on a secured, installment basis. Each time a consumer wanted to borrow money, he or she had to reapply to the bank and go through the application and approval process again."[75] With the shift to credit cards as a payment system came the transformation of most consumer debt into unsecured, revolving debt.[76]

A full scale survey of the development and spread of the increasingly varied payment systems in use is beyond the scope of this paper. The relevant points for our purposes are that (1) payment systems offer a bundle of services which include, but generally are not limited to, the transfer of value; (2) different systems offer different bundles of services and costs; and (3) different systems' bundles create different incentive structures for the participants in transactions with respect to the resolution of disputes.

Since World War II, card-based payment systems have expanded into the business-to-consumer relationship and, more recently, into the consumer-to-consumer relationship. Banks and merchants began issuing credit cards of various types in steadily increasing numbers until today's explosion of consumer credit and debit cards has filled our mailboxes with offers on an almost daily basis.[77] Card-based payment systems increasingly substitute for other payment mechanisms (e.g. cash and checks).[78] In the consumer-to-consumer market segment, PayPal offers payment systems with many of the features of card-based payment systems (although usually not including the physical card) that allow individuals to pay other individuals.[79] Our focus is on card-based payment systems. These systems offer

---

74 American Express had a small commercial credit program built around providing credit based on inventory starting in the early twentieth century. *See* GROSSMAN, *supra* note 65, at 194. The company later had a major fraud problem in connection with a "field warehousing" commercial credit inventory lending product, begun in the 1940s, when $150 million in salad oil was discovered to be missing from a tank farm due to fraud. *Id.* at 248, 318-327. The company ultimately paid out $60 million to settle claims related to that fraud. *Id.* at 327.

75 AURIEMMA, *supra* note 5, at 3.

76 *Id.*

77 *See Credit Card Offers Again Filling Mailboxes*, 14 No. 14 CONSUMER BANKR. NEWS (LRP) 2 (June 24, 2004) (reporting that U.S. households received more than 1.2 billion direct mail credit card solicitations during the first three months of 2004).

78 *See* EVANS & SCHMALENSEE, *supra* note 1, at 91 ("Payment cards are substitutes for cash, checks, and other means of exchange.").

79 As part of the dot-com expansion and later implosion, there were many companies providing a similar service to PayPal, including Citigroup-sponsored c2it and HSBC-sponsored Yahoo Direct. *See* Rogers, *supra* note 55, at 16 (discussing rapid change in e-payment systems). All of these companies offered person-to-person money transfer services, typically through email accounts. Unlike with plastic

important lessons for dispute resolution generally because of their success in reducing costs in dispute resolution and in aligning the incentives of the parties. Ultimately, their approach facilitates inexpensive resolution of the disputes.

B. *The Structure of the Technologies*

The technologies of card-based payment systems have both differences and similarities across systems. Some of these technical structures affect the incentive structure of the dispute resolution systems. In this section we examine the technical structures for their influence on the successful features of the dispute resolution systems.

1. Similarities Among Card-Based Payment Systems

In all types of card-based payment systems, both buyers (consumers) and sellers (merchants) must agree to use the system.[80] In addition, the system itself must cover its costs of operation. The system must therefore offer a bundle of services and fees that is attractive to both buyers and sellers, generates sufficient revenue to fund the system's operating costs, and be competitive with the alternatives.

Because there are two independent, distinct sets of customers with different (and often opposing) interests involved in card-based payment systems, an important feature of card-based payment systems is how they manage these conflicting interests, including whether they prevent either interest group from gaining an advantage in the system as a whole.[81] More-

---

cards, the email account acts as the alternative token to the plastic card. Presumably, transfers through PayPal are completed to offset an underlying exchange of goods or services. In cases where the payment provider does not have any information about, and is not responsible for the goods and services delivered (like the basic PayPal case), these payments are usually referred to as Quasi-Cash. *See* ERIC JACKSON, THE PAYPAL WARS: BATTLES WITH EBAY, THE MEDIA, THE MAFIA, AND THE REST OF PLANET EARTH 9 (2004) (describing market niche for PayPal).

[80] *See, e.g.*, Paymentech Merchant Application and Agreement, Terms and Conditions for Merchant Agreement, §1.3 ("You agree to comply with all Association Rules and Operating Guide procedures, and with such other procedures as we may form time to time prescribe for the creation or transmission of Sales Data. We may modify or supplement the Operating Guide in order to comply with requirements imposed by the Association Rules.") and §5 ("There may be a chargeback under any of the following circumstances, or as the Association Rules and operational requirements dictate from time to time. Consequently, additions and/or deletions to this list may occur.") (copy on file with authors).

[81] David S. Evans, *The Antitrust Economics of Multi-sided Platform Markets*, 20 YALE J. REG. 325, 331 (2003) ("A fundamental insight of the theoretical research is that these businesses need to determine an optimal pricing structure—one that balances the relative demands of the multiple customer groups—as well as optimal pricing levels. That insight has implications for many other strategic variables. Empirical examination of these industries finds that key business decisions are driven by the need

over, the competition of systems for both merchant and consumer accounts is critical to ensuring the fairness of the systems' features. We therefore describe the systems' operation from the point of view of both merchants and consumers.

To merchants, card issuers offer[82] payment guaranteed against certain forms of loss, such as consumer bankruptcy;[83] relief from the need to collect on accounts receivable; faster payment (referred to as reduced days outstanding); automatic deposit of receivables;[84] receivable financing;[85] consumer credit services;[86] fraud identification tools;[87] fraud prevention tools, such as card security features;[88] advance commitment to adopt new technologies as mandated by the payment mechanism provider, thus creating

---

to get critical levels of multiple customer groups on board and to balance complementary customer communities.").

82 For an exhaustive analysis of the payment system options available to merchants, *see* Merchant Seek, Debit Card & ATM Processing, *available at* http://www.merchantseek.com/debitcard.htm (last visited on February 14, 2005); American Express, Online Merchant Services, *available at* http://home3.americanexpress.com/uk/merchant/manage/manage_default.asp?manage_body=learnOMS_body (last visited on February 22, 2005); John Burtzloff, *Accepting Customer Payments*, Entrepreneur.com, *available at* http://www.entrepreneur.com/article/0,4621,305819,00.htm (January 13, 2003).

83 EVANS & SCHMALENSEE, *supra* note 1, at 31 ("payment cards provide a means of insuring against consumer defaults"). Insurance against consumer default is not a trivial benefit—approximately one third of VISA's issuer costs in the quarter ending 1996, for example, were due to fraud or uncollected bills to consumers, making net charge-offs the second largest expense for issuers after the cost of funds. *Id.* at 214.

84 Automatic deposit of receivables can be a significant element of the value proposition to merchants with high volumes of sales. For example, on average, in 2004, Wal-Mart earned approximately $29.2 million in sales every hour. If the electronic payment systems did not exist, Wal-Mart would have to implement a significant cash logistics system, including physical transportation, security, and more. Hourly sales are estimated by dividing Wal-Mart's annual revenue of $256.3 Billion, by 365 days per year, by 24 hours per day. This rough calculation works out to be $29.2 million. Note that in peak volume hours, Wal-Mart is likely to have substantially larger hourly revenue. *See* Wal-Mart Annual Report (2004), http://www.walmartstores.com/Files/annualreport_2004.pdf.

85 EVANS & SCHMALENSEE, *supra* note 1, at 5 (By taking charge cards, a merchant "makes a sale to someone who could not have paid cash and avoids having to offer financing.").

86 David S. Evans & Richard Schmalensee, *Economic Aspects of Payment Card Systems and Antitrust Policy Toward Joint Ventures*, 63 ANTITRUST L.J. 861, 890 (1995) ("[T]he issuers provide merchants with credit services that increase consumer demand and that the merchant might otherwise have to provide. Issuers also guarantee payment by assuming the vast bulk of credit and fraud losses."); *see also* CHUTKOW, *supra* note 4, at 58 (describing how credit cards improved consumer finance options for banks and consumers); *Id.* at 61 (quoting Bank of America official that credit cards were a "natural extension" of consumer lending); AURIEMMA, *supra* note 5, at 8 ("merchant [using credit card] has none of the risks inherent in extending credit or accepting checks").

87 EVANS & SCHMALENSEE, *supra* note 1, at 241 (describing VISA's "Cardholder Risk Identification System," which "uses computer neural network technology to predict the probability that a particular transaction is fraudulent and prompts the issuer to contact the cardholder if a certain threshold is exceeded.").

88 *See infra* notes 97 to 101.

network externalities[89] and avoiding first mover problems in technology adoption;[90] some marketing information about cardholders using the system; access to desirable groups of customers;[91] increased sales through increased credit;[92] and dispute resolution services.[93]

Competition for merchant business[94] drives merchant acquirers (including closed networks) to continually enhance the product they offer merchants, as well as compete on price and other terms.[95] For example, the

---

89 "There is a network externality when the value existing users get from the network increases when another user joins the network." EVANS & SCHMALENSEE, *supra* note 1, at 149.

90 *Id.* at 113 (associations "develop and encourage system-wide innovations in transaction processing").

91 Evans, *supra* note 81, at 353 ("There may be certain customers on one side of the market—Rochet and Tirole refer to them as 'marquee buyers'—that are extremely valuable to customers on the other side of the market. The existence of marquee buyers tends to reduce the price to all buyers and increase it to sellers. For example, American Express has been able to charge a relatively high price to merchants as compared to other card brands, because merchants viewed the American Express business clientele as extremely attractive. Corporate expense clients were 'marquee' customers that allowed American Express to raise its prices to the other side of the market, merchants.") (citations omitted).

92 CHUTKOW, *supra* note 4, at 62; AURIEMMA, *supra* note 5, at 7.

93 Many of these enhancements are discussed in more detail in section III, entitled Dispute Resolution Systems, below.

94 *See* AURIEMMA, *supra note 5,* at 27, 32 (2nd ed. 1996) (noting "fierce competition" for merchant business). In the early 1990s there were approximately 250 merchant acquirers, although the top twenty-five VISA and MasterCard merchant acquirers accounted for 79 percent of the total transaction volume. Evans & Schmalensee, *supra* note 86, at 866. Today approximately 80 merchant acquirers account for 95 percent of the market in the United States. Lloyd Constantine, Remarks at 2 Sided Market Conference, Columbia University School of Law, May 23, 2005 (notes from conference on file with authors). In addition to the competition for merchants by networks, there is fierce competition for managing merchants' private label (or store brand) cards. *See, e.g.*, W.A. Lee, *Citi Challenging Leaders in Private-Label Cards*, AMERICAN BANKER, Feb. 1, 2002, at 8 [2002 WL 4099786] (describing competition among processors and financial institutions for private label business).

95 This competition has been a feature of the marketplace almost since the beginning. For example, American Express initiated a cooperative advertising program in 1962 to win airline business, paying for airline ads that included a suggestion that passengers use their American Express card. *See* JON FRIEDMAN & JOHN MEEHAN, HOUSE OF CARDS: INSIDE THE TROUBLED EMPIRE OF AMERICAN EXPRESS 59 (1992). Similarly banks used lower fees to persuade merchants to accept debit cards in the late 1980s. As Prof. Evans explains

> [T]he debit card is an example in which different platforms made different pricing choices because they had different customers on board when they entered. In the late 1980s, ATM networks had a base of cardholders who used their cards to withdraw cash or obtain other services at ATMs. They had no merchants that took these cards. To add debit services to existing ATM cards, ATM networks charged a smaller interchange fee than did credit card systems to encourage merchants to install PIN pads. Compared to credit card systems' interchange fee of 38 cents on a typical $30 transaction, ATM networks only charged 8 cents. (On debit and credit transactions, the interchange fee is paid by the merchant's bank to the cardholder's bank. A lower interchange fee will tend to lower prices on the merchant's side and to raise them on the cardholder's side.) The PIN pads merchants installed could read the ATM cards that cardholders already had and accept the PINs they used to access ATMs. In response to ATM networks' low interchange fee, many merchants invested in the PIN pads, whose numbers increased from 53,000 in 1990 to about 3.6 million in 2001. In contrast to the credit card systems, which already had a base of merchants who took their cards and con-

physical cards have changed over time to better prevent fraud.[96] Thus, in the 1980s, the payment mechanism providers mandated the addition of the "Card Verification Value" numbers to the back of the card ("CVV1") and the magnetic strip ("CVV2").[97] The CVV numbers made it more difficult for a thief not in possession of the physical card to secure enough information to make a fraudulent purchase.[98] For example, if a thief stole a credit card receipt from a consumer's trash, the thief would have the consumer's name as it appears on the card and the card number.[99] He would not, however, have the CVV numbers, which are not printed anywhere on the receipt and should not be stored in any third party information system.[100] In response, of course, criminals created more sophisticated methods of gaining card information.[101]

The appearance of these enhancements is important because merchants, not merchant acquirers or card issuers, bear the burden of paying for

---

sumers who used them, ATM systems had to persuade banks to issue debit cards and cardholders to take these cards. Their strategy worked: The number of VISA debit cards in circulation increased from 7.6 million in 1990 to about 117 million in 2001.

Evans, *supra* note 81, at 352-53 (citations omitted). Competition extends beyond large merchants. *See, e.g.*, Steve Watkins, *Nashville, Tennessee Small Merchants Provide Steady Business to Credit Card Processor*, INVESTOR'S BUSINESS DAILY, Jan. 28, 2004, at A07 (describing iPayment, Inc.'s strategy of focusing on small business accounts with an average annual charge volume under $250,000 but with a total annual charge volume of $4 billion).

96 *See infra* note 184.

97 CHUTKOW, *supra* note 4, at 188; *id.* at 194 (describing CVV system).

98 Homer Brickey, *Credit Card Firms Battle Crooks with Technology*, THE PATRIOT LEDGER, June 24, 1995, at 27 ("[t]he best deterrent so far is a separate number used in a card-verification system, usually three digits and based on a mathematical algorithm (formula) known only to the card issuer. 'That has really put a dent in fraud,'" said John McKnight, regional director of fraud control for VISA.).

99 Beginning in 2005, the associations bar merchants from printing the full card number on the receipt. *See, e.g.*,MasterCard International Inc., Global Operations Bulleting No. 3, Mar. 1, 2005, at 88 ("Newly installed, replaced, or relocated point-of-interaction (POI) terminals, whether attended or unattended, must produce receipts that reflect only the last four digits of the primary account number (PAN). Fill characters such as X, *, or # must replace all preceding digits.") (copy on file with authors); Global Payments, Inc. Card Account Information Truncation Requirements: Suppression of Account Information on Transaction Receipts, http://www.globalpaymentsinc.com/myglobal/industry_initiatives/card_trancation_requir.html (describing association rules). State governments created a variety of requirements on account number truncation, ultimately preempted by the Fair and Accurate Credit Transactions Act of 2003, Pub. L. 108-159, 117 Stat. 1952.

100 Another example is VISA's deployment of the "Payment Service 2000" in 1992, which required members to upgrade their processing capabilities. The "Payment Service 2000" increased VISA's competitiveness against MasterCard and American Express and also reduced fraud. EVANS & SCHMALENSEE, *supra* note 1, at 204-05 (describing rollout of system).

101 Chutkow describes several of these methods. One example is the use of a separate reader in situations where the card is used out of sight of the cardholder (e.g. restaurants). The card is swiped through not only the point-of-sale terminal but a reader that captures the information from the card; this information is then used to create a fraudulent card that appears legitimate. Other schemes include the use of cameras to capture cardholders' entry of PIN numbers. CHUTKOW, *supra* note 4, at 185-86.

many types of fraud.[102] Innovations to reduce those fraud costs borne by merchants do not *directly* benefit any of the parties in direct control of the payment system: merchant acquirers, issuers, or payment mechanism providers. However, card networks that offer superior fraud prevention measures such as CVV numbers have a competitive advantage in gaining merchant business. For example, if a new anti-fraud feature is introduced by American Express, merchant acquirers who process VISA and MasterCard transactions are at a competitive disadvantage in gaining both new accounts and transactions from their existing customers.[103] Market forces have thus driven the merchant acquirers and payment mechanism providers to enhance consumers' cards for the benefit of both consumers and merchants, even though neither consumers nor merchants have a direct voice in the design and operation of the card-based payment systems, as we discuss below.

Merchant acquirers profit from their merchant accounts primarily through a variety of charges, including discount rate,[104] equipment rental fees, membership fees, service fees, and dispute resolution-related fees.[105] Fees differ substantially between institutions and between merchants. Wal-Mart, for example, likely pays lower per transaction fees than most small merchants, because Wal-Mart's volume is substantially larger, giving it leverage in negotiations with a merchant acquirer that Joe's Deli lacks.[106] Some merchant acquirers require equipment rentals, others allow merchants to purchase their own equipment. Most importantly for our purposes, merchant acquirers charge for various aspects of processing disputes.

---

102 Some types of fraud are paid for by the issuer bank. Thus when a card number is stolen and a counterfeit card is created, the issuing bank usually bears the loss. *See id.* at 186.

103 Recently, VISA introduced "Verified by Visa" and MasterCard introduced SecureCode. Both of these programs utilize an authentication technology to verify the cardholder's identity, at least verification of the registered identity on file with the issuer. When SecureCode authentication is offered by the merchant and merchant acquirer, the association rules shift liability for fraudulent transactions from the merchant to either the issuer or the cardholder. Although there are many flaws associated with these two programs, these two initiatives have made both VISA and MasterCard more attractive to certain merchants who suffer from high rates of fraud and chargebacks. The smartcard initiative (EMV) holds similar promise for certain kinds of fraud.

104 The discount rate includes an interchange fee that is passed from the merchant acquirer through the association to the issuer, to compensate the issuer for the float and certain cardholder risks. *See* MasterCard International, *MasterCard Dictionary*, December 2003, Discount Rate, at 34; and Interchange, at 52-53. *See also*, VISA U.S.A. INC. OPERATING REGULATIONS, VOLUME I—GENERAL RULES, MAY 15, 2004, [hereinafter, VISA GENERAL RULES, VOLUME I], §A, Definitions, Interchange, at A-27. *See also,* MasterCard International, *Quick Reference Booklet*, October 2003 [hereinafter, MasterCard Quick Reference Booklet], §5, Interchange Rate and Other Fee Programs, page 5-1 (describing interchange generally).

105 *See* EVANS & SCHMALENSEE, *supra* note 1, at 213 (describing various merchant fees).

106 Wal-Mart, for example, is estimated to do 10% of all PIN debit volume in the U.S. David Sibley, John Michael Stuart Centennial Professor in Economics, University of Texas in Austin, Presentation at the Columbia University School of Law Two-Sided Market Conference (May 23, 2005).

For example, typical merchant acquirers in the United States charge merchants a discount rate of between 1.5% and 3% of the purchase price.[107] Thus in a typical $100 charge by a consumer, the merchant acquirer will pay to the merchant between $97 and $98.50 of the $100, retaining $1.50 to $3.00 as its fee. The merchant acquirer will then have to forward the interchange fee to the payment mechanism provider, most of which is then passed along to the issuer.[108] The interchange expense varies by the type of merchant, risk, and other factors.[109]

In addition to monetary charges, each network has its own system of rules governing everything from the physical attributes of cards[110] to how disputes are handled.[111] Merchants must adhere to the operating rules of the payment mechanism provider, through the terms of the contract between the merchant acquirer and the merchant.[112] For example, the two leading associations (VISA and MasterCard) also require the merchant acquirer to have a merchant agreement with every merchant it services and to incorpo-

---

107 *See* MerchantSeek Merchant Account Rates, www.merchantseek.com/merchant_accounts_rates.htm (last visited April 23, 2005) ("a typical discount rate for U.S. businesses is right around 2.49% . . . Non-US businesses will pay a higher discount rates [sic] closer to the 3 % to 4% range.").

108 *See* Evans & Schmalensee, *supra* note 86, at 890 ("both MasterCard and VISA have set interchange fees so that the payment goes from the merchant side to the issuer side."); Evans, *Antitrust Economics, supra* note 81, at 376 ("A higher interchange fee tends to raise merchant fees and lower cardholder fees."); *Id.* at 375 ("Charge card systems—such as Diners Club and, historically, American Express—set these fees so that merchants contributed the preponderance of fees.").

109 *See* Tim Miller, *Explaining Credit Transaction Fees*, ENTREPRENEUR.COM, Aug. 27, 2001, http://wwww.entrepreneur.com /article/0,4621,292172,00.html (last visited April 23, 2005) ("Discount rates vary depending on the type of business, such as traditional brick-and-mortar business, a mail-order/telephone-order business, a restaurant or an e-business. Discount rates also vary depending on whether a card number is keyed into the point-of-sale terminal or swiped into the terminal."). Historically, interchange was designed to compensate issuers for the guaranteed payment, float, and the costs of extending credit to consumers. More recently, merchants have argued that they are indirectly funding cards with substantial benefit programs, far beyond the original purposes of funding float and risk. *See also* MasterCard Quick Reference Booklet, *supra* note 103, § 5 (explaining various regional and product based interchange programs).

110 *See* VISA GENERAL RULES, VOLUME I, *supra* note 104, § 10. *See also* MASTERCARD INTERNATIONAL, CARD DESIGN STANDARDS (2003).

111 *See* VISA GENERAL RULES, VOLUME I, *supra* note 104, § 7 (description of the Dispute Resolution process, generally); MASTERCARD INTERNATIONAL, CHARGEBACK GUIDE (rev. June 2004) [hereinafter MASTERCARD CHARGEBACK GUIDE] (guide to Chargebacks); and VISA U.S.A. INC., OPERATING REGULATIONS: VOLUME II, DISPUTE RESOLUTION RULES (May 15, 2004) [hereinafter VISA GENERAL RULES, VOLUME II] (rules associated with Chargebacks and Dispute Resolution).

112 *See, e.g.*, PAYMENTECH MERCHANT APPLICATION AND AGREEMENT, *supra* note 79, § 1.3 (Jan. 2004) (requiring the merchant "to comply with all Association Rules and Operating Guide procedures, and which such other procedures as [Paymentech] may from time to time prescribe for the creation or transmission of Sales Data."). *See also* VISA GENERAL RULES, VOLUME I, *supra* note 104 § 4.2.C (describing the minimum requirements of a Merchant Agreement between the Acquirer and the Association); MasterCard International, *Bylaws and Rules*, April 2004 [hereinafter, MasterCard Bylaws], § 9.1, (Apr. 2004) (description of the requirements for a Merchant Agreement and certain provisions).

rate the terms and conditions of the operating rules into that merchant agreement.[113]

On the consumer side, the value proposition is itemized billing with delayed payment, referred to in the industry as "country club style billing";[114] lifestyle financing (i.e. that you can consume a cruise today and pay for it over the next year);[115] record keeping features, including summary statements and interfacing with personal finance software;[116] fraud protection limiting losses to a small, fixed amount (imposed in some countries by government regulation);[117] point schemes; mail-order and Internet purchasing;[118] and dispute resolution processes. As on the merchant side, market forces have driven continual improvements in card features and contract terms that benefit consumers. For example, the practice of offering low introductory balance transfer interest rates to credit-worthy consumers has greatly lowered the cost of lifestyle financing for many,[119] while the use of prepaid cards has extended the convenience of card usage and the availability of alternate dispute resolution to customers with less desirable credit ratings.[120]

Issuers are paid for the value they provide to consumers through charges that include annual membership fees, interest on balances, transac-

---

113 *See* Mastercard Bylaws, *supra* note 112, §§ 9.1.1-9.1.2; VISA GENERAL RULES, VOLUME I, *supra* note 104, § 4.2.B.1.

114 AURIEMMA, *supra* note 5, at 7.

115 EVANS & SCHMALENSEE, *supra* note 1, at xi ("The millions of people who finance purchases on credit cards want to enjoy life earlier than their current incomes and savings permit."); CHUTKOW, *supra* note 4, at 61 ("enormous financial flexibility" introduced by credit cards.); AURIEMMA, *supra* note 5, at 6 ("the convenience and credit availability offered by bank credit cards were major contributors to their proliferation around the country.").

116 AURIEMMA, *supra* note 5, at 7.

117 *Id.* ("Bank cards also offer a comparatively safe means for conducting transactions. If currency is lost or stolen, the potential loss is much greater than if a credit card is stolen. It is more difficult to fraudulently use a credit card, and the cardholder's liability is limited if a card is misused.").

118 *Id.*

119 *See, e.g.*, Aviya Kushner, *Proceed with caution on balance-transfer 'deals'*, Bankrate.com, Financial Literacy in America, http://www.bakrate.com/brm/news/financial-literacy2004/balance-transfer1.asp (last visited April 25, 2005) ("'If you're a good, responsible consumer, it can be a great opportunity,' says Mark Oleson, director of Iowa State University's Financial Counseling Clinic.").

120 The use of prepaid cards, also called Stored Value Cards, is limited to the value on the card. So, individuals who cannot secure credit may still avail themselves of the payment system, by depositing funds onto such a card. These prepaid cards are proliferating into new uses that are targeting the cash economy. For example, some employers are now putting the entire value of a paycheck onto a Payroll card, for those employees who lack a bank account. The employee may then spend their paycheck, simply by using the card at a merchant, or withdrawing funds at an ATM. *See, e.g.*, Lavonne Kuykendall, *Banks Taking Baby Steps in Prepaid Debit Space,* AMERICAN BANKER, October 28, 2004, at 5.

tion fees (late fees, over limit fees, etc.).[121] Issuers are also compensated by the merchant acquirers through the interchange fee.[122] Many of these charges not only are subject to general competitive pressures—with banks competing, for example, to offer lower balance transfer costs to good potential customers—but also to individual negotiations between consumers and issuers.[123] Issuers also receive payments from organizations offering co-branded cards (e.g. professional associations, alumni associations, and the like).[124] Also as with merchant acquirers, the networks directly (closed networks) or indirectly (open networks) require consumers to agree to the relevant network rules, including the rules governing dispute resolution.[125]

All potential card users cannot bargain equally effectively with all card issuers. Consumers with better credit have greater bargaining power than those with poor credit scores.[126] Consumers that are more profitable than others also have greater bargaining power. Profitable consumers may not be the ones with the best credit scores.[127] As courts have recognized, however, the market provides many opportunities for consumers to obtain access to card-based payment systems.[128] It is not necessary that all consumers have equal bargaining power for consumer-issuer bargaining to effect

---

121 EVANS & SCHMALENSEE, *supra* note 1, at 147; AURIEMMA, *supra* note 5, at 26-27 (describing revenue sources for issuers).

122 EVANS & SCHMALENSEE, *supra* note 86, at 890 (payment of the interchange by the merchant acquirer to the issuer "reflects the fact that, in both [the VISA and MasterCard] systems, other rules require the issuers to bear far more of the costs and to take much more of the risk of producing the venture's interdependent payment service.").

123 *See, e.g., How to Negotiate a Better Credit Card Deal*, http://www.ehow.com/how_109504_negotiate-better-credit.html (last visited May 31, 2005) (describing how consumers can negotiate more favorable credit card deals.). Note that we would not expect to see identical products as consumers demand different cost structures depending on their circumstances. *See also*, EVANS & SCHMALENSEE, *supra* note 1, at 211 (distinguishing between fee structures preferred by those who pay off balances and those who carry balances).

124 AURIEMMA, *supra* note 5, at 12.

125 Issuers can modify the rules (including dispute resolution rules) even after the consumer has accepted the card's original terms in many states. *See, e.g.*, Edelist v. MBNA America Bank, 790 A.2d 1249 (Del. 2001).

126 *See* DEANNE LOONIN & CHI CHI WU, NAT'L CONSUMER LAW CTR., CREDIT DISCRIMINATION 100 (3d ed. 2002) (describing credit scoring). Once card issuers solved the problem of screening card holders to restrict the pool to those with desirable credit behavior, the list itself becomes valuable. As a result of American Express's success in this regard, for example, the brokerage firm of Shearson Loeb Rhoades accepted a merger with American Express in part to gain access to "the master list of AmEx's Gold Card holders." FRIEDMAN & MEEHAN, *supra* note 95, at 103.

127 *See Frontline: Secret History of the Credit Card* (PBS television broadcast, Nov. 23, 2004), *available at* http://www.pbs.org/wgbh/pages/frontline/shows/credit (last visited February 24, 2005).

128 *See, e.g.*, Johnson v. Chase Manhattan Bank, 231 N.Y.L.J. 19 (N.Y.Sup. Ct. March 11, 2004) (noting that "In any event, in this day and age when credit cards are rather easily available from any of a number of issuers, the fact that the customer who elected not to accept the [modification to the account agreement initiated by the issuer] would have to terminate his/her account, would not be grounds for concern.").

the terms of issuer-consumer contracts; what is important is the marginal consumer's ability to bargain. In the highly competitive issuer market,[129] this ensures that issuers continue to innovate to attract high value consumers as customers. These innovations then spread throughout the industry.

2. Key Differences Among Card-Based Payment Systems

There are two dimensions in which card-based systems differ from one another that are important for our analysis: network structure and payment timing. With respect to network structure, there are open and closed systems; these differ in the ownership and management of the relationships between the financial institutions[130] and the consumers and merchants. The entity maintaining the contractual relationship with the consumer is known as the "issuer" and the entity with the contractual relationship with the merchant is known as the "merchant acquirer" or "acquirer." An open system is a payment system where an association or third-party company maintains a contractual relationship with both the card issuer and the merchant acquirer. In an open system, issuers and acquirers may enter and exit that particular payment system according to the terms and conditions of membership.[131] Only the issuer maintains a direct contractual relationship with the consumer cardholder, and only the acquirer maintains a direct contractual relationship with the merchant. Open systems include associations such as VISA and MasterCard.[132] In a closed system, on the other hand, the issuer and the acquirer are the same financial institution; one entity main-

---

129 In the early 1990s there were approximately 7,300 VISA issuers, including approximately 100 national issuers such as Capital One, Citibank, MBNA, Bank of America, Household Bank (HSBC), US Bank, and others, who issue cards across the United States. Evans & Schmalensee, *supra* note 86, at 865-66.

130 We use the term "financial institution" to describe entities issuing cards, although not all such entities meet the federal regulatory definitions of the term. *See, e.g.*, 15 U.S.C. § 6809(3)(a) (defining financial institution as "as a business engaging in financial activities). Two major open card associations limit their membership to financial institutions who are licensed for financial activity (i.e., lending, deposit taking, etc.) in the country which they operate. *See* MasterCard Bylaws, *supra* note 112, art. I, § 1; VISA GENERAL RULES, VOLUME I, *supra* note 104, art. II, § 2.01.

131 There are terms and conditions for entry and exit into each payment system, and these conditions include fees and capitalization requirements. Entry into the MasterCard association may require a member initiation fee, a transfer fee, a sponsorship fee, and/or a portfolio acquisition fee. *See* MasterCard Bylaws, *supra* note 112, § 2.12. *See also* EVANS & SCHMALENSEE, *supra* note 1, at 233 (describing exit terms).

132 Some industry leaders assert that associations such as VISA are not "open" because membership is limited to financial institutions. Non-banks are not permitted to become a member. *See supra* note 130. The assertion that such associations are not open was recently made by Randy Gutierrez of Unicache at the Technology of Remittances Conference in San Francisco on December 11, 2004. For the purposes of this paper, we need not address this particular point in our analysis since the point is that the systems are open to entry and exit by multiple financial institutions, creating competitive pressures.

tains the relationship with both the cardholder and the merchant. Closed systems include programs such as department store cards, American Express,[133] and Discover. As we will discuss below, the incentive structures for various parties are different in open and closed systems.

With respect to payment timing, there are credit and debit cards. A credit card is a short term extension of credit by the issuer to the consumer.[134] The consumer is not obligated to immediately pay the full balance at the end of the month. If the consumer does pay the full balance owed, no interest is charged.[135] With a debit card, on the other hand, the money to pay the merchant is immediately[136] deducted from the consumer's bank account. Debit cards are expanding rapidly along with other cash-like substitutes, including prepaid and stored value debit cards where the consumer deposits cash and uses the card until the balance is zero.[137] As discussed below, different regulatory structures apply to debit and credit cards, influencing a number of attributes of the payment system. For example, the

---

133 Historically, American Express maintained an exclusive relationship with both the cardholder and merchant (a closed system). Internationally, American Express selectively issued cards under unaffiliated financial institutions. Domestically in the United States, this historical structure may be changing, as American Express recently offered a co-branded cards with MBNA and Citibank. *See MBNA Issues New American-Express Card Branded Credit Cards for More than 1,000 Affinity Groups* [hereinafter *American-Express Card Branded Credit Cards*], http://home3.americanexpress.com/corp//pc/2004/mbna_cards.asp (Last visited February 22, 2005). In this sense, American Express is moving from a closed system where it maintained the relationship with both the cardholder and merchant, to an open system, where it maintains a relationship with an issuer financial institution, who in turn maintains the relationship with the consumer.

134 AURIEMMA, *supra* note 5, at 2 ("Bank credit cards are a form of consumer loan, a revolving credit account that has a credit line of a specific amount that can be borrowed against in part or in full.").

135 For example, if a consumer has a zero balance on January 1, charges $10 on January 15, receives his statement for $10 on January 20, and pays the $10 balance before the due date (usually 30 days later), no interest applies. If, however, the consumer has a $5 balance on January 1, and the card charges a 1% per month interest rate, and then charges $10 on January 15, the consumer will owe the 1% interest on the $15. (This ignores the average daily balance method of calculating interest to simplify the example.) If the consumer charges an additional $10 on February 15 and pays the $15 balance in full by the due date, the consumer will still be charged interest on the February 15 $10 charge on the February statement. In short, the 30 day interest-free period applies only when the bills are paid completely and on time each month. *See* EVANS & SCHMALENSEE, *supra* note 1, at 141-44 (detailing differences between methods of calculating interest).

136 Generally within minutes but, depending on the technology used, it may take longer. When a PIN is used, the deduction is instantaneous. When the debit card is swiped like a credit card, it will take up to two days. *See* EVANS & SCHMALENSEE, *supra* note 1, at 299-300.

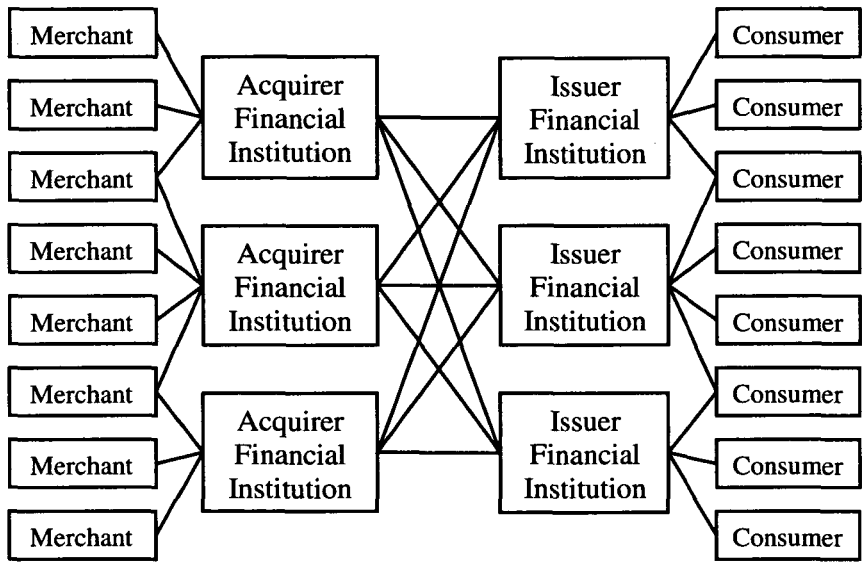
137 Stored Value cards, prepaid cards, and the like, are not regulated as depository accounts. *See* F.D.I.C. Gen. Couns. Op. No. 8 (1996), 61 Fed. Reg. 40490 (Aug. 2, 1996). *See also* F.D.I.C., Definition of "Deposit"; Stored Value Cards, 12 C.F.R. pt. 303, RIN 3064-AC80 (Apr. 26, 2004), *available at* http://www.fdic.gov/news/financial/2004/fil4404a.html (last viewed Sept. 23, 2005). For an interesting holiday shopping scenario where a family utilizes these forms of debit cards, *see* John Gould, *Ready to Explode*, http://www.intelcard.com/factsandfigures/03factsandfig.asp?A_ID=339 (last visited December 2004). A thorough account of legal issues involving stored value cards is Effross, *supra* note 62.

money flows from the consumer into the transaction pipeline more quickly than in a credit card transaction.

3. Networks

In addition to the card-based payment systems' differences, there are differences among *transactions* within the system. These differences depend upon the relationships between participants in the systems which are based on the type of network used to process the transaction. There are three types of networks over which these transactions travel: centralized, noncentralized, and hybrid. In a noncentralized network (see Figure 2), each bank has a connection to every other bank in the network.[138]

FIGURE TWO



Banks used noncentralized networks when they first began issuing card-based payment systems, with each bank contracting individually with other banks for the acceptance of the issuing bank's cards at the merchant acquirer banks' merchants. Thus, when a consumer used a decentralized net-

138 Historically, when banks privately issued their own banknotes, the settlements between them occurred through individual, bilateral exchanges of each others' notes. *See* Neal Stephenson, *The Great Simoleon Caper* (Spring 1995), http://www.cyberartsweb.org/cpace/scifi/cyberbib/Essays/SimoleonCaper.htm (last visited February 9, 2005) (explaining in detail how currency works through a fictional account of creation of private currency).

work card issued by Chase at a merchant acquired by Key Bank, the authorization and payment transactions would be routed through a private network between Chase and Key Bank. If the consumer later used the card at a merchant acquired by Wells Fargo, the transactions associated with that purchase would be routed through a private network between Chase and Wells Fargo. Similarly, if a different consumer used a card issued by Bank of America at the merchant acquired by Key Bank, those transactions would be routed through a third private network between Key Bank and Bank of America.[139] As the number of merchant acquirers and issuers contracting with each other grew, the number of private networks needed to make the noncentralized network function grew rapidly. (See Figure 2). Today, such bilateral agreements may be prohibited, at least in certain contexts, such as the processing and routing of transactions.[140]

In a centralized network, on the other hand, all contracts and transactions are with a central authority (e.g. VISA). (See Figure 3). The centralized networks offer significant cost savings to the participating institutions. Rather than individually negotiating contracts and maintaining infrastructure to support the flow of money and information between every pair of entities in the network, a participating institution has to comply with only one set of technical standards for the exchange of money and information, and negotiate only one set of contracts with the central entity.

The availability of hybrid and private networks limit the monopoly power of the payment mechanism providers both directly (in the case of open systems) and indirectly (via open systems' competition with closed systems) by allowing participants to route transactions outside any particular associations or network providers' system. It is necessary to consider how transactions are processed to fully understand this point.

When a consumer presents a card to a merchant and the merchant swipes the card through the point of sale terminal, the terminal reads various information from the card.[141] That creates an authorization transaction which is forwarded to the merchant acquirer's system. There, the authorization transaction is routed according to rules based on the bank identification number ("BIN") from the consumer's card. This number identifies the financial institution that issued the card and the appropriate 'routing,' to the merchant and merchant acquirer. The routing is the path the authorization transaction takes from this point to reach the issuer. It is dependent on the

---

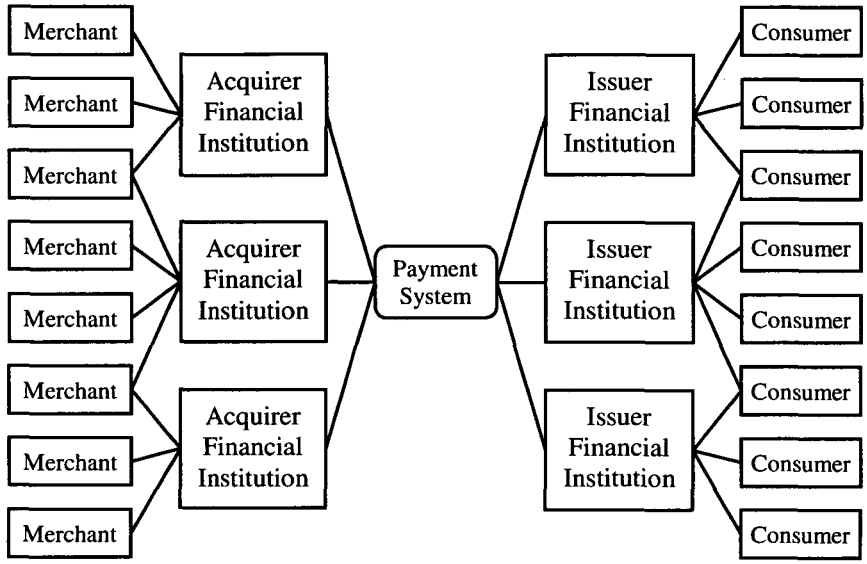
139 The dispute resolution processes, described below, would function differently under a bilateral processing approach. *See* discussion *infra* Section III, Dispute Resolution Systems.

140 *See* MasterCard International, *Cirrus Worldwide Operating Rules*, June 2004, §9, Processing Requirements (for a rule requiring certain cash machine transactions to be routed through MasterCard, having the effect of prohibiting certain bilateral processing agreements). *See also* MasterCard International, *Maestro Global Rules*, July 2004, §9; Steve Ruwe, *Required Processing of Visa Transactions through VisaNet*, Visa Bus. Rev., Visa U.S.A., Feb. 2005.

141 Terminals were introduced in the early 1980s. CHUTKOW, *supra* note 4, at 164-65 (describing introduction of the point of sale terminals).

contracts and rules of the participating institutions. For example, consider the case of a pure centralized network, such as VISA, with a consumer card issued by Chase and a merchant acquired by Fifth Third Bank. A VISA card issued by Chase and used for a charge submitted through a merchant acquired by Fifth Third Bank will have the authorization transaction routed from the merchant to Fifth Third Bank, then to VISA, then to Chase, and then the approval or rejection will reverse the path back to the merchant. The payment transaction follows the same path later that day from Chase back to the merchant's account with Fifth Third Bank. Payment by the consumer to Chase follows at a later date.

FIGURE THREE



Hybrid networks occur when an institution participating in a centralized network shifts some of its transactions from the network to a bilateral arrangement with another institution in the network, bypassing the central system for a subset of transactions. For example, Chase might determine that a large number of cardholder transactions occur with merchants in Athens, Greece acquired by the General Bank of Greece.[142] By negotiating a separate contract with General Bank of Greece and building the necessary infrastructure, including dedicated communications links, General Bank of Greece can dynamically route transactions involving Chase customers outside of the VISA network over a private network between General Bank of

[142] We have chosen actual bank names for our examples. The transactions described, however, are all hypothetical.

Greece and Chase. To do this, General Bank of Greece programs its computer to route transactions with Chase BINs directly to Chase rather than through the VISA network.[143] Chase and General Bank of Greece can then divide the savings from shifting those transactions away from the VISA network among themselves.[144] Although at present most of such agreements are between banks in different countries, because the savings from removing transactions from the network are greater when the network for charge for currency conversions is also present, there is no reason why such hybrid networks cannot exist domestically as well.[145]

Centralized networks benefit from the network externalities and transaction cost-saving features of centralization. There is thus a strong incentive for institutions to join centralized networks as networks grow in size and complexity. This incentive creates fears about monopoly abuses by the networks.[146] At least while financial institutions are not restricted in routing transactions across any network (of which that financial institution is a member), monopoly rents the network can obtain are limited by the potential competition allowed by hybridization of the network.[147] If the network abuses its position,[148] institutions will shift more business outside the network. Furthermore, most financial institutions are members of both VISA and MasterCard.[149] So, even in environments that restrict dynamic transac-

---

143 For example, the BIN table will direct that all card transactions prefixed with VISA BINS are routed to the VISA connection except BINs belonging to Chase, which are routed through the special connection.

144 At times, networks have attempted to prevent this by charging a fee regardless of whether the network was used. For example, Interlink attempted to impose such a charge in the debit card market but was forced by competition from Mastercard's Maestro network to drop it. *See* Balto, *supra* note 63, at 1096-97.

145 Such networks do exist for ATM/debit cards, where the merchant may route the transaction based on the cost of using a particular network. *See* EVANS & SCHMALENSEE, *supra* note 1, at 310. *See also* Ruwe, *supra* note 140 (notice that, with some exceptions, domestic Visa branded transactions must be processed through VisaNet, effectively prohibiting domestic bilateral transaction routing arrangements).

146 *See* Randy E. Barnett, *Pursuing Justice in a Free Society, Part One: Power vs. Liberty,* CRIM. JUST. ETHICS, Summer/Fall 1985; Randy E. Barnett, *Pursuing Justice in a Free Society, Part Two: Crime Prevention and the Legal Order,* CRIM. JUST. ETHICS, Winter/Spring 1986 (discussing problem of monopolizing entity in a competitive legal market).

147 Payment card system networks are vulnerable to competition. As Professors Evans and Schmalensee note,

> In network industries, as in industries with important scale or learning economies, we also tend to see temporary dominance of one or a few firms or networks. There is a temptation to think that such dominance will be permanent. But most network industries are not like manufacturing industries, in which ownership of capital-intensive capacity or key proprietary technology may give rise to long-lived dominance. As we shall see, the shares of payment card networks and individual issuers have varied considerably over time.

EVANS & SCHMALENSEE, PAYING, *supra* note 1, at 153.

148 We define "abuse" as seeking to obtain profits above the competitive level.

149 Since many financial institutions are members (owners) of both MasterCard and Visa, the charge is often made that these two associations must be directly colluding together. *See* EVANS &

tion rerouting among competing networks, financial institutions are able to threaten to, or to actually alter their portfolios of cards to prevent either network or both from extracting monopoly rents. Even financial institutions that are only a member of either VISA or MasterCard may be able to issue cards under other payment brands, such as American Express.[150] Moreover, because the network is negotiating with the issuers and merchant acquirers, who are themselves in a competitive market for consumer and merchant accounts, the network's ability to shift costs to individual consumers or merchants is limited by the market pressures in the issuer and merchant acquirer markets. In effect, consumers and merchants negotiate with the network through the issuers and merchant acquirers. Those institutions can be thought of as "bundling" individual consumer and merchant bargaining power into larger units, since an issuer or merchant acquirer which negotiates more favorable terms for its customers from the network will gain a competitive advantage over other institutions. Competitive pressures from the individual account market thus are an important limit on the ability of networks to rent-seek at the expense of merchants or cardholders.

C. *Applying the Technology: The Payment Transaction*

A payment transaction in a card-based payment system has two logical components: the authorization and the settlement transactions.[151] In a sim-

---

SCHMALENSEE, *supra* note 1, at 193-06 (debunking this claim). Some countries, such as Canada, address this concern head-on by prohibiting a financial institution from issuing under both the MasterCard and Visa brands. *See* David A. Balto, *Networks and Exclusivity: Antitrust Analysis to Promote Network Competition*, 7 GEO. MASON L. REV. 523, 538 (1999) (describing Canadian ban on duality). We have not found any study that addresses whether prohibiting an institution from issuing multiple brands either limits or promotes competition. If anything, we suspect that such limitations on card issuing, bilateral processing arrangements, or dynamic re-routing of card transactions, increase the transactions costs of switching from one association to another. This increase in switching costs correspondingly increases the monopoly-like rents that both associations may assess, to just below the amount required to change the card portfolio from one association to another.

150 *See American-Express Card Branded Credit Cards*, *supra* note 133, at 38.

151 In a credit-card transaction, those two components occur separately (referred to as a "dual message transaction"); in a debit card transaction those two components occur simultaneously (a "single message transaction."). The difference between single message and dual message approaches to payments creates potential issues for financial institutions. These issues are not relevant for the private dispute resolution mechanism, and, therefore, are ignored here. Also, the electronic path the transaction components take will vary depending on the type of network (open or closed) and the relationships between the particular banks and other parties. General descriptions of the payment card industry are contained in the opinions of the courts in U.S. v. VISA, U.S.A. Inc., 163 F.Supp. 322 (S.D.N.Y. 2001) and Schwartz v. VISA International Corp., 2003 WL 1870370 (Cal. Super. 2005). Although we disagree with the courts' legal reasoning and conclusions in these cases, the basic presentation of the technical details of the card systems is reasonably accurate.

ple closed network transaction, a consumer uses her American Express card at Best Buy. The card is swiped at the Best Buy cash register, the card reader transmits the authorization request[152] to American Express, and American Express checks the consumer's credit and responds accordingly. If the transaction is approved, the consumer signs for her purchase and the sale is complete. Later that day,[153] the merchant uploads all its American Express sales transactions for the day to American Express. American Express posts the purchase to the consumer's account and transfers the funds to the merchant's account. At the end of the billing cycle, the consumer receives and pays her bill from American Express, but the funds (net of American Express's charges) have previously been sent to the merchant by American Express, a short term extension of credit to the consumer by American Express.

In an open system where the merchant acquirer and issuer institutions are different, the transaction is similar, except that the authorization transaction flows from the merchant to its merchant acquirer, then to the appropriate association,[154] which routes the transaction to the issuer, who then responds via the same route. The settlement transaction flows later that day in reverse through this same route.

What appears to be an open network transaction can actually be a closed network transaction when the issuer and merchant acquirer institutions are the same. Thus, if a Citibank MasterCard is used to purchase a cellular phone at a Cingular store and Cingular's merchant acquirer is also Citibank, then the transaction may not be routed through the association but instead be handled internally by Citibank entirely within a Citibank network in the same manner as American Express would have handled the transaction if the American Express card had been used.[155] When a transaction is internally routed, the incentive structures for dispute resolution change, as described below.

---

152 The data sent are the card number, expiration date, amount of transaction, date of transaction, security code, and merchant number. Some transactions also include additional security information such as cardholder addresses (e.g. Internet transactions).

153 There are generally several windows each day in which transactions are posted. Large merchants (e.g. Wal-Mart) may settle transactions multiple times within a day; smaller merchants may settle once a day.

154 In international transactions, there is an additional step in which the information flows between the appropriate national associations. Thus, when a U.S. consumer uses a U.S. issued VISA card in Europe, Visa Europe (a distinct legal and business entity) will pass the information from the European merchant acquirer to Visa U.S.A, who in turn will pass the transaction on to the US issuer. Inter-regional transaction handoffs occur when the region (within a single association) utilizes its own transaction authorization and settlement systems.

155 Of course, there may be differences in internal procedures between banks and other institutions. Our point is merely that the open network will resemble a closed network in these instances. When a transaction in an Open network is internally routed, the decision making incentives associated with dispute resolution more closely resemble the incentive structure of a Closed network. *See* discussion *infra* Section III, Dispute Resolution Systems.

Further, some institutions subcontract data services with third party transaction processors. The rise of third-party entities was an important development in the 1990s.[156] First Data, Total Systems, eFunds, EDS, Atos Origin, Alliance Data Systems, Sema, Arksys, FiServe, Euronet, and other companies contract with financial institutions to handle various aspects of processing card transactions.[157] If the same company has contracted with both the merchant acquirer and issuer, the transaction may not be routed through the association, but stay entirely within the third party transaction processor's network.[158]

These differences can have consequences for the dispute resolution process, since a transaction that is not routed through the association is not necessarily governed by the association's rules;[159] indeed, the association is likely to be completely unaware that the transaction even exists. When a transaction is routed outside the association, the institution(s) involved may opt to provide its (their) own dispute resolution process.[160] This process substitution may be indistinguishable to the merchant and the cardholder

---

156 Brian G. Olsen, *Bank Credit Card Organization, in* AMERICAN BANKERS ASSOCIATION, THE BANK CREDIT CARD BUSINESS 15, 22 (2d ed. 1996).

157 Third party transaction processing companies provide technology services to the financial institution. In effect, the financial institution is outsourcing their financial operations to that technology company. First Data began as a regional association of banks to handle processing. It was later spun off and became a for-profit company. *Id.* at 21.

158 This routing alternative is an issue in that it juxtaposes the legal and technology relationships between associations and financial institutions. Some associations contend that not routing the transaction through the association will violate association rules. It certainly deprives the association of its fee for the transaction. This ability to reroute transactions can give large players with merchant and consumer accounts an advantage. *See* EVANS & SCHMALENSEE, *supra* note 1, at 276 (describing how National Bancard Corporation saw this as a threat in Chicago in competition with First Chicago, which the company argued could undercut its price because it did not have to pay the interchange fee on transactions between merchant customers and its own bank card holders.) Evans & Schmalensee argue that this does not necessarily advantage the First Chicago-type banks, because they still must cover the network costs on their own network. *Id.* at 277. While they are correct, they do not take account of the function of such competition in limiting the interchange fee that the associations can charge. See, also, VISA's announcement that *all* signature debit transactions must be routed through VISA's network. *See, e.g.*, Steve Ruwe, *Required Processing of Visa Transactions through VisaNet*, Visa Business Review, Visa U.S.A., February 2005, Issue No. 050215 (explaining that the Visa U.S.A Board of Directors has directed Visa U.S.A to revise its Operating Regulations to require all Visa transactions be processed through VisaNet).

159 The multi-party nature of the payment system creates disagreements about which party owns, manages, and regulates cardholder data, merchant data, and transactional data. *See* Paul Wenske, *Some Debit Card Users Receive a Little More Protection*, KANSAS CITY STAR, Jan. 6, 2005, at C1, *available at* http://www.kansascity/business/10575372.htm?1c (noting VISA chargeback protection is provided only on VISA's network.).

160 The motivation for routing around the association is the savings of the association's fees and expenses associated with the transaction. In the international context, the process, earnings, and expenses related to converting transacting currency into cardholder currency (also known as foreign exchange) can be a substantial motivation for routing around the association.

from the association process, but it need not be. Regardless, the parties to the transaction are unlikely to know the *corporate jurisdiction* to which they are committed at the time of the transaction.[161] And, the cardholder is unlikely to be aware of the differences *ex ante* because cardholders rarely know who a merchant's merchant acquirer is when making a purchase.[162] Similarly, the merchant's employee conducting the transaction may be aware of the name of the consumer's issuer (which is generally imprinted on the card)[163] but store clerks are unlikely to be aware of the consequences for fees or dispute resolution rules of accepting one card over another. No party to a particular transaction may thus be aware simply by entering into the transaction of either the specific legal rules to which they are agreeing to use in the event of a dispute, or the private jurisdiction that will apply those rules. Despite consumers' ignorance of the system, however, they are not disadvantaged because of the competition among issuers and networks.

ATM network transactions are similar with one key difference. In an ATM transaction, the interchange fees flow in reverse. Fees are paid by the issuer to the ATM owner, instead of by the merchant acquirer to the issuer.[164] As with credit and debit card transactions, if a cardholder uses a machine owned by the issuer, the transaction is probably handled internally; if the cardholder uses a machine owned by another party, then the transaction flows over a network. Depending on the relationships between the ATM owner and the issuer, a transaction may flow through an association or through an alternative network, as described above. There are more alternative networks for ATM machines than for credit and debit card transactions at merchants.[165] Indeed, most debit cards provide debit services

---

161 A similar issue arises with debit cards, where consumers may not be aware of whether they are using an online or offline system. *See, e.g.*, Effross, *supra* note 62, at 362. Some commentators argue that such lack of knowledge is a justification for intervention. *See, e.g.*, Malla Pollack, *Opt-In Government: Using the Internet to Empower Choice—Privacy Application*, 50 CATH. U. L. REV. 653, 670 (2001) ("Market choice activation of the 'invisible hand' requires transparency. Consumers cannot choose x unless they can tell when x is, and is not, part of the offer.").

162 Even if provided with this information, however, consumers may not be aware of the consequences of using particular financial associations with a merchant.

163 Many issuers use multiple brands on their cards, however. Thus AT&T credit cards are currently issued by Citibank. Jane Adler, *Troubles for Cobranded Cards*, CREDIT CARD MGMT., Jan. 2005, at 12, 14, 2005 WLNR 58777. The Citibank name appears only in small print on the back of the card, where it is unlikely to be noticed by the merchant's employee.

164 Fees flow in reverse because the ATM machine is a substitute for a branch of a financial institution. Therefore, when an issuer's cardholder utilizes another financial institution's ATM machine, the Issuer compensates that other financial institution for making their ATM available to their cardholders. Also, there is a cost for holding cash in an ATM machine. Thus, the interchange flows in the opposite direction for ATM transactions, from the Issuer to the Acquirer. Reverse interchange does create some unusual incentives for merchants, with regards to merchant ownership and placement of the ATM within the merchant's location.

165 For example, major ATM networks include NYCE, Star, Pulse, MAC, and MoneyStation.

across several platforms.[166] A bank, for example, may issue a debit card that works in its own ATM network, through a regional ATM network, and through MasterCard or VISA.[167]

Again, the crucial characteristic is the ease with which existing or new institutions can enter the market. Establishing a new connection between a merchant acquirer and an issuer requires only negotiation of an agreement between the institutions, acquisition of telecommunications channels, and modification of the relevant computer programs that route transactions. Although such costs are not zero,[168] for large numbers of transactions the per transaction cost is effectively or close to zero, given the low marginal cost of telecommunications and the ability to amortize the fixed costs over substantial numbers of transactions.[169] This ease of entry allows both actual[170] and potential competition[171] to discipline the networks' behavior to-

166 *See* EVANS & SCHMALENSEE, *supra* note 1, at 297.

167 During the 1980s, ATM networks became more interconnected. EVANS & SCHMALENSEE, *supra* note 1, at 303 ("In the 1980s, ATMs not only increased in number, they became more interconnected."); CHUTKOW, *supra* note 4, at 231-32 (describing expansion of ATM networks). Some transaction processors such as First Data have purchased networks such as Star, which in effect makes First Data capable of internally routing transactions amongst their transaction processing customers without using the VISA, MasterCard or other traditional card-based system networks, a practice sometimes called sub-switching. Richard Mitchell, *The Future of Visa and MasterCard,* CREDIT CARD MGMT., June 2004, at 36, 38, 2004 WLNR 205830 ("First Data sought to leverage its connections with issuers and acquirers to create a closed-loop processing system, and enable clients that send V[ISA] transactions through First Data Net to bypass V[ISA]'s VisaNet authorization, clearing and settlement system and avoid some of the association's fees."). VISA initially saw the ATM networks as a threat. *See* CHUTKOW, *supra* note 4, at 231-32 (describing opposition from VISA to development of member bank ATM networks). The practice is recognized as a serious threat to the associations in the industry. Mitchell, *supra*, at 40 ("[T]he ability of First Data to route transactions internally as 'on-us' between its processing clients that are both issuers and acquirers not only is financially attractive, but theoretically allows the processor to provide customized loyalty programs, special pricing outside of 'inflexible' interchange schemes and other innovations not generally available through traditional networks.").

168 Determining the price structure for a multisided network such as a card-based payment system is a complex problem, whose solution can be costly. However new entrants have the benefit of existing price structures, which can serve as a model for the new entrant. *See* Evans, *supra* note 81, at 363 ("Multi-sided platform markets are also hard to get into because firms must solve quite complex business problems. That complexity may, however, give subsequent entrants an advantage; they can look to the pricing structures and business models adopted by successful incumbents. When American Express entered the charge card business in 1958, for example, it could observe the success of the pricing structure that Diners Club had adopted when it entered in 1950.").

169 *See* EVANS & SCHMALENSEE, *supra* note 1, at 127 (noting declines in transmission and processing costs); Robin Sidel & Joseph T. Hallinan, *MasterCard Swipes Big Debit Account from Visa,* WALL ST. J., Jan. 6, 2005, at C3 (noting competition between MasterCard and Visa in debit market and suggesting Discover is preparing to enter market).

170 EVANS & SCHMALENSEE, *supra* note 1, at 228 ("Entering the payment card industry as an issuer is fairly easy.").

171 *See* EVANS & SCHMALENSEE, *supra* note 1, at 116 ("The prominence of First Data in both the issuing and acquiring business has led some to suggest that it has the potential to become a competing payment card system.").

ward the merchant acquirers and issuers. The competitive markets for merchant and individual accounts in turn prevents rent-seeking by those institutions.[172]

D. *The Private Legal Structure*

The rules governing card-based transactions are largely created through a series of private contracts. The constraints both shape and are shaped by the payment systems' technologies. Merchant acquirers negotiate contracts with merchants to obtain the merchants' sales transactions. The degree to which terms vary depends, of course, on the bargaining power of the parties. Small merchants have less leverage than, for example, Wal-Mart.[173] The existence of alternative networks and the low transactions costs of rerouting transactions, for example, by reprogramming the card readers to route transactions from particular card issuers over particular networks limits merchant acquirers' ability to shift costs onto merchants. Issuers negotiate contracts with consumers, although these negotiations generally center on the interest rate, annual fees, and rewards programs rather than on the remainder of the terms of the contract.[174] Both issuers and merchant acquirers negotiate contracts with the payment mechanism providers. These contracts are generally boilerplate, but side agreements between the payment mechanism providers and particular financial institutions occur, where the financial institution controls a particularly desirable block of accounts.[175] Moreover, their members, who are the issuers and acquirers, control the associations.

The associations also impose rules on their member institutions. These rules are created by the associations' boards and membership. Member institutions get votes on the association boards based on their relative size, with larger institutions which provide the association with a greater

---

172 *See* EVANS & SCHMALENSEE, *supra* note 1, at 171 (noting that "system wars . . . have raged since 1958 when American Express introduced its charge card to challenge Diners Club, then the dominant payment card. These wars have provided long-run benefits to consumers and merchants through lower prices, faster service, and enhanced features."); *id.* at 226-28 (concluding that industry is competitive); *id.* at 239 ("Prices [received by issuers on balances] declined by almost 35 percent between the first quarter of 1984 and the fourth quarter of 1996. The price that the typical consumer actually paid, however, decreased by 7 percent over the same period, and would have decreased significantly more had tax laws not been changed.").

173 James J. Daly, *Visa's Trillion-Dollar Year*, CREDIT CARD MGMT., May 3, 2004, at 38, 2004 WLNR 183657 (Wal-Mart is "widely rumored to have a custom interchange plan with Visa," but the association won't confirm it.).

174 *See, e.g.*, SCOTT BILKER, TALK YOUR WAY OUT OF CREDIT CARD DEBT!: PHONE CALLS TO BANKS THAT SAVED MORE THAN $43,000 IN INTEREST CHARGES AND FEES (2003).

175 *See, e.g.*, Glenn Cheney, *A Credit Card Collision for the Ages*, BANK TECHNOLOGY NEWS, April 1, 1999, at 8 (describing dispute between VISA and Citibank that led Citi to switch its primary allegiance to MasterCard.)

number of transactions and dollar volume of transactions receiving more votes.[176] These rules provide the "law" that governs much of the relationship between financial institutions and both merchants and cardholders.[177] We discuss the characterization of these contractual terms as law below.

The card based payment systems have bylaws and operating rules that cover a wide range of topics at various levels of detail. The bylaws and operating rules may even be in a variety of different documents. For example, some relatively mundane but practical issues are defined, such as the physical appearance of the cards, including the position of the association logo, font, inclusion of security measures such as holograms and security numbers, and a card design approval process.[178] The operating rules also specify how cards should be packaged, transported, and stored.[179] The association bylaws and operation rules also cover standard corporate matters, including the duties and responsibilities of the board of directors and various advisory boards and councils.[180] Some payment associations are not really payment associations in the classic sense because they are owned and operated by an unaffiliated non-bank company.[181] These payment entities typically have councils or advisory boards that have the authority to amend the rules.[182]

In any given dispute between a merchant and a cardholder, the dispute is potentially governed by the contracts between the cardholder and the issuer, between the issuer and the payment mechanism provider, between the merchant and the merchant acquirer, and between the merchant acquirer and the payment mechanism provider plus the rules of the payment mecha-

---

176 *See, e.g.*, Jason Fargo, *Behind Citi's Feud with VISA*, CREDIT CARD MGMT., April 1, 1999, at 28, (describing Citibank's gain in representation by its shift to MasterCard.)

177 *See* David V. Snyder, *Private Lawmaking*, 64 OHIO ST. L. J. 371, 405-06 (2003) (analyzing Visa and Mastercard rules as law).

178 *See generally* VISA GENERAL RULES, VOLUME I, *supra* note 104, §§ 10.1-10.4; MasterCard Bylaws, *supra* note 112, §§ 5.1-5.8; NYCE Network Operating Rules, § 4.6 (January 1, 2005); Star Network Operating Rules, Version 1.4, app. D (October 2004); Quest Operating Rules, Version 1.3, ch. 2 (May 2001); ACCEL/Exchange Network Operating Rules, app. D (March 2003). *See also*, MasterCard International, MasterCard Card Design Standards (2003); Visa Check Card: Design and Branding Standards (2004); Visa PrePaid Card: Design and Branding Standards (2004).

179 *See, e.g.*, American Express, Global Network Services, *Business and Operational Policies*, § 8.3, Shipments from Manufacturer to Issuer (October 31, 2004).

180 *See e.g.*, MasterCard Bylaws, *supra* note 112, art. IV; VISA GENERAL RULES, VOLUME I, *supra* note 104, art. V; NYCE Network Operating Rules, § 2.2 (January 1, 2005); Quest Operating Rules, Version 1.3, App. I (May 2001); and, ACCEL/Exchange Network Operating Rules, § 1.3.2.2 (March 2003).

181 There are many examples of such payment systems. Some examples include: American Express which is owned by American Express; NYCE which is owned by Metavente; Star which is owned by First Data; Exchange which is owned by FiServ; Pulse which is owned by Morgan Stanley; and, Discover which is also owned by Morgan Stanley.

182 *See supra* note 178. Note that the Star ATM/Debit network does not have such an advisory board.

nism provider. In most cases, however, the critical document for dispute resolution will be the payment mechanism provider's rules governing disputes, since the payment mechanism providers will require the issuers and merchant acquirers to incorporate those rules into their contracts with consumers and merchants.

These rules are the law of the relationship. They go beyond a contract because the payment mechanism provider agreement is not only a contractual agreement to existing rules but an agreement to a process which generates future rules.[183] These as-yet-unknown rules are binding on the members of the association. To avoid these rules' application requires exit from the association. The introduction of holograms is an example of the imposition of "new law" by the associations. Both VISA and MasterCard began requiring card issuers to alter their cards to include various security measures such as holograms in the 1990s.[184] All card issuers had to include this security feature, which was not a trivial expense.[185]

Most critically, the rules require most disputes to be handled within the institutions established by the rules themselves. That is, consumers, merchants, merchant acquirers, and issuers must make use of the institutions created by the web of contracts amongst them and, in this case, largely derivative of the payment mechanism providers' contracts with the issuers and merchant acquirers.

E. *The Context*

To briefly summarize, we have shown that card-based payment systems provide a mechanism to convert a diverse set of interactions, many of

---

183 *See* Chris Sagers, *The Evolving Federal Approach to Private Legislation and the Twilight of Government*, (unpublished manuscript, on file with author) ("The more one thinks about it, the more difficult it seems to find any robust and meaningful distinction between 'law' and 'standard' except to say that 'standards' are created by private bodies.").

184 *See* CHUTKOW, *supra* note 4, at 180-82 (describing card security measures, including microprinting on logo and holograms); *see also* CHUTKOW, *supra* note 3, at 188-89 (describing other security features); Homer Brickey, *Credit Card Firms Battling Crooks with Technology*, THE PATRIOT LEDGER, June 24, 1995 (describing security features).

185 Industry players hold pricing and cost information as highly confidential and competitive information. However, it is possible to estimate the magnitude of this expense. In a recently 10-Q filing, one of the leading providers of holograms, American Bank Note Holographics Inc., indicated that a "significant portion of the Company's business is derived from orders placed by certain credit card companies, including MasterCard and manufacturers of VISA brand credit cards, and variations in the timing of such orders can cause significant fluctuations in the Company's sales. Sales to MasterCard were approximately 37% and 45% of sales for the three months ended September 30, 2004 and 2003." Sales for the quarter ending on September 30, 2003 totaled $4.6 million. Sales for the quarter ending on September 30, 2004 totaled $5.8 million. *See* American Bank Note Holographics Inc., Quarterly Report (Form 10-Q) (Nov. 12, 2004), *available at* http://biz.yahoo.com/e/041112/abhh.ob10-q.html (last visited February 24, 2004).

which are single episode transactions, into a series of repeat relationships. Thus individual purchases by consumers from separate merchants are transformed into a repeat relationship between the consumer and her issuer and between the merchant and its merchant acquirer. The issuers and merchant acquirers themselves interact in repeat relationships (or are a single entity). Moreover, these relationships are subject to simultaneous competitive forces in many distinct markets: the consumer relationship;[186] the merchant relationship;[187] the relationships with the financial institutions;[188] the choice of payment,[189] and the terms of that choice. The result of the competitive environment is pressure on the payment mechanism providers, the issuers, and the merchant acquirers to improve the quality and reduce the costs of their products. One important set of improvements produced by this competition is aimed at reducing the frequency of fraudulent use.

By virtue of the web of contracts connecting issuers, consumers, merchants, merchant acquirers and payment mechanism providers, the parties to all these contracts are governed by a set of institutions created in the payment mechanism providers' contracts with issuers and merchant acquirers by reference to the payment mechanism providers' rules. These institutions include rule generation institutions (e.g. "legislatures") and dispute resolution institutions (e.g. "courts"). In the next section we discuss the details of these dispute resolution institutions. This section has shown that several key characteristics are the result of the underlying technology and

---

186 There are many financial institutions that compete for the consumer, her payments, and her borrowing choices. This competition is not limited to financial institutions licensed by VISA or MasterCard, but also includes other card based payment systems such as American Express, Carte Blanche, Diners Club, Discover, JCB, and a host of privately issued payment cards.

187 There is also an intense competition for merchant acceptance. If the merchant does not accept the payment mechanism/mode/choice, no consumer will want that payment vehicle. Likewise, if no consumers have a particular payment mechanism, no merchant will want to accept it.

188 Each association and even many private companies offering payment vehicles that compete for the attention of the financial institution. Some of these associations are jointly owned by financial institutions (i.e., MasterCard, VISA, The Clearing House, etc.) and others are private companies (i.e., NYCE/Metavente, Star/First Data, Pulse/Morgan Stanley, etc.). The Clearing House, a private company jointly owned by many financial institutions, competes with the Federal Reserve to clear checks while simultaneously competing with card based payment mechanisms to extend the reach of checks into new markets. See http://www.theclearinghouse.org/payment_services/000229f.php (last visited May 27, 2005), for new check-based product offerings currently being pitched to financial institutions. Likewise, ATM networks, such as Star and NYCE, compete with VISA and MasterCard over the routing of debit card traffic, while simultaneously competing for the consumer's choice with other payment mechanisms (cash, check, etc.).

189 Competition extends beyond card based payment systems into other payment systems entirely. At the most fundamental level, the consumer may elect to utilize cash. The consumer may also elect to pay with check. The source of funds for that check has exploded over the years to include many nontraditional sources of funds. A check may draw on: equity in and be secured by a consumer's residence; an unsecured personal line of credit that may not even be associated with a card, at all; a brokerage account, including any line of credit established to purchase equities; corporate bonds (i.e., The General Motors' GMAC Demand Notes investment vehicle); and any other source of value.

the competitive environment amongst the players in the card-based payment system industry.

First, competition restrains the behavior of the various parties. Because there is real or potential competition at each stage of the web of relationships that comprise the card-based payment systems, no individual entity can exert monopoly power to force disadvantageous terms on others. In effect, competition serves as a substitute for the combination of checks and balances and due process limitations imposed by constitutions in political systems to restrict the scope of their law-making power.

Second, much of the web of contracts is opaque to individual entities and participants. An entity which is solely an issuer, for example, may not know anything about the terms of merchant acquirer-merchant contracts; consumers are likely to know little about the terms of issuer-association contracts, and so forth. In one sense, it is as if the federalism provisions of the United States Constitution were unknown to the residents of the states, and known only to the state governments. Despite this opacity, however, those involved in contracts are restrained by competition from colluding to extract rents from the parties to whom the contracts are opaque. Continuing our political analogy, (perhaps past the point of reasonableness), it is as if individual U.S. states had the option of switching from membership in the United States to membership in Canada, Mexico, the E.U., or a new combination of states at will (and without a civil war). Under such circumstances, the parties to the contracts have an interest in promoting "federalism" even if state citizens are unaware of the terms of the federal constitution. Competition thus overcomes opacity.

Third, the transformation by card-based payment systems of diverse transactions into a limited set of transactions, facilitates the standardization of dispute resolution procedures through the imposition of ex ante procedures which prevent the creation of disputes (e.g. deterring fraud), create an objective basis for resolving disputes (e.g. creating physical proof of authorization of a charge through signatures on charge slips), and restrict the domain of potential disputes by defining possible grounds for disputing charges in the contracts. The ability to affect potential disputants' behavior by contract *before* a dispute arises is a significant advantage for the resulting dispute resolution mechanism over public legal systems' *ex post* scope for inducing behavior changes.

Fourth, card-based payment systems' transformation of discrete transactions into repeat interactions allows them to harness reputational effects to make the system work. At the most basic level, all participants gain substantial benefits from the use of card-based payment systems. The potential loss of these benefits motivates parties to (in most cases) comply with the association rules. Thus consumers can lose their cards, merchants their accounts, issuers and merchant acquirers their memberships in associations, and associations their members. The threat of being excluded from future mutually beneficial trades thus motivates participants to behave.

III. DISPUTE RESOLUTION SYSTEMS

One of the interesting features of card-based systems is their ability to reduce the universe of possible complaints to a limited, workable, and finite number of dispute types. Cardholders might claim that the charge was not authorized (e.g. "I did not buy a TV set from that merchant"), that the item delivered does not meet the promised quality ("I bought the TV set but it does not work as promised"), or that the item delivered is unsatisfactory for other reasons ("When I got it home, I realized that the TV set is unattractive in my living room and I do not want it."). The card-based payment systems' dispute resolution processes vary their procedures according to the type of dispute. If the question is only whether the charge was authorized, the factual issues are straightforward: was there an authorized use of the card at the merchant in question? The merchant must demonstrate that it has proof that an authorized use took place (e.g. a signed charge slip) and that it complied with the payment mechanism provider's rules in authorizing the charge (e.g. that it sought authorization, verified the signature on the card, etc.). If the merchant cannot prove the use was authorized (and the exact proof will vary depending on whether the transaction was at a bricks and mortar facility, online, or by telephone), the consumer is likely to prevail. If the merchant can prove authorization and rule-compliance, the merchant is likely to prevail. Our other examples of disputes require different approaches to fact gathering. In the following section, we will explore the process itself, various aspects of automation, the legal basis for the system, fees and incentives, and resulting behavior of the players in the payment system.

A. *The Process and Supporting Systems*

Most disputes are initiated by the consumer, once the consumer becomes aware of some difficulty with the transaction. Difficulties generally arise in one of two circumstances: the receipt of the cardholder statement[190] or when the product or service fails to meet the consumer's expectations. On the cardholder statement, a consumer may notice either that there is an unauthorized charge or that the amount of a particular charge is incorrect. Here the consumer is likely to complain directly to the financial institution issuing his card. With a quality issue, at some point after purchase, the product or service becomes defective (at least from the point of view of the

[190] We are using the term "statement" in a very general sense. In today's electronic era, it is important to note that a consumer receives a statement from their financial institution through many different mechanisms, including web based notification, mobile phone notification, paper statements, and on screen display at ATMs, among others. Through any of these mechanisms, the consumer may realize that there is a problem that needs resolution.

consumer). If the consumer fails to secure an acceptable outcome from the offending merchant, the consumer may escalate the process and complain to the financial institution who issued the card. In both circumstances, the consumer will initiate a dispute.

1. Initiating a Dispute

The consumer initiates a dispute by contacting either the merchant or the financial institution that issued his or her card.[191] The consumer initiates the formal process by contacting the financial institution that issued the card. A customer service employee of that financial institution will categorize the complaint with a code to indicate the reason that most closely resembles the substance of the consumer complaint. Based on the code that is selected, different steps in the resolution process and information may be required. Some of these steps or provision of information will be required prior to proceeding on to the next step. For example, if the consumer disputes a charge because the cardholder either does not recognize the transaction or denies that the transaction occurred, the consumer must notify the issuer in writing that she is contesting the charge.[192] There are usually no formal requirements for the format of the written charge contest; as long as the consumer includes the information necessary to identify the charge in question and makes the basis of the dispute clear, the issuer accepts the written contest.[193] Not all disputes require a written letter complaint from the consumer, however.[194]

---

[191] Of course, if contacted, the merchant may issue a full refund or partial refund to the consumer, by submitting a credit transaction through to the payment mechanism, in the same way that the consumer was initially charged. The merchant may also exchange defective merchandise for non-defective merchandise. If, upon contact from the consumer, the merchant corrects the problem to the satisfaction of the consumer, the dispute ends. There are no fees assessed by either the card issuer or merchant acquirer to any other party for settled disputes. This part of the process may appear unrealistic to those who believe banks have an inherent advantage over individual consumers or that the size of a particular party renders any bargain with an individual unfair. However, as described below, the dispute resolution fee structures create significant financial incentives for the merchant to resolve the dispute with the consumer directly, before the consumer complains to her financial institution.

[192] If the consumer does not initially notify the issuer in writing (e.g. she calls the issuer rather than writing the complaint) for certain types of disputes, the issuer notifies the consumer that she must put the dispute in writing. An increasing variety of "writings" are acceptable, including web fill-in forms, email, and the like.

[193] The dispute resolutions by VISA and MasterCard do not address in great detail how the issuer should interact with the consumer for dispute resolution. MasterCard does require a consumer complaint in writing, but, indicates that an unedited email is acceptable. *See* MASTERCARD CHARGEBACK GUIDE, *supra* note 111, 1-31 . For certain chargeback codes, MasterCard specifies that the consumer complaint must make certain affirmations to be acceptable. One type of affirmation that may be required is that the cardholder engaged in the transaction. Another type of affirmation is that the con-

2. Classifying a Dispute

All payment mechanism providers have their own unique dispute reason codes and resolution requirements for each of these codes. MasterCard uses approximately thirty reason codes to characterize the dispute;[195] VISA uses approximately forty.[196] These reason codes are important to all parties involved in any dispute because each code specifies any unique process steps, the required information necessary to resolve the dispute, and, sometimes, the decision criteria for resolution. Failure to follow a particular process step, failure to provide a specific piece of information, or failure to act within prescribed timeframes all may result in permanently losing the dispute on process (as opposed to substantive) grounds.

There are differences between systems.[197] For example, if a consumer

---

sumer first attempted to resolve the dispute with merchant. Not all affirmations are required for every dispute. *See id.* at 3-6.

[194] Some card issuers now accept web based complaints on card transactions. Indeed, the web based card transaction complaint may substantially improve the accuracy of the complaint process. Instead of searching for a transaction from a consumer written letter complaint, the consumer searches and self-identifies the specific problematic transaction. Once the consumer identifies the specific transaction, the consumer next self selects the reason for the dispute from a web based drop-down box, effectively classifying the dispute for the issuer. After selecting the reason, a customized web screen with relevant fields may appear next. For example, a series of fields requesting confirmation that the consumer has already attempted to resolve the dispute with the merchant with a blank date field and contact name at the merchant, would only appear if the consumer had a product quality related dispute. That series of fields would not appear if the consumer denied making the charge completely, because such fields are both unnecessary and irrelevant. After all of the relevant information is collected, it is electronically forwarded through the network to the merchant acquirer and onto the merchant, with no initial human interaction, other than the consumer's completion of the dynamic web based forms. *See, e.g.*, American Express, http://www.americanexpress.com (last visited Feb.1 2005) (note: forms are available only to cardholders who are actually disputing a charge). Consider how such dynamic forms (that vary the information requested, depending on the dispute) prevent a consumer plaintiff from taking two positions that are internally inconsistent or contradictory.

[195] MASTERCARD CHARGEBACK GUIDE, *supra* note 111.

[196] VISA GENERAL RULES, VOLUME II, *supra* note 111.

[197] Depending on business strategy, some financial institutions who issue cards issue only one brand, MasterCard or VISA. Given the many associations and alternative brands in the market, and the control mechanisms for each, there are many business strategies that financial institutions elect to maximize value from these associations. For example, a financial institution may elect to issue only MasterCard cards. Such strategies are important to industry structure but do not impact the dispute resolution function. As such, they are beyond the scope of this analysis. Those financial institutions need not concern themselves with the differences among payment mechanism providers. Indeed, those institutions who issue more than one brand must expend additional effort to understand some of the subtleties underlying various reason codes. Unfortunately for financial institutions who acquire for merchants, competitive pressures in most countries demand that acquirers handle multiple payment mechanisms from multiple payment mechanism providers.

Typically, merchants demand that their financial institution support all payment choices they elect to accept from consumers. Otherwise, that merchant would have multiple accounts at multiple financial

contacts her financial institution complaining that a purchased product no longer functions and the transaction was a VISA transaction,[198] the issuer institution should classify the complaint with Reason Code 56, Defective Merchandise.[199] This reason code has a time limit of 120 days from the transaction and certain process steps and information items are required, as discussed below. If the consumer brought the same complaint, and the transaction was a MasterCard transaction, the issuer should classify the complaint as Reason Code 4853, Defective/Not as Described.[200] As with VISA, this MasterCard reason code has a time limit of 120 days, and certain specific process steps and information items are required, as discussed below.

3. Gathering Information from the Cardholder

Once notified, the issuer conducts a preliminary investigation of the contest. This preliminary investigation revolves around the process steps and information gathering required by the classification of a consumer dispute (into a particular reason code). In our example, if the product does not function as the cardholder expected, MasterCard requires a written statement from the cardholder. That written statement must: (1) indicate that the cardholder did in fact engage in the transaction; (2) indicate that the cardholder contacted the merchant to resolve the dispute and the merchant declined to resolve the problem; (3) indicate that the cardholder returned or

---

institutions, and may have multiple sets of point of sale equipment (for each payment system). This scenario imposes substantial additional overhead onto the merchant. For example, the total incoming revenue would need to be reconciled with multiple bank accounts and reconciliation problems become much more difficult to isolate to one particular payment system. For historical and sometime anti-trust reasons, this scenario exists in some countries.

These multi-payment mechanism provider/brand acquirer financial institutions must understand the requirements of all institutions for which they acquire transactions from merchants. Since most merchants elect to accept cards from multiple payment mechanism providers, to successfully defend themselves these merchants must understand the rules for each payment mechanism. Misclassifying a dispute can be costly to the issuer, both in terms of the fees that are passed back and forth, and more importantly, the issuer may lose the dispute because of the coding error.

198 A payment system will only handle dispute transactions associated with that payment system. Therefore, a transaction that is internally routed where the Issuer financial institution is the same as the Acquirer financial institution (on-us) is not eligible for dispute resolution by Visa, MasterCard, or any other payment system. Likewise, transactions routed under a bi-lateral agreement between two financial institutions are also ineligible for dispute resolution by the brand on the face of the card. *See e.g.*, VISA GENERAL RULES, VOLUME II, *supra* note 111, §1.1 (defines Visa rules as uniformly governing disputes for cards with the Visa brand, but indicates that the rules only apply to transactions made using the brand).

199 *Id.* §1-224, Reason Code 56.

200 MASTERCARD CHARGEBACK GUIDE, *supra* note 111, § 3-113, 3.19, Message Reason Code 4853.

attempted to return the goods; (4) provide a description of the goods or services disputed; (5) explain how the goods did not conform, specifically stating if the goods were of a different quality, quantity, color, or size than expected, or (6) whether the goods were damaged during the shipping process; and, (7) provide any additional documentation that may be necessary to resolve the dispute.[201] VISA's requirements for its equivalent chargeback reason code are substantially similar.[202]

Since all customer service employees of a financial institution are not likely to know or understand the requirements of every rule, the issuer may have employed technology in their customer service system to explicitly script these steps and information requirements for the customer service agent.[203] After the customer service agent classifies this dispute as a Defective Merchandise dispute, a series of questions will appear on the agent's screen for the customer service agent to ask the consumer. These questions will mirror the requirements of that particular reason code.[204]

In this example, after the dispute is classified the issuer is likely to ask the cardholder whether they have contacted the merchant. If the cardholder has not yet contacted the merchant and the dispute is submitted for formal resolution, the issuer and cardholder will lose this dispute.[205] Therefore, the issuer will advise the cardholder to contact the merchant first to resolve the problem and, if and only if the merchant is not willing to resolve the problem, then open a formal chargeback to resolve this problem.

Assuming all of the requirements for this reason code are met using the technology and expertise of the issuer's financial institution, the item will be formally charged back to the merchant; we describe the chargeback process below.

---

201 *Id.* § 3.19.

202 VISA GENERAL RULES, VOLUME II, *supra* note 111, § 1.

203 Interview with William Green, Manager, Chargebacks and Retrievals Processing, Electronic Data Systems, in Westlake, Ohio (Feb. 6, 2004) [hereinafter, Interview, Green] (describing how EDS implemented various work queues and scripts to simultaneously comply with both Visa and MasterCard's rules, to maximize the likelihood that EDS' institutional customers would win the disputes, or at least, minimize the number of lost disputes); *see also* Interview with Robert Sadeckas, Executive, Business Process Management, Electronic Data Systems, in Westlake, Ohio (Feb. 6, 2004) [hereinafter, Interview, Sadeckas] *and* Interview with Loretta Hui, Assistant Vice President, Claims and Adjustments, Citishare Corporation, New York, N.Y. (Feb. 15, 2005) [hereinafter, Interview, Hui].

204 *See* Interview, Green, *supra* note 203; *see also* Interview Sadeckas, *supra* note 203, *see also* Interview, Hui, *supra* note 203.

205 *See, e.g.*, VISA GENERAL RULES, VOLUME II, *supra* note 111, at 417; MASTERCARD CHARGEBACK GUIDE, supra, note 111, §3.19.1, Message Reason Code 4853. *See also,* Interview, Green, *supra* note 203; Interview, Sadeckas, *supra* note 203; Interview, Hui, *supra* note 203.

4. Gathering Information from the Merchant: The Retrieval Request

Certain disputes require a retrieval request. A retrieval request is the process the issuer may utilize to retrieve a copy of the transaction receipt from a merchant or merchant acquirer.[206] When the issuer initiates a retrieval request, the request is passed onto the association, which in turn passes it onto the acquiring financial institution, which then either responds on behalf of the merchant, or passes the request onto the merchant.

Since the receipt[207] ultimately documents the payment obligation, the issuer may initiate a retrieval request sua sponte for its own investigations of fraud, rules violations, or as a result of a cardholder request or dispute. To prevent abuse, the payment mechanism providers have imposed some limits to retrieval requests. For example, VISA has prohibited requests for *original* transaction receipts.[208]

Another limit on retrieval requests is the requirement that the issuer provide certain minimum information elements to secure a retrieval request. Some of these elements include a reference number, account number, transaction date, category code, transaction amount, and merchant location.[209] By requiring that the issuer supply this information for a retrieval request, the payment mechanism provider curtails the sort of open-ended requests that are common in the discovery portion of public dispute resolution. For example, the structure of retrieval system makes it challenging for an issuer to request copies of *all* charge authorizations during a particular period. The issuer would need a great deal of information to make such a request, some of which, it simply does not possess. This limitation on the "discovery" permitted or enabled is an important factor in holding down the costs of the process.

Chargeback rules may require retrieval requests for some disputes because of the nature of the dispute itself. For example, if an incoming transaction does not have enough information to process the transaction, an issuer may only chargeback this item after attempting to reconstruct the missing item, through a retrieval request.[210]

---

206 *See* MasterCard Dictionary, *supra* note 40, at 91.

207 We are using "receipt" very generally here. Historically speaking, merchants used to send their card based payment system receipts into their financial institution for payment, as with pre-Check 21 drafts. Now, the receipt is primarily electronic and both the merchant and cardholder receive copies for their own record keeping. Notwithstanding the improvements in back-office processing that eliminated the need for merchant submission of most paper receipts, the rules still impose requirements on the creation of a receipt. *See, e.g.*, VISA GENERAL RULES, VOLUME I, *supra* note 104, §4.2.J, (requiring that a POS terminal must generate an Electronic Transaction Receipt).

208 *See, e.g.*, VISA GENERAL RULES, VOLUME II, *supra* note 111, §1.2.B.2, Requests for Transaction Receipt Originals.

209 *See, e.g.*, VISA, GENERAL RULES, VOLUME II, *supra* note 111, §1.2.D, Dispute Resolution Rules, Retrieval Requests. *See also*, MASTERCARD CHARGEBACK GUIDE, *supra* note 111, at D-1.

210 *See* MASTERCARD CHARGEBACK GUIDE, *supra* note 111, Chargeback Code 4802.

Other than the processing costs (see below), a retrieval request by itself does not have a monetary impact on any of the parties involved with the payment. At this point, the consumer's account is not credited for the amount of the transaction and the merchant's account is not debited for the transaction amount.

In our example of a consumer complaining about defective merchandise, no retrieval request is required by rules because the consumer is not denying the transaction itself. Indeed, the chargeback rules about defective merchandise require the consumer to affirmatively state that the cardholder participated in the transaction. That affirmative statement eliminates reason codes related to transactions that the consumer denies making altogether.[211]

For retrieval requests, an acquirer may charge an issuer a fee based on the timeliness of the acquirer's response. Rapid responses from the acquirer are rewarded with larger fees. For example, an acquirer may charge an issuer $8 for certain responses within five days of the issuer request and $0 for responses that take more than 21 days.[212] VISA's rule is even more punitive to acquirers because if an acquirer does not meet the retrieval and fulfillment standards, VISA will assess an increased transaction fee for all future fulfillments.[213] Fees also vary depending upon the reason for the request. If the retrieval request was required because the merchant information was incorrect (i.e., the name of the merchant was different than the actual name, the location was incorrect, etc.), the issuer may collect a nominal punitive fee from the acquirer for misinformation.[214]

5. Charging a Transaction Back to the Merchant

Unlike a retrieval request, a chargeback transaction has a monetary impact on the consumer, issuer, association or other payment mechanism provider, acquirer, and merchant. Upon initiating a chargeback to the merchant acquirer, the issuer temporarily removes the charge from the consumer's account, stopping the accrual of interest and fees associated with

---

211 Consider how such affirmations prevent multiple contradictory positions that are permitted in public dispute resolution forums.

212 MASTERCARD CHARGEBACK GUIDE, *supra* note 111, MasterCom fees, Retrieval Requests and Fulfillments, 6-21.

213 *See* VISA GENERAL RULES, VOLUME II, *supra* note 111, §1.2.I, at 1-15.

214 *See* MASTERCARD CHARGEBACK GUIDE, *supra* note 111, Part I, §F, Message Code 7612 Retrieval Handling Fee; For Issuer Use to Penalize an Acquirer for Incorrect Information Verified by the Retrieval Request Document; VISA GENERAL RULES, VOLUME II, *supra* note 111, §1.2.1.2.a at 1-15 ("The Issuer may collect a [nominal] handling fee from the Acquirer . . . for a Retrieval Request resulting from a significantly different Merchant name or incorrect city, state, foreign country, or Transaction Date in the Clearing Record.").

the charge.[215] Likewise, upon receipt of a chargeback, the acquirer financial institution will debit the merchant's account, reduce incoming settlements by the amount or hold the amount in a reserve account. The monetary impact of the chargeback gives the merchant a strong incentive to resolve the dispute quickly. Moreover, the ability of the issuer and merchant acquirer to control the fund eliminates the need to worry about enforcing the decision.[216] In addition, the issuer may collect a nominal handling fee from the acquirer for each chargeback processed. This handling fee is in addition to the amount of the transaction itself. The handling fee is designed to cover the information gathering and complaint codification process. It also shifts some of the handling costs to the merchant acquirer.

6. Representment

Based on its investigation, the merchant acquirer and merchant determine whether or not to accept the chargeback. If it accepts the chargeback, the merchant acquirer removes the amount in question from the merchant's account permanently. If the merchant acquirer rejects the chargeback, it represents the charge to the issuer.

When the merchant submits a representment of a transaction, the merchant must have a reason and the associated documentation required for that reason code, to represent the transaction. The payment mechanism providers have codified these reason codes. Fundamentally, the representment reason codes are grounded in the rules for any given chargeback. If the issuer did not follow the chargeback rules or if the chargeback was substantively improper, the acquirer or the merchant may represent the item to the issuer. In our example of a consumer complaint about defective merchandise, the merchant may represent if the initial chargeback was unsubstantiated. In this case, an unsubstantiated chargeback would not include the required written consumer letter or, if a letter was included, not contain all of the required elements. The acquirer may collect a more significant handling fee from the issuer for each representment processed. This fee is designed to cover the research costs of the merchant acquirer and to shift some of the response cost to the issuer. In this way, the issuer has an incentive to make the complaint as accurate as possible when initially submitted.

---

215 The issuer may still reduce the cardholder's credit line for the amount, because until the temporary credit is permanent, the charge may be reinstated.

216 *See* Perritt, *supra* note 42, at 676 ("Many forms of ADR involve a readily available fund (usually the payment for the disputed transaction) as a way of satisfying a decision for either disputant. The availability of the fund often is underestimated as a consideration. This consideration may explain why intermediary-provided dispute resolution, such as credit card chargebacks and escrow arrangements, prove more attractive in practice than independent third-party mechanisms such as arbitration or mediation. The successful party to an arbitration must still be concerned about the enforceability of an arbitration award against a reluctant loser.").

7. Acceptance or Rejection of Representment and Further Chargeback Rights

The issuer then decides whether or not to accept the representment. If it accepts the representment, the issuer releases the funds back to the acquirer bank. If rejected, this stage of the process ends. The issuer may then repost the amount in question to the consumer's account or decide to pay the charge itself, depending on the facts of the case and the value of the consumer's relationship with the issuer. The issuer may also collect a handling fee in addition to any fees previously assessed to the issuer by the acquirer if the issuer elects to chargeback and escalate to arbitration, as discussed below. In this way, the acquirer has an incentive not to represent items that are likely to be escalated to arbitration.

8. Association Arbitration and Mediation

If the issuer does not accept the representment, the issuer may appeal to the association. This is the first point in the process involving a neutral decision maker. Analysis of the substance of the dispute begins and ends with the underlying merchant-consumer transaction. When a consumer purchases goods or services from a merchant, using a card, that transaction is governed by any explicit or implicit contract between the consumer and merchant, including the implicit or explicit warranty and the warranty's limitations.

If a dispute is appealed, the association investigates and makes a determination. The investigation is typically limited to the materials provided by each member to the other member during the earlier stages of the dispute resolution process. Indeed, MasterCard's rules explicitly state that it will discard any materials submitted outside of the normal retrieval, chargeback, representment, and arbitration processes.[217]

In the arbitration process, the associations are not limited to awarding the amount in contest; the association may fine either or both parties for any errors it uncovers in the course of the investigation and may investigate rule violations related to the transaction regardless of their connection with the merits of the dispute.[218]

The party which loses the appeal to the association can then appeal once more within the association based on claimed errors in interpreting the rules. The loser can also make an equity-based claim (i.e., "It isn't fair that

---

217 *See* MASTERCARD CHARGEBACK GUIDE, *supra* note 111, §4.4, at 4-5.

218 In any circumstance where a member financial institution believes that another member financial institution is out of compliance regarding a particular transaction or set of transactions, that member institution may bring a compliance case against the other member even without a dispute.

I lost") against the other financial institution outside the formal process. The winner has complete discretion over whether to grant such a claim.[219]

In the association arbitration processes, the appeals follow the English rule, with the loser paying the costs of the appeal process. The associations typically charge a significant filing fee[220] and an even more significant administrative fee to the member found responsible for the case.[221] In addition, the association may assess fees for technical violations during the dispute resolution process regardless of whether the member won or lost the dispute. Such technical fee violations include persisting with an invalid chargeback; submitting an invalid representment; submitting invalid documentation; and processing a chargeback beyond the permitted time limits.[222]

B. *Incentives*

An issuer bears the primary costs of consumer behavior, at least for those consumers that are that financial institution's customers. Therefore, the issuer will question the consumer early in the process to ensure that the dispute is legitimate and that the required documentation is available prior to submitting a dispute. If the dispute appears legitimate, and the consumer is a profitable consumer, to avoid the fees the issuer may simply remove the charge from the consumer's statement, without actually charging back the amount in question.

With the narrow exception of civil rights laws,[223] there is no legal requirement that an issuer serve all people without regard to cost or profitability of that consumer. Therefore, an issuer has several options with consumers who are costly to service. The issuer may impose higher fees, including higher annual fees, late fees, and higher interest rates to expensive consumers. The issuer may elect not to renew a consumer account or the issuer may terminate an existing account.[224] Given the financial incentives de-

---

219 Statistics for such scenarios are not available. However, evidence exists that winners occasionally grant internal reviews of such disputes. *See* Interview, Hui, *supra* note 203. The mere fact that such a review possibility exists and is sometimes granted without any force whatsoever, is evidence of a process where the potential for and the value of repeat interactions is quite significant. In Hui's view, "it is important to be standing on principle in any decision, and to be sure that the other [financial institution] believes it, even if the result is unfavorable to that party." *Id.*

220 *See* MASTERCARD CHARGEBACK GUIDE, *supra* note 111, §4.5, Fees and Assessments, at 4-6.

221 *Id.*

222 *See, e.g.*, VISA GENERAL RULES, VOLUME II, *supra* note 111,§ 2, at 3.G.4.c ("Visa, USA assesses the following fees to the responsible Member . . . [significant] penalty fee for each technical violation of the *VISA U.S.A. Inc. Operating Regulations.*").

223 The Equal Credit Opportunity Act and Regulation B are two examples of such laws. *See generally*, Equal Credit Opportunity Act, 15 U.S.C.A § 1691 (2005); *see generally*, 12 C.F.R. §§202.1 - 202.17 (2005).

224 Although cardholder initiated disputes are currently not reported to the credit bureaus, the fact that an account was closed by the issuer is reported and recorded by the credit bureaus. Other unrelated

scribed above, as a consumer submits disputes for resolution, that consumer becomes less profitable.

Likewise, merchant acquirers bear the costs associated with merchants with chargebacks. If an acquirer receives a chargeback, and the merchant is a highly profitable merchant, the acquirer may simply accept the chargeback and not deduct the amount from the merchant's account. The acquirer may take this action as an act of goodwill.

Merchants with more disputes are more expensive to service and high dispute rates also indicate that there may be other problems with that merchant. An acquirer financial institution may elect to substantially increase the discount rate for its merchant customers who receive too many disputes. Merchants with particularly high rates of chargebacks may be eliminated from the payment system altogether. Merchants who commit fraud[225] can be permanently banished by the acquirer financial institution with either or both VISA and MasterCard through the maintenance of a terminated merchant file. Both associations require all merchant acquirers to check the terminated merchant file prior to opening a merchant account.[226] This file acts as a substantial deterrent for those merchants and their owners who elect to deliberately defraud the payment system.

IV. COMPETITION, REGULATION, & THE EVOLUTION OF THE SYSTEMS

As we discussed above, we contend that two of the crucial elements that make the card-based payment systems' dispute resolution systems successful are their ability to transform single-instance transactions (e.g. potential disputes) into repeat-play transactions, through the insertion of the association, issuer and merchant acquirer into the consumer-merchant trans-

---

lenders are likely to ask a borrower why the issuer closed their account. This information sharing creates further incentives to prevent abuse of the chargeback system.

225 There are numerous merchant frauds that are likely to result in global permanent banishment from a particular payment system. One such fraud is the acceptance and submission of transactions from known stolen cards, and splitting of the proceeds with the criminals who pilfered the card. This fraud works (in the short term) because not all stolen cards are known at the time of the card theft and not all consumers complain about small transactions. Therefore, merchants who are caught colluding with criminals to submit false transactions to collect the proceeds, may be permanently banished from accepting cards. Another such fraud is the deliberate acceptance of cards and non-shipment of goods. Merchants who commit this kind of fraud withdraw the incoming deposits and disappear. Eventually, cardholders may charge these purchases back to the merchant, but, there will be no merchant left to accept the chargebacks. Therefore, the merchant acquirer is responsible for this fraud and will likely permanently banish the owners of that merchant from future participation in the payment system. Often, these frauds initially appear in the chargebacks process.

226 MasterCard Bylaws, *supra* note 112, § 9.5.2.4 at 9-6 (requiring acquirers to report merchants to association); VISA GENERAL RULES, VOLUME I, *supra* note 104, §4, at 2.D.1.b ("An Acquirer must: query the terminated merchant file to determine if the prospective Merchant has been terminated for cause . . . .").

action, thereby harnessing reputational incentives and imposing *ex ante* constraints on potential disputants' behavior through requiring structured behavior by participants to the transactions. Neither feature was an intentionally designed characteristic of the payment systems. Both characteristics evolved in response to market pressures during the multiple attempts at constructing card businesses by banks and other financial institutions in the 1950s.

In addition to the market pressures, card-based payment systems have been affected by regulatory pressures. While we believe, as we argue below, that these regulatory pressures have not been determinative of the success of the card-based payment systems' dispute resolution systems, we briefly describe the major regulatory events and analyze their influence as well.

A. *The Creation of Card-Based Payment Systems*

Hotels, oil companies and department stores all began to issue cards to their customers before World War I, but these systems were limited to specific merchants rather than general purpose systems.[227] The first major step in the creation of the modern card-based payment systems was the beginning of networks in 1948 when a group of New York City department stores banded together to make their cards interchangeable across merchants.[228] Card-based payment systems did not become widespread until the development of the Diners Club (1950) and American Express (1958) charge cards in the 1950s.[229] Other companies also attempted to create general purpose cards during the 1950s but all but these two failed.[230] Diners Club initially targeted wealthy residents of Manhattan for use at restaurants.[231] The companies' value proposition to prospective customers was

---

227 EVANS & SCHMALENSEE, *supra* note 1, at 61.

228 EVANS & SCHMALENSEE, *supra* note 1, at 62.

229 *See* GROSSMAN, *supra* note 65, at 262-263 (describing start of Diners Club); *id.* at 280-285 (describing start of American Express card). Although American Express did not produce a card until 1958, the company considered the idea as early as July 1946 but rejected it because of fear it would compete with the travelers' check market. *Id.* at 264-266. Later the company rejected moving into the market because it saw Diners Club and other early card companies as "shlock" operations. *Id.* at 265-266. When it finally decided to enter the market, American Express initially considered buying Diners Club. *Id.* at 274.

230 EVANS & SCHMALENSEE, *supra* note 1, at 63.

231 *See* GROSSMAN, *supra* note 65, at 262 (Diners Club initially intended as "a universal restaurant card that would be accepted at all major New York restaurants."); EVANS & SCHMALENSEE, *supra* note 1, at 84 (describing early industry as targeting "selected Manhattan gourmets"); *id.* at 85 ("Credit cards have led the way in taking a product that was originally targeted to well-off restaurant goers in Manhattan and making it available to the masses.").

that the cards freed the customer from carrying cash,[232] provided thirty day payment terms,[233] and gave the prestige of membership in an elite group of card-holders.[234] For merchants (Diners Club targeted primarily restaurants at first but soon expanded to "florists, gourmet shops, motel chains," and Hertz rental cars),[235] the companies would automatically deposit the charged funds in the merchants' accounts, freeing them from handling incoming cash, and also handle collections from card-holders.[236] In particular, Diners Club and American Express assumed the risk that cardholders would not pay their monthly bills. If a cardholder did not pay, the restaurant retained the payment from the card issuer.

By 1957, Diners Club had almost 500,000 cardholders and charge volume of $7 million per month.[237] When American Express entered the market in 1957, it purchased the American Hotel Association and Gourmet magazine's charge cards, giving it 190,000 cardholders before it even began operations.[238] Applications brought in another 60,000 cardholders by the first day.[239] The company had more than 17,500 establishments committed to accept the card by the first day as well, helped by the company's prestige.[240] The card grew quickly, reaching 900,000 cardholders and 82,000 merchants in 1962.[241] Both cards were sufficiently valued by consumers that the companies were able to charge an annual membership fee.[242] They were also valued by businesses, allowing the card companies to charge a "discount fee" of a percentage of the transaction.[243] Diners Club initially

---

232 *See* EVANS & SCHMALENSEE, *supra* note 1, at 94 ("By reducing the need for cash balances, payment cards provide a potentially enormous benefit to consumers."); GROSSMAN, *supra* note 65, at 262 (Diners Club was founded when Frank McNamara "had just finished a meal at a restaurant when he realized to his dismay that he could not pay the check."); EVANS & SCHMALENSEE, *supra* note 1, at 212 ("Simply put, payment cards are much easier to carry around and use than are most of the other payment methods.").

233 *See* GROSSMAN, *supra* note 65, at 262 ("In the card, club members had blank-check, interest-free charge privileges and a notable convenience.").

234 *Id.* at 262.

235 *Id.* at 263.

236 *Id.* at 262.

237 *Id.* at 274.

238 *Id.* at 283.

239 GROSSMAN, *supra* note 65, at 284.

240 *Id.* at 285.

241 FRIEDMAN & MEEHAN, *supra* note 95, at 59.

242 GROSSMAN, *supra* note 65, at 263.

243 "Transaction" is a term of art in the payment system industry. It refers to the individual exchange that occurs when a consumer utilizes a payment mechanism to transfer value to a merchant (e.g. when a consumer swipes their credit or debit card through a merchant's terminal). We use the term in this paper as it is used in the industry rather than to mean a deal (e.g. acquisition of one company by another), as lawyers and investment bankers often do. *See also,* MASTERCARD DICTIONARY 104-106 (for thirteen different definitions of various types of transactions or items related to the core concept of a transaction).

charged a 7 percent merchant discount; more recently typical merchant discount rates ranged from 1.6 percent for Discover to 2.75 percent for American Express.[244] (Diners Club is now co-branded MasterCard,[245] and therefore, the discount rate has fallen to the MasterCard levels.)

In effect, the early payment cards were simply the extension of the type of credit provision made possible in industries selling physical inventory by factoring to the consumer-restaurant transaction. As we described earlier, in a factoring transaction a financial intermediary provides a business with credit secured by the business's inventory, while that inventory is being sold to customers.[246] Factoring required, however, a relatively constant level of inventory of durable assets to secure the credit and costly physical monitoring[247] to ensure the merchant kept the required level of inventory. Providing credit on this model to a restaurant would not, of course, be possible because the restaurant's "inventory" is the accounts receivable from individual diner's consumption of food. Without physical, durable assets, restaurants lacked the collateral to obtain loans based on value.

Diners Club changed that. When Diners Club handled a consumer purchase of a meal in a restaurant, Diners Club paid the restaurant the bill for the charged meal (less the discount) before Diners Club received its payment from the consumer. When Diners Club issued the card to the consumer, it selected only people it thought most likely to be able to ultimately pay the bill, although its initial credit screening was quite crude.[248] Diners Club then essentially offered the restaurant financing for those portions of the receivables charged to the card, monetizing the cardholders' promises to pay in the future for the meal they had already eaten. Thus not only were consumers able to buy meals on credit, but the restaurant was able to finance its receivables.

As this brief account makes clear, the credit card business was not created to harness reputation or to structure transactions to reduce the frequency of disputes. It was created, like most businesses, in the hopes of making money by offering a service. Nonetheless, the nature of the busi-

---

244 *See* EVANS & SCHMALENSEE, *supra* note 1, at 129.

245 Diners Club and MasterCard announced a relationship where Diners Club cards would become co-branded MasterCard. *See* http://www.dinersclubnewsroom.com/view_release.cfm?id=183 (April 29, 2005) and http://www.dinersclubnewsroom.com/view_release.cfm?id=199 (September 20, 2005). In effect, the Diners Club card is effectively a MasterCard. When the card is presented to a merchant, that card is treated as a MasterCard, with the discount rates applicable to MasterCards.

246 *See* note 73 *supra.*

247 There are also problems associated with merchants who secure more than one loan on the same inventory, particularly because of historic information gaps associated with collateral. These information gaps disappear with financing based on a specific consumer transaction.

248 GROSSMAN, *supra* note 65, at 263 (noting that initial list of prospects was simply a mailing list of 5,000 sales managers).

ness produced problems that led the industry to develop measures which introduced reputation and structure.

The new charge cards quickly developed problems. Abuse and fraud were major risks, large enough that many early observers did not believe Diners Club could survive.[249] Because of the lack of real-time connections between the merchants and Diners Club and American Express, there was no way to immediately stop a card's use if a cardholder began to abuse it.[250] American Express lost $4 million on the charge card operation in its first two years of operation, largely because of a lack of credit screening of prospective cardholders.[251] Indeed, by 1960 losses had reached $10 million and senior executives were not convinced that the card would survive.[252]

At the same time as Diners Club and American Express were creating their closed networks, banks were also attempting to enter the credit card market. In 1951 Franklin Bank in New York expanded the market by creating an applicant screening process that allowed it to issue revolving cards outside the narrow demographic relied on by American Express and Diners Club.[253] Franklin began by simply sending cards to prospective cardholders without credit screening.[254] Eventually, Franklin developed a profile of

---

249 *Id.*

250 *Id.* ("[M]any people doubted that the company would survive [due to credit risk] . . . A large number of card abuses would bankrupt the Diners' Club.")

251 *Id.* at 286 ("The company had done a poor job in evaluating credit risks, which was understandable since it had no experience with handing out blank-check credit, or any credit for that matter.") Screening card holders remained an issue for issuers into the 1970s. As Chutkow's history of VISA notes, "many banks issued cards indiscriminately, and not only to their own customers. Some banks bought mailing lists and issued cards to everyone on them, without any credit analysis or screening. Mass mailings led to massive thefts, often directly from mailboxes, and that led to massive fraud." CHUTKOW, *supra* note 4, at 154.

252 GROSSMAN, *supra* note 65, at 299. Starting a new card operation remains expensive. Dean Witter "incurred substantial initial losses as it spent money prospecting for customers and building merchant acceptance" but ultimately became profitable. EVANS & SCHMALENSEE, *supra* note 1, at 281. To cope with the losses, American Express raised its fees from $6 to $10 a year and increased the merchant discount rate. The losses stopped, and by 1967 card volume reached $1.1 billion, the number of cardholders reached 2 million, and the card earned American Express $6.5 million in profits. GROSSMAN, *supra* note 65, at 303. This problem continued to plague card issuers, with Citibank losing more than $500 million over three years after it introduced a national marketing campaign for its credit cards and American Express experiencing twice the industry average uncollectible debts with its Optima card. *See* FRIEDMAN & MEEHAN, *supra* note 95, at 65; *see* Rob Wells, *American Express Chief Brings Stability to Card Giant,* SEATTLE TIMES, Oct. 3, 1993, at D7 [1993 WLNR 1149106].

253 CHUTKOW, *supra* note 4, at 59-60.

254 AURIEMMA, *supra* note 5, at 4. Banks issuing early credit cards had several problems. First, in many cases they issued cards without prescreening the cardholders for creditworthiness. AURIEMMA, *supra* note 5, at 9 ("The most aggressive banks sent cards to deposit customers, loan customers, safe deposit customers, and any other customers whose addresses they could obtain. Many banks bought mailing lists consisting of names from magazine subscriptions, driver's license registrations, and the like. Not surprisingly, some of the individuals who were issued cards did not manage the credit well."). Second, even when they did screen, banks sometimes used inappropriate criteria. *Id.* ("Some banks

customers it believed likely to pay and vastly expanded the number of people with access to a card-based payment system.[255] Franklin's card, however, was useful only in the New York metropolitan area.[256]

Banks across the country began to issue proprietary cards.[257] By 1953 there were nearly 100 banks issuing proprietary credit cards in Manhattan alone. [258] While the markets for these cards were geographically limited (e.g. a New York bank's card was accepted only by merchants in the New York area, and a San Francisco bank's card was accepted only in the San Francisco area) there was intense competition among banks within each geographic region.[259]

The proliferation of cards led to complications on the merchant side, however, since each bank had to individually create a relationship with each merchant to allow the merchant to accept its card. Merchants were reluctant to accept additional cards unless there was a substantial base of consumers with that bank's proprietary card.[260] To get the necessary cardholder base, banks offered successively more generous terms and features to cardholders.

The transactions costs of the proliferation of cards for merchants were substantial, since each proprietary card required a separate contract, bank account, and processing with the issuing bank. Within a few years many of the banks exited the market because they lacked sufficient cardholder and/or merchant bases to profitably operate proprietary card systems.[261] By 1960, the boom was over but a competitive market in proprietary cards with

---

were inexperienced with extending unsecured credit—especially revolving credit associated with a credit card. The credit approval criteria that had served banks well when making installment loans, such as automobile or home loans, proved to be inadequate for extending credit through cards.") Similar problems arose when banks began converting ATM cards into debit cards. *See* BALTO, *supra* note 63, at 1102 ("A large part of the [fraud] problem arises when offline debit cards are issued in an unsolicited manner. Banks often reissue their online ATM cards as a VISA check card which can be used as either an ATM or an offline debit card.").

255 CHUTKOW, *supra* note 4, at 60 (describing credit screening).

256 Steve Rhode, The History of Credit and Debt, http://www.myvesta.org/history/history_creditcard.htm (last visited May 31, 2005).

257 EVANS & SCHMALENSEE, *supra* note 1, at 63.

258 CHUTKOW, *supra* note 4, at 60 ("By 1955, about 100 banks were operating card programs. . . Most of these bank cards were usable only in a small local area, and few generated enough transaction volume to be profitable."); AURIEMMA, *supra* note 5, at 4 ("Franklin National Bank's credit card program was copied by hundreds of other banks in the late 1950s and early 1960s.").

259 *See* EVANS & SCHMALENSEE, *supra* note 1, at 63-64 (describing competitive environment that saw banks losing large amounts of money competing in the new business).

260 *See* Evans & Schmalensee, *supra* note 86, at 887 ("payment cards were useful to consumers only if they were accepted by many merchants, and they were useful to merchants only if they were carried and used by many consumers").

261 EVANS & SCHMALENSEE, *supra* note 1, at 63-64 ("By 1962, many bank payment card plans had fled the field, including Chase Manhattan's.").

local merchant bases existed throughout the country in most major metropolitan areas.[262]

The proprietary era thus had two critical weaknesses. First, the transactions costs of merchant-bank relationships were high because of the need for individual contracts with each issuer. The individual nature of the merchant-bank relationships also reduced the scope for banks to impose structure on the credit card transactions, as merchants would not be willing to follow different procedures for each card they accepted. Second, the issuer-consumer relationship lacked sufficient reputational guarantees. Banks had selected cardholders based on little more than the magazines the cardholders subscribed to; unsurprisingly, the cardholders had "defected" from the deal by not paying their bills in a timely way.

B. *The Rise of Associations*

In the mid-1960s two crucial innovations developed that took opposite approaches to solving the problems created by the regional-based proprietary cards. In California, Bank of America had the most extensive set of relationships with merchants, allowing its cardholders to use their Bank of America credit card at the widest range of merchants in California.[263] Realizing that this set of relationships with merchants had value to other banks as well, and that it had one of the few profitable bank card operations, in 1966 Bank of America decided to license access to its merchant portfolio to banks from other states, creating "BankAmericard." [264] Within two years, forty-one banks were issuing BankAmericards and another 1,823 banks were signing up merchants or issuing cards as licensees of the forty-one.[265] Bank of America limited the fraud problem by giving each card a credit limit ($300 for "simple" cards and $500 for preferred customers) and requiring merchants to call in for authorization for transactions over that amount.[266]

Banks who licensed the merchant portfolio from Bank of America physically added the BankAmericard logo to their credit cards, in addition to maintaining their own name on the card.[267] Thus, for example, a consumer in Nevada would have a First National Bank of Reno card which was also a BankAmericard. The franchise model had serious problems, with franchisees losing hundreds of millions of dollars in the first years.[268] Later

---

262 EVANS & SCHMALENSEE, *supra* note 1, at 64.

263 EVANS & SCHMALENSEE, *supra* note 1, at 64.

264 CHUTKOW, *supra* note 4, at 69-70 (describing how Bank of America decided to license the use of the system to other banks).

265 *Id.* at 70.

266 *Id.* at 64.

267 EVANS & SCHMALENSEE, *supra* note 1, at 65.

268 CHUTKOW, *supra* note 4, at 92.

Bank of America spun off the card network as National BankAmericard Inc. ("NBI") in 1970[269] (which became VISA in 1976)[270] because the franchise model was unworkable.[271] The following year, the newly independent network had $3.7 billion in charges, up from $2.7 billion the year before.[272]

Bank of America charged the licensing banks a variety of fees. Not only did it charge licensees for the use of its trademark, it introduced a concept it called "Interchange" which was a fee for each transaction at a Bank of America merchant.[273] The development of the interchange was critical to making the open network function.[274] Merchant acquirers cannot predict in advance which issuer's cards will be used at their merchants, requiring them to either negotiate *ex ante* a price for the obligation generated by the underlying transaction with every issuer in the network or to negotiate *ex post* once the issuer was identified. Of course, in an *ex post* negotiation the merchant acquirer would be in a difficult bargaining position since the only buyer available would be the cardholder's issuer.[275] Alternatively, the merchant or merchant's bank could negotiate bilateral agreements with every other bank in the system. If, for the sake of argument, there are 15,000 participating financial institutions, and the merchant desired 100% coverage, that would translate to numerous individually negotiated contracts.[276] By setting a system-wide pricing and negotiating structure for the obligations, the networks solve this coordination problem inexpensively, both in terms of the upfront setup and the ongoing maintenance.

At about the same time, banks in other areas began to form associations that truly shared the merchant portfolios.[277] For example in 1965, the Midwest Bank Card Association was formed by four Chicago banks,[278] in 1966 Interbank was formed by fourteen eastern banks,[279] and the Western States Bank Card Association ("WSBCA") was formed by four big Califor-

---

269 *Id.* at 92-109 (describing spin-off).

270 *Id.* at 159-160 (describing name change).

271 EVANS & SCHMALENSEE, *supra* note 1, at 157.

272 CHUTKOW, *supra* note 4, at 117.

273 The interchange fee compensates the issuer bank "for the 'free' period between settlement, or payment, to the acquirer (the merchant bank) for cardholder purchases and billing to cardholders." Olsen, *supra* note 156, at 17.

274 Evans, *supra* note 81, at 375-76 ("Members of cooperative systems such as MasterCard and VISA compete for cardholders and merchants. Absent coordination there is no way for these members to determine pricing structure and thereby internalize the indirect network externalities created by merchants for cardholders and vice versa. A centrally set interchange fee enables the cooperatives to establish a pricing structure.") (citations omitted).

275 *See* Evans & Schmalensee, *supra* note 86, at 889-90.

276 To be specific, under the assumption of 15,000 financial institutions and a target of 100% coverage, there would be 15000! (factorial) individual contracts.

277 EVANS & SCHMALENSEE, *supra* note 1, at 65-66.

278 CHUTKOW, *supra* note 4, at 70.

279 *Id.* at 70.

nia banks. By the end of the 1960s the WSBCA became MasterCharge.[280] In the 1970s, the regional associations began to merge into larger, more national and even international associations.[281]

The association approach differed from Bank of America's approach in that, rather than licensing an asset to another bank, members in the association contributed assets (e.g. access to their merchant and consumer portfolios) to the association, receiving back reciprocal access to the other members' merchant portfolios. There were also similarities. The associations implemented fee structures for transactions out of a bank's card portfolio in another member bank's merchant portfolio that paralleled Bank of America's charges for the use of Bank of America's merchant portfolio.

One major change in the association structure came in 1976 when antitrust regulators forced a change in both associations' rules that prohibited members from participating in both.[282] With those rules withdrawn, many banks joined both associations.[283] Today VISA and MasterCard are both owned by many of the same banks, although the percentage share each bank has in the two networks may vary. Despite this overlap in ownership, the networks are extremely competitive.[284] Today the competition is largely on the card issuer side, with a few banks and non-banks dominating the merchant acquirer business.[285] In the 1990s, non-banks began to enter the credit card issuer market. Companies like AT&T, General Motors and others created co-branded cards through partners and financial subsidiaries.[286]

Although the major event of the 1970s was the rise of the two national associations, MasterCard and VISA, proprietary networks continued to thrive as well. Not only did Diners Club and American Express expand out of their original market niches among the Manhattan elites,[287] other proprietary networks (e.g. gasoline and department store cards) also grew.[288] In-

---

280 *Id.* According to Robert D. Manning, the Western States Bank Card Association, the Interbank Card Association, and the Midwest Bank Card Association were all specifically formed as a response to Bank of America's licensing of BankAmericard. And, later, Western States Bank Card Association became MasterCharge. *See* ROBERT D. MANNING, CREDIT CARD NATION: THE CONSEQUENCES OF AMERICA'S ADDICTION TO CREDIT 84 (2000).

281 WSBCA first licensed its know-how to Europay in Europe, ultimately merging with it in the 1970s. *See* CHUTKOW, *supra* note 4, at 224-27 (on rise of Visa International).

282 EVANS & SCHMALENSEE, *supra* note 1, at 69-70 ("Faced with an ambivalent Antitrust Division [of the Department of Justice] and the possibility of expensive litigation, Visa removed all restrictions on dual membership in mid-1976.").

283 Olsen, *supra* note 156, at 19 ("Duality has eliminated most of the profits from the merchant business" due to competition.).

284 EVANS & SCHMALENSEE, *supra* note 1, at 70-71 ("Considerable competition" between networks explained).

285 Olsen, *supra* note 156, at 19.

286 Cynthia R. Whiteman, *Marketing, in* AMERICAN BANKERS ASSOCIATION, THE BANK CREDIT CARD BUSINESS 38, 39 (2d ed. 1996).

287 EVANS & SCHMALENSEE, *supra* note 1, at 85.

288 *Id.* at 85-94.

deed, new closed networks have regularly appeared since then, the largest and most widely recognized of which was the creation of Discover by Sears, Roebuck & Co. in 1985.[289]

By the 1970s, the proprietary bank card world of the 1950s and 1960s was transformed. VISA and MasterCard's innovations in both business model and technology had reduced transactions costs to a fraction of Diners Club's and American Express's initial charges, association networks linked merchants and consumers, association rules governed virtually every aspect of transactions, and advances in credit reporting began to make it possible for issuers to track the reputation of their customers. Issuers and merchant acquirers competed for business by cutting costs and innovating to improve the quality of their products.

C. *The Modern Era*

Since the 1970s, the evolution of card-based payment systems has increased in pace. Important security features, which reduce fraud, and features which reduce errors and speed processing appeared during the 1970s,[290] including the introduction of the magnetic strip on the back of the card in the 1970s,[291] which facilitates swiping the card through an electronic mechanism, reducing errors in transactions and increasing security. New anti-fraud technology continued to appear, such as the addition of the CVV and CVV2 numbers in the 1990s.[292] Holograms on cards were introduced in 1983, making it harder to counterfeit.[293] Tamper resistant signature panels on the back of cards were introduced in 1989.[294]

Perhaps the two most important technological developments were the appearance of PIN-based debit cards in the 1980s[295] and the shift to electronic processing in the mid-1970s.[296] As noted earlier, debit cards operate on a single message system, combining the authorization and settlement

---

289 *Id.* at 281.

290 *See* CHUTKOW, *supra* note 4, at 153-54 (describing development of computer systems to reduce fraud losses).

291 *See* Jim Collins, *Hidden Identity,* ATTACHÉ (June 2004), available at http://www.attachemag.com/archives/06-04/informed/infos1.htm (describing history of magnetic strip).

292 These numbers are created by proprietary formulas by issuers; the association rules in open associations specify how complex the formulas must be.

293 CHUTKOW, *supra* note 4, at 183.

294 MasterCard, History of Firsts, http://www.mastercardinternational.com/corporate/history_firsts.html (last visited April 23, 2005).

295 EVANS & SCHMALENSEE, *supra* note 1, at 297-302 (describing growth of the debit card).

296 VISA introduced electronic processing in 1973 and American Express and MasterCard soon followed. CHUTKOW, *supra* note 4, at 158-159. *See also* EVANS & SCHMALENSEE, *supra* note 1, at 176 ("[P]erhaps the most important series of innovations that have taken place over time have involved improvements in processing transactions.").

messages into a single transaction.[297] Electronic processing made card based systems less like check processing and allowed innovations such as the introduction of real time authorization and rule-based systems for detecting potential fraud.[298]

One key aspect of many of the innovations in the industry is that they reduce fraud. The burden of fraud had been contractually shifted to the merchants by the merchant acquirers. Indeed, since the financial institutions write the rules, they have every incentive to shift losses to others. Yet card-based payment systems continue to innovate to reduce losses that no party to the association contract bears. Moreover, the industry continues to be highly competitive,[299] with competition regularly appearing in new areas and in driving new technologies. For example, the associations make use of financial incentives to encourage the adoption of new technologies, with interchange fees varying with the merchant's authorization mechanism (e.g. POS terminal, allowing rapid settlement, or paper).[300] Moreover, while the 1990s saw a great deal of consolidation among financial institutions[301] and technical services companies' operating networks,[302] it also saw the rise of independent sales organizations (ISO). Historically the ISOs acted as sales arms of merchant acquirers, going door to door from merchant to merchant

---

297 *See supra* notes 136-137.

298 For example, the Falcon system, originally developed by HNC from technology utilized to recognize friendly tanks on the battlefield, compares authorization requests against the prior use pattern for that specific card holder. Thus if a card holder has not traveled internationally in the past ten years use of her card overseas is more likely to trigger a real-time hold or a post transaction alert (depending on the fraud and risk control policies of the issuer) than if she has frequently traveled outside her home country in the past. If the card holder has never made a major jewelry purchase, then a request for authorization for a significant purchase at a jeweler's is more likely to trigger a real-time hold or a post transaction alert than if the cardholder frequently purchases jewelry. *See* http://www.fairissac.com (for a description of Falcon; HNC was purchased by Fair Issac.). *See also* CHUTKOW, *supra* note 4, at 188-89 (describing VisaNet's similar system). By comparison, prior to the development of electronic verification,

> "authorizations involved checking an account number against numbers listed on a merchant warning bulletin. Putting a number on a merchant warning bulletin could take several weeks. And even after the listing, so long as the delinquent customer kept he amount of purchase below the floor limit, the merchant had no way of knowing that the cardholder's charge privileges had been suspended. Delinquent borrowers could make numerous purchases of this nature before the account appeared on the bulletin."

AURIEMMA, *supra* note 5, at 10.

299 The modern era is also marked by fierce competition among networks. *See, e.g.*, FRIEDMAN & MEEHAN, *supra* note 95, at 110-111 (describing American Express's competition with VISA and MasterCard in the 1980s); *id.* at 253-54 (describing competition for merchants with VISA and MasterCard, especially the 1991 restaurant "revolt" in Boston).

300 Olsen, *supra* note 156, at 17.

301 *See* Simon Kwan, *Banking Consolidation*, FRBSF ECONOMIC LETTER, June 18, 2004, available at http://www.frbsf.org/publications/economics/letter/2004/el2004-15.pdf.

302 First Data, for example, has an impressive market share and continues to expand. *See* Olga Kharif, *Why First Data Is Ready to Roll*, BUS. WK., Jan. 5, 2004, available at http://yahoo.businessweek.com/bwdaily/dnflash/jan2004/nf2004015_7492_db014.htm.

convincing the merchants to sign contracts with the merchant acquirer the ISO represented. More recently, however, ISOs used their relationships with merchants to persuade merchants to accept ISO-owned ATM machines in the merchant's facilities. These ATM machines route their transactions through a merchant acquirer via a contract between the ISO and the merchant acquirer.[303] The ISOs have thus introduced a new form of competition. Similarly, the growth of firms such as First Data provides the associations with potential competition across networks.[304] Network competition suggests the continuation of substantial differences in services across the networks.[305]

D. *Competition-Driven Evolution*

At some points in time, the evolution of card-based payment systems appears to be the paradigmatic story of an industry evolving into oligopoly. A fiercely competitive market of individual small players competes itself to the point of bankruptcy, consolidates into a few major players (the associations, American Express), and then imposes detailed rules on the entire industry. Just another "dog bites man," or, rather, a "capitalist bites consumer" tale—one of many told in the legal academy.

Closer examination reveals the dominance of competition at virtually every turn. Competition drives MasterCard and VISA, despite the common ownership of the two networks. Competition drives individual banks within both associations. Competition drives American Express to challenge VISA and MasterCard and vice versa. Competition drives First Data to build a network that could challenge the associations. We think it is fair to conclude, therefore, that the card-based payment systems market has as its primary characteristic competition. This competition is aimed at making money. To make money, the various players in the industry have harnessed reputational capital and technology to reduce losses. As a byproduct of this effort they have created a dispute resolution system that offers important advantages over public court systems.

---

303 The merchant acquirer must be a bank and a member of the network through which the ISO wishes to route the transactions.

304 *See supra* notes 142-50.

305 EVANS & SCHMALENSEE, *supra* note 1, at 151 ("When network externalities are important, multiple networks that do not interconnect can survive only if they are offering consumers substantially different services. In the case of payment cards, multiple networks exist in part because these networks offer consumers and merchants somewhat different products.").

E. *Regulation-Driven Evolution*

In this section, we review the interrelationship among the private dispute resolution system and some of the impacts of federal regulation of lending. In 1968, Congress passed the Consumer Protection Act;[306] the components most relevant for this analysis are the sections that are now referred to as the Truth-In-Lending Act (TILA). This act attempted to protect consumers by providing more transparent and clear disclosure of terms and conditions associated with lending transactions.[307] (Its success in achieving this is open to question.) Over the years, this act has been amended to expand the protections beyond the disclosure of key terms, rates, and fees.[308] Using the regulatory authority granted to it by the TILA, the Board of Governors of the Federal Reserve System adopted Regulation Z.[309] Both the TILA provisions and corresponding Regulation Z provisions directly regulate some aspects of card-based payment systems' dispute resolution systems.[310] Three provisions are particularly relevant.

First, the TILA and Regulation Z limit the liability of the holder of a credit card to $50 for unauthorized charges.[311] Capped liability thus fre-

---

306 Truth-In-Lending Act, Pub. L. 90-321, May 29, 1968 [hereinafter, TILA] 15 USC 1601 et seq.

307 Financial institutions complied with the extensive disclosure requirements by providing the mandatory notices and explanations. Consumers and even lawyers have not reacted well to these extensive disclosures mandated by the act. Harvard Law Professor Elizabeth Warren, a noted expert on contract law, claimed on public television that "I have read my credit card agreement, and I can't figure out the terms. I teach contract law. And, the underlying premise of contract law is that the two parties to the contract understand what the terms are." *See Frontline: Secret History of the Credit Card, Chapter Three, "Credit Reporting Agencies / Traps in the Fine Print"*, (PBS television broadcast Jan. 2005), http://www.pbs.org/wgbh/frontline/shows/credit/view. A discussion of whether the extensive disclosure requirements of this act should be repealed to enable better and clearer contracts is beyond the scope of this analysis on dispute resolution. *See* Richard Hynes & Eric A. Posner, *The Law and Economics of Consumer Finance*, 4 AM. L. & ECON. REV. 168, 193-95 (2002) (analyzing TILA provisions).

308 In 1974, protections were added for inaccurate and unfair credit billing and credit card practices. Pub. L. 93-495, Oct. 28, 1974. In 1980, the act was reorganized and some requirements were eliminated. Pub. L. 96-221, Mar. 31, 1980. Further changes that are beyond the scope of this analysis were made in the decades that followed. *See* Pub. L. 104-12, Mar. 19, 1995 (changes to class action suits under this act). *See also* Pub. L. 104-29, Sept. 30, 1995.

309 *See* 12 C.F.R. 226 (known in the industry as Regulation Z or sometimes Reg Z).

310 *See* 15 U.S.C. § 1643 (2005) Liability of the holder of a Credit Card; 15 U.S.C. § 1666 (2005) Correction of Billing Errors; and 15 U.S.C. § 1666i (2005) Assertion by cardholder against card issuer of claims and defenses. Within Reg Z, these key sections are implemented through, respectively: Liability of cardholder for unauthorized use (12 C.F.R. § 226.12(b)); Billing Error Resolution (12 C.F.R. § 226.13); and Right of Cardholder to assert claims and defenses against the Issuer (12 C.F.R. § 226.12(c)).

311 15 U.S.C. § 1643 ("A cardholder shall be liable for the unauthorized use of a credit card only if—(A) the card is an accepted credit card; (B) the liability is not in excess of $ 50; (C) the card issuer gives adequate notice to the cardholder of the potential liability; (D) the card issuer has provided the cardholder with a description of a means by which the card issuer may be notified of loss or theft of the card . . . ."); ("(b) Liability of cardholder for unauthorized use-(1) Limitation on amount. The liability

quently turns on what is actually authorized by the cardholder.[312] The courts have largely interpreted this to put absolute responsibility on the issuer to know and understand its consumer's intent, and therefore, if there is a question about authorization, the issuer has generally been held liable.[313]

As a result, issuers implemented rules, processes, and technologies to enable them to prove that the cardholder authorized the transaction. For example, to comply with the notification method component of the statute, issuers placed a "Lost/Stolen" telephone number on the reverse side of every card. To ensure that the end consumer actually physically received a card, when issuing and mailing out new credit cards, issuers implemented card activation technologies that required the recipient to take steps to prove to the issuer that the card had been received by the intended recipient.[314] In addition, sometimes the issuer will contact the cardholder after the first purchase using a new card, to ensure that the intended consumer actually made that purchase.

Later, if and when a consumer initiates a dispute with the complaint that a charge was not authorized, these processes and technologies can either support or refute the consumer's claim with additional information. For example, in the initiation of the dispute, if the consumer claims that she never received the card, the issuer will check the activation records. If that card was activated from the cardholder's phone number, the issuer knows

---

of a cardholder for unauthorized use of a credit card shall not exceed the lesser of $50 or the amount of money, property, labor, or services obtained by the unauthorized use before notification to the card issuer under paragraph (b)(3) of this section.") Reg Z defines "unauthorized use" as "the use of a credit card by a person, other than the cardholder, who does not have actual, implied, or apparent authority for such use, and from which the cardholder receives no benefit." 12 C.F.R. § 226.12(b)).

312 *See* Universal Bank v. McCafferty, 88 Ohio App.3d 556 (1993), 624 N.E.2d 358 (McCafferty not liable on friend's charges when issuer sent card to friend at McCafferty's request because McCafferty did not authorize friend's use of the card). The sentence construction of the act starts with the declarative statement "the cardholder shall be liable for unauthorized use of a credit card only if" and then provides a laundry list of requirements. The laundry list utilizes the conjunctive "and" between each of the provisions to indicate when a consumer is actually liable for an authorized charge. Arguably therefore, if any of the items on the laundry list fail to be true, then AND the charge is unauthorized, the cardholder cannot be held liable. Therefore, liability typically turns on authorization.

313 *See* discussion *supra* note 312.

314 The service works because a card mailed to a consumer is not "live" in the sense that the card cannot be utilized to make purchases until the consumer activates it through an activation process offered by the issuer financial institution. Although not foolproof, this activation service employs various technologies to ensure that the person receiving the card, did in fact receive the card. For example, most card activation systems require the consumer to dial a toll-free 800 style number and answer a few questions with a computer known as an Interactive Voice Response Unit. Questions may request the consumer's social security number, birth date, and other information that is not generally known. This computer also utilizes the Automated Number Identification service to collect secondary information from the phone company. Automated Number Identification is roughly equivalent to Caller Id. The incoming telephone number from the Card Activation Call is then matched and cross referenced to the telephone number on the card application and to the name the phone company has on record.

that someone in the household activated the card, perhaps even the cardholder.[315] And, if the issuer has held a previous conversation with the consumer about her first purchase, it will be much more difficult for the consumer to claim they never received the card.

A second important regulatory measure is the requirement that issuers correct billing errors.[316] Simply stated, this provision requires the card issuer to credit the consumer cardholder's account and investigate the problem when it receives a complaint about a billing error. If the consumer validates that the charge is actually correct, the correction process is undone and the charge is reinstated onto the consumer's account. The statute uses an expansive definition of billing error, including many types of claims beyond accounting and mathematical errors in the definition.[317]

---

315 Although possible, it is highly unlikely that a card thief would break into a house solely to complete the card activation from the actual consumer's home phone number.

316 15 U.S.C. § 1666 (2005) (Written notice by obligor to creditor; time for and contents of notice; procedure upon receipt of notice by creditor: "[I]f a creditor, within sixty days after having transmitted to an obligor a statement of the obligor's account in connection with an extension of consumer credit, receives . . . a written notice . . . from the obligor in which the obligor— (1) sets forth or otherwise enables the creditor to identify the name and account number (if any) of the obligor, (2) indicates the obligor's belief that the statement contains a billing error and the amount of such billing error, and (3) sets forth the reasons for the obligor's belief . . . the creditor shall, unless the obligor has, after giving such written notice and before the expiration of the time limits herein specified, agreed that the statement was correct— (A) not later than thirty days after the receipt of the notice, send a written acknowledgment thereof to the obligor, unless the action required in subparagraph (B) is taken within such thirty-day period, and (B) not later than two complete billing cycles of the creditor (in no event later than ninety days) after the receipt of the notice and prior to taking any action to collect the amount, or any part thereof, indicated by the obligor under paragraph (2) either— (i) make appropriate corrections in the account of the obligor, including the crediting of any finance charges on amounts erroneously billed, and transmit to the obligor a notification of such corrections and the creditor's explanation of any change in the amount indicated by the obligor under paragraph (2) and, if any such change is made and the obligor so requests, copies of documentary evidence of the obligor's indebtedness; or (ii) send a written explanation or clarification to the obligor, after having conducted an investigation, setting forth to the extent applicable the reasons why the creditor believes the account of the obligor was correctly shown in the statement and, upon request of the obligor, provide copies of documentary evidence of the obligor's indebtedness.") *See also* 12 C.F.R. § 226.13.

317 15 U.S.C. § 1666 ("For the purpose of this section, a "billing error" consists of any of the following: (1) A reflection on a statement of an extension of credit which was not made to the obligor or, if made, was not in the amount reflected on such statement. (2) A reflection on a statement of an extension of credit for which the obligor requests additional clarification including documentary evidence thereof. (3) A reflection on a statement of goods or services not accepted by the obligor or his designee or not delivered to the obligor or his designee in accordance with the agreement made at the time of a transaction. (4) The creditor's failure to reflect properly on a statement a payment made by the obligor or a credit issued to the obligor. (5) A computation error or similar error of an accounting nature of the creditor on a statement. (6) Failure to transmit the statement . . . to the last address of the obligor which has been disclosed to the creditor, unless that address was furnished less than twenty days before the end of the billing cycle for which the statement is required. (7) Any other error described in regulations of the Board.").

So, in the dispute resolution process, the issuer will classify a consumer dispute that is brought[318] under this section of the TILA and Regulation Z according to the association rules. As described earlier, each dispute classification will initiate specific process steps and information gathering needs. With the exception of non-delivery of goods or the statement itself, the TILA requirements affect disputes which require the sales receipt from the merchant. A billing error ultimately leads to either a consumer who denies making or simply does not recognize the charge, or to a dispute in which the amount is incorrect.[319] Therefore, a retrieval request, as described earlier, is initiated to provide the information necessary to resolve these classes of disputes.[320] Under the TILA provisions, consumers are time limited to raising billing error disputes to sixty days from the statement mailing date.[321] TILA also bars issuers from assessing interest or penalties on disputes or reporting the disputes to credit bureaus until the dispute resolution process is completed.

The third important TILA impact comes from its provisions permitting cardholders to assert claims and defenses against card issuer provisions.[322]

---

318 Of course, a consumer is unlikely to indicate that they are making a claim under the Billing Errors section of the Truth in Lending Act. Rather, that consumer will simply call the issuer and complain that the charges are incorrect. The issuer will classify the complaint.

319 Sometimes the "tip" amount at restaurants is incorrect. Therefore, the consumer recognizes the restaurant charge, but does not recognize the total amount, as accurate.

320 Of course, if the merchant does not respond to the retrieval request, the merchant may ultimately lose a chargeback dispute and the credit to the consumer's account will become permanent. If the retrieval request produces a receipt that the consumer recognizes and accepts as legitimate, the charge is reinstated onto the account and the process ends. Or, the remainder of the dispute resolution process is followed, as described above.

321 Given the timing of any particular card charge item and the printing of the statement, the consumer probably has closer to ninety days from the date of the actual charge to complain about an error.

322 15 U.S.C. § 1666i (2005) ("(a) Claims and defenses assertable. Subject to the limitation contained in subsection (b), a card issuer who has issued a credit card to a cardholder pursuant to an open end consumer credit plan shall be subject to all claims (other than tort claims) and defenses arising out of any transaction in which the credit card is used as a method of payment or extension of credit if (1) the obligor has made a good faith attempt to obtain satisfactory resolution of a disagreement or problem relative to the transaction from the person honoring the credit card; (2) the amount of the initial transaction exceeds $ 50; and (3) the place where the initial transaction occurred was in the same State as the mailing address previously provided by the cardholder or was within 100 miles from such address, except that the limitations set forth in clauses (2) and (3) with respect to an obligor's right to assert claims and defenses against a card issuer shall not be applicable to any transaction in which the person honoring the credit card . . . (E) has obtained the order for such transaction through a mail solicitation made by or participated in by the card issuer in which the cardholder is solicited to enter into such transaction by using the credit card issued by the card issuer. (b) Amount of claims and defenses assertable. The amount of claims for defenses asserted by the cardholder may not exceed the amount of credit outstanding with respect to such transaction at the time the cardholder first notifies the card issuer or the person honoring the credit card of such claim or defense.") *See also* 12 C.F.R. § 226.12 (2005). Note that the implementing regulations specifically exclude debit cards and similar non-credit cards. Although the statute is based on the extension of credit, the question of what is a credit card is left unan-

The statute and regulations provide that a cardholder may assert a claim or defense on the underlying transaction to the issuer, if that cardholder has first made a good faith attempt to resolve the dispute with the merchant. Note that the assertion of a claim or defense on the underlying transaction is mutually exclusive with authorization or billing error disputes, described above. So, in asserting a claim or defense, the consumer is admitting that the transaction occurred, and that the consumer authorized that transaction.

The net effect of TILA and Regulation Z is to shift responsibility associated with the payment system away from the consumer toward the financial institution that issues that card. Ultimately, however, the issuer does not actually bear the costs associated with these disputes. By contract and rule, losses associated with disputes that the cardholder wins are shifted to the acquirer and ultimately to the merchant, and thence to the consumer through higher prices.[323]

Federal regulation, through TILA and Regulation Z, has had an impact on card-based payment systems' dispute resolution mechanisms. The regulatory provisions described above provided reasons for the development of particular provisions in those dispute resolution systems. We are unconvinced, however, that either TILA or Regulation Z deserves more than minimal credit for the success of the card-based payment systems' dispute resolution procedures.[324] Regulation mandated a limit on consumer card losses, provided incentives for some minor card features (e.g. the 800 "lost or stolen" number on the back), and required some procedure to assert defenses against charges.

It is far from clear, however, that these provisions (or something like them) would not have been adopted in response to competitive pressures. As we described earlier, the early credit cards were particularly vulnerable to fraudulent use, as issuers exercised little to no control over the credit worthiness of cardholders and made indiscriminate mailings of cards. Fraud prevention measures resulted in the first instance from the financial losses experienced by issuers as a result of fraud. Moreover, since consum-

---

swered. There is a legitimate argument that a debit card would fit into the credit focus of the statute because there are temporary extensions of credit granted in the payment process of moving money through the chain of participants from a consumer, through a debit card issuer, an association, a debit card acquirer, and ultimately to a merchant's settlement account.

323 Of course, if the dispute resolution processes that the associations have implemented lead to an accurate result, the merchant would have had to pay these damages anyway except that the consumer would have had to resort to formal public legal means, including lawyers' fees and potentially punitive damages.

324 Much of the legal literature seems to accept that Regulation Z is responsible for the success of the dispute resolution system. *See, e.g.*, Jane K. Winn, *Making XML Pay: Revising Existing Electronic Payments Law to Accommodate Innovation*, 53 S.M.U. L. REV. 1477, 1491-92 (2000) ("Regulation Z provides not merely a billing error resolution procedure and protection from liability for unauthorized use of the credit card, but it also provides a simple and effective alternative dispute resolution process in the event the consumer is unhappy with the transaction itself.").

ers had to be persuaded initially to adopt credit cards as a means of payment and since competing payment systems were less vulnerable to fraud (e.g. checks, cash), adoption of dispute resolution provisions and antifraud measures generally (although not any particular provision) would have been compelled by market pressures.[325] Finally, the major card networks have extended provision of dispute resolution procedures internationally, well beyond the reach of Regulation Z.[326] Similarly, domestic debit card networks and issuers have generally provided equivalent dispute resolution procedures, despite the inapplicability of Regulation Z, to protect their brand names.[327] We conclude, therefore, that TILA and Regulation Z had some impact on the shaping of the dispute resolution systems associated with card-based payment systems, but the regulatory impact was less important than the impact of competition.

V. CONCLUSION

We opened this paper by proposing a "radical rethinking of dispute resolution" based on card-based payment systems' dispute resolution procedures. Those who have read this far (without skipping!) would undoubtedly like to know what that radical rethinking is.

Card-based payment systems' procedures for resolving disputes look nothing like the procedures used by the public legal system. In place of lawyers, judges, and juries, card-based payment systems use clerical employees, simple processes, and technology. In place of notice pleading, they use something that more closely resembles the old common law forms of action than anything else we have encountered in the modern world. In place of liberal discovery rules, they use restrictive rules providing for limited discovery. In place of clever lawyering, they use structured, semi-automated interviews. Despite all these differences, card-based payment systems' dispute resolution systems do not seem to be sparking any signifi-

---

325 EVANS & SCHMALENSEE, *supra* note 1, at 324 ("Payment cards have not flourished just because they provide a convenient alternative to cash and checks. Over time, entrepreneurs have discovered that they can integrate other products and services into payment cards and thereby make these cards more valuable for consumers and merchants.").

326 *See, e.g.*, FTC filings by VISA and American Express quoted in Perritt, *supra* note 42, at 690, n.70 (quoting VISA Senior Vice President that "The chargeback reasons permitted under VISA's rules for international transactions have been adopted to enable issuers of VISA cards to address the fundamental consumer concerns of their cardholders, and incidentally to reinforce the reputation of VISA Cards as the best way to pay" and quoting American Express Group Counsel that ""[w]hile U.S. law requires us to institute these practices, as a card issuer, we have adopted a policy of applying them consistently outside the U.S. as well.").

327 *See* Balto, *supra* note 63, at 1104-05 (describing voluntary steps by VISA and MasterCard; although Mr. Balto does not see these steps as sufficient and calls for regulation, we disagree with his assessment.).

cant consumer or merchant revolts. Either the systems do not harm consumers or merchants, or any harm they cause is overwhelmed by the benefits of the system.[328] There is, therefore, some evidence that consumers and merchants alike accept this sort of system.

Is it better? The systems we describe above handle a high volume of disputes (generally low value disputes, to be sure) and handle them quickly and cheaply. For at least some classes of disputes, therefore, we think this type of system is clearly better, so long as we define "better" as "cheaper." While we are sensitive to the traditional claims that the legal system serves a higher function of justice, fairness, and consistency, rather than to simply cheaply resolve disputes among private parties, we are also skeptical about the frequency with which that argument is used to justify what appears to be rent-seeking by those with an interest in the current public legal system (lawyers, legislatures, judges, etc.)[329]

We think the card-based payment systems model is better in ways other than being simply cheaper. In particular, the model provided by card-based payment systems is better for disputes which arise from the strategic behavior of one or more of the parties. Such behavior is rampant in the public legal system because of the structure of the litigation process. [330] Because these systems incorporate information about disputes generally, as well as about specific parties, they are better able to identify and correct strategic behavior by disputants. Moreover, because the card-based payment systems can change the rules governing the use of the cards, card issuers can learn from present disputes how to avoid future disputes and implement new rules which prevent such disputes from arising in the future.

One of the reasons the card-based payment systems' dispute resolution systems succeed may initially appear counter-intuitive for law professors (at least it did for the ones who coauthored this piece). By simplifying disputes into categories and then applying managerial expertise to ruthlessly drive down costs, these dispute resolution systems not only do not use lawyers, they have no room for them. As Professor Hadfield notes, "Brilliant

---

328 If the dispute resolution systems did harm either consumers or merchants but left net benefits positive, presumably competitive pressures would push some card to offer superior dispute resolution. Dean Perritt reaches a similar conclusion from his observation of disputes: "Although good empirical data is lacking, it appears that the system satisfies both consumers and merchants. Almost no reported cases in the regular courts exist, suggesting that consumers rarely are motivated to go beyond the chargeback process to more formal forms of dispute resolution." Perritt, *supra* note 42, at 691.

329 Astoundingly, one article referred to the American Bar Association as "one of the few neutral, non-stakeholding but nongovernmental or intergovernmental entities" involved in dispute resolution, despite acknowledging that "a portion of its members are obviously interested in the provision of ADR services and therefore have some vested interest." Louise Ellen Teitz, *Providing Legal Services for the Middle Class in Cyberspace: The Promise and Challenge of On-Line Dispute Resolution*, 70 FORDHAM L. REV. 985, 1005 (2001).

330 Hadfield, *supra* note 11, at 972 ("The process of litigation is a series of strategic moves and countermoves: sophisticated moves require even more sophisticated responses.").

lawyering is the art of drawing out and then persuading others of the saliency of distinctions and similarities that were not previously recognized."[331] The essence of the dispute resolution systems described here, however, is that they reject the introduction of "distinctions and similarities" not embedded in the rules. In short, if there is no code for a dispute, there is no dispute.[332] If there is a code, everything from the acceptable evidence to the time limits for the process are dictated by the code. The code is applied not by a highly trained lawyer, but by a clerk. We think this is critical to reducing the costs of dispute resolution.[333] Moreover, card-based systems' reliance on simple procedures eliminates an important cost to complexity: the difficulty of explaining complex systems to consumers.[334] Professor Hadfield raises an important point about such solutions. After identifying complexity as a key problem in the public legal system,[335] she notes that while reducing complexity is a natural area for improvements, doing so raises "deep philosophical and practical questions which all come down to this: is legal reasoning, as we know it, what law and justice is?"[336] We are not sure that it is. After talking with many participants in the dispute resolution system of card-based payment systems, both on and off the record, we do not see injustice in its often inflexible rules. In other words, adding lawyers to the process would not obviously lead to an improvement in any dimension.

The card-based payment systems' dispute resolution processes described here do not meet the traditional public legal system-oriented definition of due process.[337] For example, Judge Henry J. Friendly defined eleven

---

331 *Id.* at 966.

332 In a sense, we are echoing the "most significant principle to emerge from the academic study of law on the Internet . . . the ideas that software code . . . is broadly substitutable for legal code . . . Code is law; architecture is control; software is power." R. Polk Wagner, *On Software Regulation*, 78 S. CAL. L. REV. 457, 459 (2005).

333 Prof. Hadfield notes that "the hours required to resolve a legal matter are not fixed by abstract and immutable principles of justice. They are determined by the procedures and reasoning requirements established and implemented by members of the profession (lawyers and judges and legislators) in an antagonistic, interactive process." Hadfield, *supra* note 11, at 965. It is by the use of forms and procedures that the card-based payment systems are able to control the dynamic to complicate disputes that Hadfield observes in the public legal system.

334 This simplification is a significant concern of "consumer advocates" in their critiques of alternative dispute resolution systems. *See, e.g.*, Krause, *supra* note 6, at 480 ("Many consumer advocates voice one central, underlying concern. At what point does the imposition of a private dispute-resolution requirement place too big a burden on the average consumer. That is, when is the sophistication and savvy of the average Internet user not enough to level the playing field.").

335 Hadfield, *supra* note 11, at 995 ("The complexity of legal reasoning and process is fundamental to the entire market. It is the source of direct cost, as we have seen. But more importantly it plays a central role in a host of indirect distortions.").

336 *Id.* at 1002.

337 For example, Prof. Gibbons is critical of arbitration in the consumer context because it is 'unfair': limited discovery, lack of a jury trial or a right to appeal, repeat-player advantages in selecting arbitrators, no class relief, and excessive fees unfairly disadvantage individuals bringing claims." Gib-

elements of procedural due process as: (1) an unbiased decision maker; (2) notice and a statement of the reasons for the initial action that causes the dispute; (3) an opportunity to present reasons why the action should not be taken; (4) an opportunity to present evidence, including witnesses; (5) the right to know opposing evidence; (6) the right to cross-examine opposing witnesses; (7) limiting the decision to the evidence in the record; (8) the right to be represented by counsel; (9) a record of evidence prepared by the decision maker; (10) the decision maker gives reasons for the decision; and (11) the availability of appellate review.[338] Some have argued that these elements should also apply, at least in part, to non-judicial forms of dispute resolution.[339]

Many of Friendly's eleven elements do not apply to the card-based payment systems' dispute resolution processes. There is no right to counsel, no opportunity to cross-examine opposing witnesses, no right to know opposing evidence, no record available to the parties, and no written decision given to parties stating reasons for the decision. Moreover, it is at least arguable that the decision maker is insufficiently neutral to qualify in Friendly's definition, or that either the record or the internal appeals process would satisfy his definitions. At most, therefore, six of his eleven criteria are satisfied and possibly as few as four. Rather than forming a basis for condemning the card-based payment systems, we suggest that these differences should prompt a radical rethinking of the value of traditional due process in dispute resolution. In the case of card-based payment systems' dispute resolution procedures, we conclude that the incentives provided by competition serve as an effective substitute for formal procedural due process structures.

Another criticism of card-based systems is that consumers lack information or the incentive to bargain with card issuers, making government regulation necessary to even the playing field.[340] We contend that such arguments fail to grasp the power of competition to induce fair outcomes. In the case of card-based payment systems, the competitive pressures on card issuers, the introduction of merchant acquirers as repeat players, and networks' role provide incentives for card issuers to treat cardholders fairly without requiring cardholders to invest in knowledge about the details of

---

bons, *supra* note 49, at 15. *See also* Lucille M. Ponte, *Boosting Consumer Confidence in E-Business: Recommendations for Establishing Fair and Effective Dispute Resolution Programs for B2C Online Transactions*, 12 ALB. L.J. SCI. & TECH. 441 (2002) (describing an elaborate set of principles proposed by the American Arbitration Association's National Consumer Disputes Advisory Council, few of which are met by the card-based systems).

338 Henry J. Friendly, *Some Kind of Hearing*, 123 U. PA. L. REV. 1267, 1279-95 (1975).

339 *See* Perritt, *supra* note 42, at 679-83.

340 *See, e.g.*, Effross, *supra* note 62, at 376 ("given consumers' lack of incentive or knowledge to bargain, and likely reluctance to litigate, especially where their adversaries would tend to be sophisticated financial institutions and the amount at issue relatively small, federal regulation is warranted.") (citations omitted).

the system. Demanding that such systems replicate the institutions of the public legal system only ensures that they cannot innovate and so will have the same general failings and successes as the public system.

Consider for example the idea of an unbiased decision maker. In the public legal system this is ensured by providing the judiciary with independence.[341] In card-based systems it is provided by the competitive pressures on the networks by other networks. For example, if a network treats a merchant acquirer or issuer unfairly on a regular basis, that entity will switch its allegiance generally, or a greater share of its transactions, to a competing network and issuers and merchant acquirers who treat cardholders or merchants unfairly will lose market share to competitors.

Lawyers have generally been able to maintain control of alternatives to the public legal system because most alternatives depend upon the public legal system to enforce their decisions.[342] Card-based systems show that this need not be true where the dispute resolution mechanism is part of a good or service desired independently. Merchants and consumers (usually) accept the results of the dispute resolution process because they want to continue to participate in the payment system. Instances in which either seek redress in the courts are relatively rare, (admittedly a judgment largely based on the infrequency of reported opinions, given the number of cards, merchants, and cardholders). Moreover, the card-based payment system actors (issuers, acquirers, networks, card companies) profit by using their dispute resolution procedures to lure customers to their networks.[343]

The strength of the card-based payment system lies in competition's incentives to develop better, more accurate, cheaper, and faster processes and its ability to harness reputation and learn from experience. Neither characteristic is a feature of the public legal system. We therefore contend that expanding those characteristics would likely lead to better dispute resolution processes for disputes currently in the public legal system.

How can this type of dispute resolution system be expanded beyond the card context? Wherever repeat players analogous to the associations and proprietary networks exist in a competitive environment, there is potential for extending these systems. Ironically, some of the easiest may be in

---

341 Daniel Klerman and Paul Mahoney provide a concise definition of independence: "A fully independent judiciary is one in which judges enjoy tenure during good behavior, a salary sufficient to shield them from pressure from either government or private parties, sufficient prestige that the hope of promotion to a more prominent post is not a large motivator, a system of prerequisites (location and appointments of offices, etc.) that is hard for the government to manipulate, and rules regarding jurisdiction over cases that are resistant to executive and legislative meddling, among others." Daniel M. Klerman and Paul G. Mahoney, *The Value of Judicial Independence: Evidence from 18th Century England*, Am. L & Econ Rev. (forthcoming) (available at http://ssm.com/abstract=587383).

342 Hadfield, *supra* note 11, at 994.

343 *See* Gibbons, *supra* note 5050, at 3 ("American Express®, Visa, MasterCard®, Discover®, JCB®, and other credit card issuers are arbitraging the risk [of e-commerce] by being the dispute resolution mechanism of last resort for most B2C e-commerce transactions.").

the context of what we traditionally view as interactions between strangers. Automobile accidents, for example, generally occur between parties who have already contracted with insurance companies.[344] Treaties between insurance companies could institute dispute resolution processes with characteristics like those of the card-based payment systems. Medical insurance, if it could be freed from the employer linkage created by the tax deductibility of insurance premiums, offers another potential vehicle for extending the model.

For these reasons, we advocate radically rethinking assumptions surrounding dispute resolution.

344 Today, when a car hits a pedestrian, both parties may not be insured. If, however, systems of dispute resolution develop (as described in this section), a strong incentive for pedestrians to insure themselves may develop, so that pedestrians too, would be able to reap the benefits of such a radical efficiency orientation.

WHO'S TO PROTECT CYBERSPACE?

Christopher J. Coyne, Ph.D. & Peter T. Leeson, Ph.D.***

ABSTRACT

Until now, the evolution of cyber security has been largely driven by market demand and has developed in the absence of formal governance. However, in the post-9/11 world and with an increase in cyber attacks, government's role in cyber security has become a major policy issue. This paper contends that economic principles have been excluded from the debate about who should provide cyber security. This paper seeks to fill this gap. We postulate that an analysis of cyber security in the absence of economic considerations is incomplete. Toward this end, we employ several economic concepts in order to offer insight to policymakers involved in this debate. In doing so, we hope to shed light on the most effective means of securing the Internet.

1. INTRODUCTION

Over the past decade, the growth of cyberspace has enabled individuals across the world to become increasingly connected. Table 1, which shows Internet access for different languages, highlights the extent of Internet expansion across borders and cultures:

Language	Internet Access (millions)	Percentage World Population Online	2004 (est. millions)
English	**262.3**	**35.6**	**280**
European Languages	257.4	34.9	328
Asian Languages	216.9	29.4	263
Total Non-English	**474.3**	**64.4**	**680**
Total World	**679.7**		**940**

Table 1: *Global Internet Statistics by Language (2003)*[1]

* Department of Economics, Hampden-Sydney College. Email: ccoyne@hsc.edu.
** Department of Economics, West Virginia University. Email: pete.leeson@mail.wvu.edu.

The development and expansion of the Internet has created innumerable new opportunities for access to information, personal interaction and entrepreneurial ventures.[2] Not only have the costs of communication fallen considerably but also, perhaps even more importantly, the sphere of potential trading partners has expanded dramatically creating immense new gains from exchange. Consider, for instance, the increase in eCommerce over the last four years, as illustrated in Table 2:

	2000	2001	2002	2003	Estimated 2004
Total $ (B)	$657.0	$1,233.6	$2,231.2	$3,979.7	$6,789.8

Table 2: *Worldwide eCommerce Growth*[3]

This is a tenfold increase over a four-year period. The online banking industry also highlights the increasing reach of cyberspace. The number of individuals using online banking services has increased 80 percent, from 13 million to 23.2 million, in the period from September 2001 to September 2003.[4] These rising trends illustrate the general fact that the lives of average citizens are becoming increasingly connected to cyberspace. This interconnectedness goes beyond direct interaction with cyberspace and extends to indirect interaction as well. Many of the services that the average individual relies on—water, electricity, mass transportation and other "critical infrastructure"—are linked to cyberspace although the end user may never realize it.[5] From direct interactions on personal computers and business networks to indirect interactions through critical infrastructure, the existence and development of cyber security is of the utmost importance for cyberspace to achieve its full potential.

---

[1] Source: Global Reach (http://www.glreach.com/globstats/index.php3). Note that the "Total World" does not equal the sum of "Total English" and "Total Non-English." This discrepancy is due to an overlap between English and non-English figures. Many users access the Internet in two languages twice. The "Total World" row is lower than the sum to correct for this overlap. For more on the methodology see: http://global-reach.biz/globstats/refs.php3#overlap.

[2] Varian et al conclude that the world wide web contains a textual content equivalent to that contained in 10 to twenty million books (McMillan 2002, p. 156).

[3] Source: Global Reach (http://www.glreach.com/eng/ed/art/2004.ecommerce.php3).

[4] *Nashville Business Journal*, September 22, 2003 (http://www.bizjournals.com/nashville/stories/2003/09/22/daily5.html).

[5] The Patriot Act defines critical infrastructure as: "Systems and assets, whether physical or virtual, so vital to the United States that the incapacity or destruction of such systems and assets would have a debilitating impact on security, national economic security, national public health or safety, or any combination of those matters."

Cyber security involves freedom from the risk of danger when interacting in cyberspace. As indicated, we consider participation in cyberspace to encompass a wide-range of activities including both direct and indirect interactions. Security takes on many different forms in cyberspace including encryption techniques, firewalls, virus-scanning software, intrusion detection systems and secure payment systems. In the absence of security, the full potential of information technologies cannot be realized because users will be fearful of malicious activities (Cheswick and Bellovin 1994). From simple searches, downloads and communication on the Internet to more complex transactions, individuals require security for their hardware, software, personal information and online exchanges. In addition to the range of activities that require security, there is also a range of Internet users demanding a secure environment. These users include private individuals, businesses and government.

The increasing interconnectedness discussed above does come with the possibility of significant losses through cyber crime. For instance, in 2003, hacker-created computer viruses alone cost businesses $55 billion. This is nearly double the damage they inflicted in 2002 (SecurityStats.com 2004). In a 2004 survey by the Computer Security Institute (CSI), over half of respondents indicated some form of computer security breach over the past twelve months and 100 percent of respondents indicated a website-related incident over that same period (CSI 2004).

In the post-9/11 world, Internet security has become a major policy issue, specifically in the context of national security. Consider for instance the following from Tom Ridge, the former Director of Homeland Security:

> "When people think of critical infrastructure, they have a tendency to think of bricks and mortar . . . . But given the interdependency of just about every physical piece of critical infrastructure, energy, telecommunications, financial institutions and the like with the Internet and the cyber side of their business, we need to be focused on both and will be . . . . We [the government] need to do a national overview of our infrastructure, map vulnerabilities, then set priorities, and then work with the private sector to reduce vulnerabilities based on our priorities" (Quoted in Verton 2003, p. 235).

One of our main aims in this paper is to provide a realistic understanding of how cyber security fits in with national security. Is it our contention that in the context of cyberspace, individual security, as it relates to each and every user, and "national security" are inseparable. Just as security at the personal level involves the absence of risk of danger, so too does national security. Indeed, neatly categorizing national security as its own distinct category, separate from cyber security is a difficult task. This is largely due to the fact that national security is directly dependent upon security at the lowest levels of cyber usage.

We often think of national security as a single good provided by government, national defense being one example. Cyber security, however, is distinctly different than this because at the national level it is simply the

sum of dispersed decisions of individual users and businesses. Highlighting the role that individual users play, Verton writes, "Millions of home computer users with high-speed Internet connections fail to secure their connections, and become potential 'jumping off' points for terrorists and malicious hackers" (2003, p. x). The very essence of the Internet is interconnectivity. What this means is that national security concerns are directly linked to the most basic security issues that the average user faces.

In light of this, it is easy to see why cyber security is currently one of the main policy topics of discussion. The development of cyber security and growth of cyberspace in general has taken place with little central direction. According to its inventor, Tim Berners-Lee, the Internet grew "by the grassroots effort of thousands."[6] Currently, it is estimated that eighty percent of what is deemed "critical infrastructure" is privately owned (Verton 2003, p. x). Potential problems arise, it is argued, specifically because of the Internet's decentralized nature. In short, no one user will be looking out for the national interest and hence national security. It is increasingly common nowadays to hear that the absence of coordinated efforts to protect cyberspace means vulnerabilities will persist. Given this, the conclusion often drawn is that the government must play an active role in protecting cyberspace against cyber crime and cyber terrorism.[7] The exact role that government is to take is still being debated.

As the title of this paper suggests, we focus on answering the question, "Who's to protect cyberspace?" Our core thesis is as follows: Although economic issues are at the center of cyber security, economic considerations have been largely absent from the policy debate. Economics can contribute to adjudicating between the various courses of action in determining policy toward cyber security. Toward this end we employ several basic economic concepts in order to offer insight to policymakers involved in this debate. In doing so we hope to shed light on the most effective means of securing the Internet.

Those in the legal profession have focused on governance issues related to cyberspace, which are closely linked to the issue of security. For instance, Johnson and Post (1996a, 1996b) postulate that since the Internet is not linked to any geographical polity, governance will take place via privately provided rules that lead to the emergence of common standards. Reidberg (1996) argues that the primary source of governance in cyberspace is technology developers. It is his contention that the hardware and software that allows users to operate in cyberspace imposes a set of default rules. Neither of these works, though, incorporates explicit economic analysis into their work. Our paper can be seen as contributing to this dis-

---

[6] *San Jose Mercury News*, January 30, 2001, books section, p. 2.

[7] Pollit (1997) defines cyber-terrorism as: "The premeditated, politically motivated attack against information, computer systems, computer programs, and data which results in violence against noncombatant targets by subnational groups or clandestine agents."

cussion on governance, its new contribution being a focus on the economic aspects of cyber governance and security. There is also a growing body of literature in the area of the economics of information security (see for instance Anderson 2001; Camp and Lewis 2004). While the insights from this literature are extremely relevant to this debate, they have been largely neglected in both the private and policy realms.[8] Given this, and in light of increasing calls for government involvement in cyber security, it makes sense to highlight what economics can contribute.

This paper proceeds as follows. We first apply the economic concepts of marginal costs, marginal benefits and efficiency to the issue of Internet security. Section 3 discusses and applies the concepts of externalities and market failure to cyberspace. In light of this discussion, Section 4 highlights some ways that the market can overcome problems stemming from externalities. Section 5 considers the concept of government failure and the implications for government regulation of cyberspace. Section 6 discusses the policy implications stemming from our analysis. Section 7 concludes by reiterating the main points of our analysis.

2. MARGINAL COSTS, MARGINAL BENEFITS AND THE EFFICIENT LEVEL OF INTERNET SECURITY

When considering any potential course of action, economists focus on weighing the benefits of the action versus its costs. More specifically, economists are concerned with the costs and benefits of undertaking an additional, or marginal, unit of the activity in question. If there is a net gain, where the marginal benefits outweigh the marginal costs, the activity should be undertaken, the result being an economic improvement. Likewise if the marginal costs outweigh the marginal benefits, the activity in question should not be undertaken. Economists refer to a situation as efficient if all possible improvements have been made such that no further improvements are possible.

The logic of efficiency has clear implications for cyber governance and security. If asked, most people would say that the optimal level of cyber breaches is zero.[9] But economics tells us otherwise. From an economic standpoint, what we want is the *efficient level* of cyber breaches. If the damage done by a breach is greater than the cost of the cheapest means of preventing it, than the breach is inefficient and should be eliminated. Likewise, if the cost of the cheapest means of preventing the breach is

---

8 See for instance, "The New Economics of Information Security," Information Week, March 29, 2004. Available at: http://www.informationweek.com/story/showArticle.jhtml?articleID=18402633 (last accessed 7/12/04).

9 We use the term "breaches" here in the broadest possible sense to include such things as hacking, viruses, fraud, cyber terrorism, etc.

greater than the benefit gained, the breach is efficient. Ultimately, what this means is that the efficient level of cyber breaches is not necessarily zero. For instance, if it costs $1 million to prevent a virus or cyber attack that only causes $500,000 worth of damage, the prevention should not be undertaken. In this example, the costs of prevention outweigh the benefits, and it is an efficient cyber breach.[10] We now have a general economic rule for considering the efficient level of computer security. Security efforts should only be undertaken if the marginal benefits outweigh the marginal costs. In general, the efficient level of cyber breaches is where the marginal costs of prevention exactly offset the marginal benefits of prevention.

In many cases, security efforts will be undertaken to prevent potential attacks, which may or may not in fact occur. For example, many of the current efforts undertaken by the government against cyber terrorism are done to prevent a potential attack from occurring. In such cases one can determine an expected probability that such an attack will in fact occur and calculate the expected cost and expected benefit of undertaking the security measure to prevent that attack from occurring.

The immediate implication of applying the basic concepts of marginal costs, marginal benefits and efficiency to cyber security is that the end goal of policy is not necessarily to reduce the level of cyber breaches to zero. Instead, we should aim for a policy mix that yields the efficient level of breaches. Ultimately, what we want to achieve is a policy that sets the punishment for a breach equal to the cost of damage. If this can be achieved, only efficient breaches will be undertaken. In other words, those engaged in breaches will only commit breaches when the benefit they receive is greater than the cost (i.e., damage). Another implication is that considering only the aggregate number of breaches as a metric of the general cyber environment is not informative from an economic standpoint. The number of breaches tells us nothing about the cost they impose or the benefit of preventing them.[11]

The main difficulty with the cost-benefit approach is obtaining the relevant information to determine actual costs and benefits. This becomes even more difficult when attempting to perform this analysis on breaches that may or may not occur because this involves some degree of speculation, not only regarding the probability of a breach, but also the damage it will cause.[12] As we will discuss below, the market is one means of generat-

---

[10] There have been several attempts at measuring the costs of cyber breaches. See for instance, PricewaterhouseCoopers (2000).

[11] For instance, part of the hacker subculture consists of hackers who breach a system and without doing any damage report the security holes to the system administrator. In this sense, they actually provide a benefit in repairing security holes before malicious hackers can take advantage of them. This benefit is not captured when one considers the total number of breaches and it is not clear that one would want to expend resources in preventing these breaches.

[12] The efficient level of security has been debated by among others Anderson (2002) and Schneier (2002).

ing the knowledge required for cyber security investments. Despite these difficulties, we now have a framework in place to judge the efficiency of security efforts.[13] One thing that is clear is that ignoring costs and benefits leads to an incomplete analysis and can potentially lead to wasted resources.

3. THE THEORY OF EXTERNALITIES AND MARKET FAILURE

The notion of externalities is also extremely relevant to the discussion of cyber security. Economists define an externality as a net cost or benefit that an activity imposes on those outside (i.e., external to) the activity. The problem stemming from externalities is that an individual only considers the costs and benefits directly relevant to him. In other words, an individual's decision excludes the costs and benefits that the activity imposes on others.

Externalities can be either positive or negative depending on whether they yield an external benefit or cost. A common example of a positive externality is a scientific research breakthrough. In this case, the good produces a positive externality that has large spillover benefits to those outside the individuals actually engaged in the scientific research. In the case of positive externalities, the primary actor does not internalize all benefits of his action. Theoretically, positive externalities will be undersupplied on the market due to the free-rider problem stemming from non-excludability and pricing issues related to non-rivalry. One common example of a negative externality is pollution from a factory. In such cases, the primary actor does not internalize all costs of his action. Theoretically, negative externalities will be oversupplied because the producer will internalize all benefits of the activity but not all of the costs.

Externalities are said to lead to market failure because the market fails to efficiently distribute costs and benefits such that they are fully internalized. In other words, the market, left to its own devices, will fail to provide the incentives to produce the socially optimal level of goods with positive or negative externalities. The standard conclusion is that government must either be involved in producing the good or service, or must regulate the activity in question in order to align costs and benefits and to ensure externalities are internalized. In the case of negative externalities, government usually penalizes the behavior, while in the case of positive externalities it usually encourages the behavior through subsidies or other incentives.

Given the above rendering of externalities, we can now place cyber security within this context. First, it must be noted that the Internet pro-

---

[13] It should be noted that there is software, for example CORA, which allows firms to calculate the return on a security investment. The software analyzes the costs of security breaches in terms of recovery time and weighs those costs against the benefits of investing in the prevention activity.

duces what economists refer to as a network externality in that the value of each connection increases as the total number of connections increases. For instance, while one Internet connection may allow the user to search for specific information, the value of the connection increases as others begin to use the Internet as well. With more connections, there are more users to interact with, whether the purposes are commerce, information or entertainment.

Given the interconnectedness of cyberspace, the actions taken by users will spill over and affect other users. These spillovers can be either positive or negative depending on how we look at the issue. The failure to undertake security measures can potentially have large negative effects on other users. If two users are connected and one fails to secure their system, he is putting the other user at risk as well. Likewise, security efforts undertaken by some users will provide a positive spillover to other users. To understand why, consider an analogy with vaccines. The prevention of communicable disease yields enormous spillover benefits to all members of a society. In other words, each member of a community benefits (i.e., receives a large positive benefit) if the other members of the community are vaccinated against a disease because they do not have to be concerned that they will catch the disease. A potential problem arises though because there is an incentive to free ride. If each individual believes that all others will be vaccinated, there is no reason for them to be vaccinated as well. The case with cyber security can be seen in a similar light. If everyone else's computer is vaccinated against viruses and protected against breaches, other members of the cyber community benefit as well and don't need to take steps to protect their system. For instance, those interacting with the uninfected user who regularly scans his computer do not have to be concerned with receiving a virus infection from that user.

As such, when individual users or businesses take steps to make their own computer or business more secure, they make the general cyber environment more secure as well, thus benefiting all users. Given this, economic theory predicts that individual decision calculus will yield too little security. The individual undertaking the security precautions does not internalize all the benefits, and will seek to free-ride off of the efforts taken by others. Similarly, when users fail to undertake security measures, they only incur part of the cost of their actions. Therefore, theory predicts that security will be undersupplied on the market and vulnerability, or a lack of security, will be oversupplied on the market.

Although not using the exact terminology specified above, policymakers often view cyber security within this framework. To illustrate this, consider the following quote from former Governor James Gilmore who led the Advisory Panel to Assess Domestic Response Capabilities for Terrorism Involving Weapons of Mass Destruction: "So far, pure public/private partnerships and market forces are not acting . . . to protect the cybercommunity. Relying on the private sector's willingness to do the right thing when

it comes to security is simply not an answer." (Quoted in Verton 2003, p. 26). In economic terms, Gilmore is indicating that a market failure exists due to a lack of incentive on the unhampered market to "do the right thing" and provide the optimal level of cyber security. Indeed, the notion of externalities and market failure underlies all claims that the market will underproduce cyber security and that the government must intervene and regulate to makeup for the shortfall. Consider the following from Richard Clarke, the former cyber security czar:

> I went around saying that regulation was a bad thing because the government was stupid and would do it badly . . . . But the thing about regulation is that there was always a footnote—like, unless there's market failure, we don't want regulation. If the market doesn't cause voluntary processes [to change], then government gets involved.[14]

The immediate concern that results from issues of externalities and market failure are how these problem can best be remedied. There are at least two possibilities for dealing with the problem. One involves considering possible ways for the market to privately solve externality problems. The second is for government to intervene via regulation. In the next two sections, we treat each of these potential solutions in turn.

4. PRIVATE SOLUTIONS TO EXTERNALITIES

Given that cyber security measures have large positive spillovers, economic theory predicts that these measures will be undersupplied on the market. The question then becomes whether economic theory's predictions are correct or if there are means through which the market can internalize the related externalities. Typically, there are several avenues through which goods possessing strong externalities can be privately supplied.

The key realization is that not all benefits have to be internalized for a good with externalities to be produced at the optimal level. Indeed, nearly every activity has some related externality. The good can be privately produced provided that there are solutions that allow *enough* of the benefits to be fenced off and internalized by the producer. Similarly, the presence of spillovers is itself not enough to prevent some producers from providing a needed good. Some producers may be motivated by good-will or act for other reasons unconnected to monetary rewards and therefore are willing to incur the cost of providing say, a public good, even though they gain little (or even lose) from a profit and loss perspective. In the following subsections we consider these two avenues through which goods possessing positive externalities are privately supplied in the context of cyber security.

---

[14] Source of quote: "RSA: Can regulation cure security's ills?", available at: http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci953148,00.html (last accessed 6/7/04).

4.1 Private Provision via Voluntary Donation

Voluntary donations are one method of funding goods with large positive externalities. Donations of money and artwork to museums, contributions to listener and viewer-supported radio and television stations, and donations to health research all serve as some readily apparent examples. While economic theory would predict free-riding in such situations, we observe many individuals making such donations nonetheless.

There are several instances of the private provision of cyber security by the voluntary donation of time and/or money, completely separate from any government organizations encouraging this behavior. One example of this is CyberAngels, an organization that was founded in 1995 by Curtis Sliwa, head of the Guardian Angels. CyberAngels is a completely voluntary program whose goals include: (1) preventing online crimes through education, (2) assisting victims who have suffered from Internet crimes and (3) monitoring legal issues as they relate to the Internet across borders.[15] In line with these goals, the activities of the CyberAngels include searching for online fraud and scams, finding and reporting sites that use children in sexually provocative ways, monitoring children in child chat rooms, offering online classes and assisting victims of online harassment, stalking, fraud and hacking.[16] CyberAngels is funded through private donations from various donors ranging from individuals to corporations.

Microsoft's bounty program provides another illustration of the private provision of cyber security through private donations. In November of 2003, Microsoft announced that it was creating an anti-virus reward program backed by $5 million of its own cash. Under the program, a reward will be offered for information that leads to the arrest of the writers of computer viruses. The first two bounties announced were two $250,000 rewards for information leading to the arrest of the writers of Blaster worm and SoBig.F email viruses. Even more recently, Microsoft offered a $250,000 bounty on the creator of the MyDoom.B virus.[17]

The cases of CyberAngels and Microsoft's anti-virus reward program illustrate that while the free-rider incentive may indeed be present, it is not necessarily the strongest incentive. Other incentives such as good will, a feeling of civic duty or pride, or some notion of fairness or morality may be present as well. The key insight is that while it is appropriate for economic theory to assume a strict self-interestedness among the agents that populate its models, it is inappropriate to maintain that goods with large positive

---

[15] For more on the mission statement of the CyberAngels, see: http://www.cyberangels.org/mission/index.html.

[16] The main website of the CyberAngels program (http://www.cyberangels.org/index.html) is available in four languages.

[17] For details on this program see: http://www.microsoft.com/presspass/press/2003/nov03/11-05AntiVirusRewardsPR.asp.

spillovers will not be supplied privately in the real world based on this assumption. While theory requires the simplification that reducing motivation to a single element entails, we must keep in mind that the world in which we find ourselves is considerably more complex and involves innumerable motivations that may completely outweigh the countervailing motivation of self-interest.[18] Clearly these donations are not, at their current levels, enough to protect cyberspace in its entirety. The main point though is that, contrary to theory, they do in fact exist. As the Internet continues to grow, there is no reason to expect that these types of voluntary donations will not increase as well.

Yet another example of the private provision of cyber security through voluntary donation is open source code. Open source code has a long history in the development of the Internet. In its early stages, the Internet was a simple protocol for exchanging data. The early versions of this protocol included the file transfer protocol (FTP) and the electronic message protocol (SMTP). The subsequent development of the "Gopher" protocol allowed for directories to be depicted graphically. The hypertext transfer protocol (HTTP) and the hypertext markup language (HTML) were created in 1991 and are the foundation of the Internet as we know it today. These protocols were available to all users (i.e., open) and were used to develop many additional applications. Much of the subsequent software and applications developed were "open"—i.e., the source code and object code were available to all other users.[19] The rapid growth of the Internet has been attributed to this early openness of code (Lessig 1999, p. 103). Users could view the code of others and either improve or build upon it. In this regard, open source code can be seen as a good with significant positive externalities that is privately provided.[20] Individual users "donate" or allow for the code they developed privately to be open for all Internet users to view, copy

---

[18] Also of note is the market for "ethcial hackers" which are hired by companies to hack into their systems before "unethical hackers" can. Gartner Inc., a market research firm in Stamford, Connecticut, estimates this to be a $1.8 billion industry for the year 2002 with expected growth of 28% for the next three years. Some ethical hackers focus on one specific operating system such as eEye Digital Security (http://www. eeye.com/html/) that specializes in Microsoft Windows. In addition to assisting their clients, eEye voluntarily reports any holes in Windows to Microsoft, although they have no formal relationship, and doesn't publicly release the information on the security flaw until Microsoft develops a patch. See, Nick Wingfield, "It Takes a Hacker," *The Wall Street Journal*, March 11, 2002 and Brad Stone, "An eEye on Microsoft," Newsweek, March 22, 2004.

[19] Source code is the code that computer programmers write in. Object code is machine-readable (Lessig 1999, p. 103).

[20] Indeed, open source software would be an example of what economists call a pure public good. Once made public, it both non-excludable—all users can access it—and non-rivalrous—one users consumption of the code does not reduce the amount available for others. The notion of public goods and externalities are closely related. A public good possesses large positive externalities and a public bad large negative externalities. For more on open source code as the private provision of a public good, see James Besson, "Open Source Software: Free Provision of Complex Public Goods" available at: http://www.researchoninnovation.org/opensrc.pdf (last accessed 7/7/04).

and improve upon. Today, a mixture of open and closed code exists on the Internet. Nonetheless, open source code still plays a critical role in cyberspace and in Internet security.[21]

Open source code relates to the issue of cyber security on two fronts. On the one hand, there are specific security programs based on open source code that are publicly available for downloading by all users. To a greater extent though, security is an issue with all open source code programs. With open source programs, the underlying code is available to all—both benevolent users as well as criminals. As a result, questions of security arise for open source programs given that all users have access to the code.

There is much debate regarding the viability of open source code from a security standpoint. Critics argue that open source code provides potential criminals with the blueprints of the security system. Advocates counter that the constant peer review actually makes programs based on open source code more stable and reliable as compared to commercial code. For instance, Vincent Rijmen, an award winning developer, believes that the open nature of Linux is preferable from a security standpoint, "not only because more people can look at it, but, more importantly, because the model forces people to write more clear code, and to adhere to standards. This in turn facilitates security review."[22] In any case, clearly all users of open source code receive a large positive spillover. Specifically, they gain a large benefit from the initial availability of the code as well as from improvements made to open source code by other programmers.

Another response to critics of open source security code is that those seeking security can take existing open source security code and make minor adjustments that customize the program specifically for the user. These adjustments can be open or closed code but the foundation is available through the initial open source code that existed from the work of others.[23] Several companies now offer security packages based on open source code including Guardent (http://www.guardent.com/), Covalent (http://www.covalent.net) and Astaro Corporation (www.astaro.com), to name a few.[24]

---

[21] To support this claim, consider that the Apache system, the number-one server on the Internet, is open code as is SENDMAIL, one of the most widely used programs for forwarding email (Lessig 1999, p. 104). During the first three years of Apache system's existence, 388 developers contributed 6,092 enhancements and corrected 695 bugs (Mockus et al. 2000). This rate clearly exceeds that of commercially provided software which relies on closed code (Mockus et al 2000, Table 1).

[22] Interview with Vincent Rijman, available at: http://www.linuxsecurity.com/feature_stories/interview-aes-3.html.

[23] A survey by Franke and von Hippel (2002) found that over 19% of the firms who used the Apache system had modified the code while another 33% customized the system by adding on security modules obtained from third parties. Indeed, it is because of the open source code that add-on modules have been developed. As of January 2004, there were over 300 modules developed. See http://modules.apache.org/.

[24] The U.S. Navy also uses an open source security program, SHADOW. See http://www.techweb.com/wire/story/TWB19981008S0010.

In addition to the benefits discussed above, security based on open source code has the additional benefit of being lower cost, as the user does not have to pay licensing fees.

Open source software is clearly an example of a good with significant spillover effects that is nonetheless privately provided. Once it is written and the contribution is made available or "donated" to the cyber community, all users are able to access it and benefit. Although standard economic theory predicts that such goods will fail to be produced on the unhampered market, we observe the opposite. There are several potential incentives that lead to the provision of open source code. One is that those who make their code public benefit from others who improve on their initial code. There is also the potential for fame within the programming sub-culture.[25] While anyone can contribute by posting code, the reputation or fame mechanism serves as a sorting device for other users. Fame provides enough of a benefit for these programmers to provide code to the rest of the cyber community. Open source code has allowed for the continual innovation and development of new applications and programs. While there are both potential costs and benefits to using open source code, it is a clear example of a private solution to the production of a good with significant spillover effects.

4.2 The Private Provision of Internet Security via By-Product

The free-rider problem can also be overcome if it is possible to tie a by-product to the externality. Television commercials are one example of this mechanism. Financing for commercial television comes mostly from private sponsors who pay for advertising to be aired during television programming. The by-product of the externality—here the television program—is the captive viewing audience. We see many analogous examples in cyberspace.

Many Internet applications offer security features free of charge, but tie in other features allowing providers to earn a profit. For instance, most free email applications (e.g., Hotmail, Yahoo mail, etc.) contain virus scan features that check incoming/outgoing emails and attachments for viruses. In order to benefit from these security features, users must register with the provider. The providers make profits through advertisers who target the users of the application. For instance, Hotmail members receive emails from sellers in their inbox. Yahoo offers a pop-up blocker free of charge, but the user must have an account and a companion bar is placed at the top of the Internet browser, providing links to other Yahoo services connected to advertisers.

In order to increase the number of users and garner profits from advertisers, these providers must make their products attractive. Because part of

[25] On the issue of fame, see the *Economist* article, "An Open and Shut Case," May 10, 2001.

the attractiveness is security, producers offer this feature. Once again, security increases the value of cyberspace for all users. In this context, cyber security is privately provided because the captive audience has a value that advertisers are willing to pay for. As with advertisers on television, advertisers on the Internet are willing to pay to reach as many people as possible.

In a similar vein, some providers of security software offer one version of their application free of charge, but charge the user for an upgrade. They provide a basic level of security with no charge but include in the package advertisements for the premium versions of their software. A good example of this is Ad-Aware which is developed and distributed by Lavasoft.[26]

The Ad-Aware software erases spyware from a user's computer. Spyware is programming that is tied into downloads—often the user is unaware that it is associated with the download. Once downloaded, spyware uses the available Internet connection to send information from the user's computer to the spyware company. One form of spyware - commercial spyware - tracks the websites visited by the user. Commercial spyware is often associated with adware, which uses the information to send pop-up advertisements that fit with the information related to the user. A second and more dangerous form of spyware - domestic spyware - tracks and captures the activities of the user via their keystrokes. This form is analogous to a wiretap and sensitive information such as passwords and private email and instant messenger conversations are at risk (Mitnick and Simon 2002, p. 203-8). Ad-Aware scans the user's computer memory, registry and hard drives for commercial spyware components and allows for their safe removal.

While the basic version is free of charge, Lavasoft offers two other versions—Ad-Aware Plus and Ad-Aware Professional for a charge. These versions contain more features than the basic version. In this context, the positive externality is the free security software and the by-product is the captive audience that downloads the free version. The captive audience is enough in terms of potential profitability for Lavasoft to provide the basic version free of charge. There are other examples as well. For instance the basic version of ZoneAlarm, a firewall software product, is free of charge to any user. Similar to Ad-Aware, ZoneAlarm charges customers for more advanced versions of its software.

Internet security provided by most firms also falls into this category. Most businesses that utilize cyberspace invest resources in cyber security. It is in their interest to do so for several reasons. For one, as noted in the Introduction, breaches are costly. In economic terms firms should be willing to invest in cyber security up to the point where the costs are equal to the benefits. Moreover, consumers demand that their information and transactions be protected. In order to attract customers, online businesses

---

[26] For more on Lavasoft see: http://www.lavasoft.de/default.shtml.en.

must offer certain security measures. In the absence of minimal levels of security, we would expect the customer base of online firms to decrease significantly. The by-product of the externality—here cyber security, are the customers that are willing to offer the firm business. The key point is that these customers are willing to do so only if a secure environment is provided. The secure environment has significant spillover effects to parties outside the immediate transaction. Despite the fact that firms do not capture all of the benefits, they offer security because they secure enough monetary benefits through their direct interaction with customers providing them with business.

Consider, for instance, the case of formal online payment mechanisms such as PayPal and BidPay. These services allow buyers to make secure payments, via credit card or through their bank account, to sellers. Given that they are dealing with sensitive information regarding their customers, security is of the utmost importance. Given this, PayPal and BidPay make use of encryption technology to protect the information of their customers—both buyers and sellers.[27] The services offered by these middlemen who provide payment mechanisms do provide significant positive externalities. As discussed earlier, the Internet is a network externality which increases in value the more others are connected and able to participate online. By providing the potential for secure transactions, these services increase the value of the Internet to other users by lowering transaction costs.[28] They provide security despite the fact that there are positive spillovers that they do not capture because it is the only way to maintain and increase their customer base and profitability.

Understanding that private businesses have an incentive to invest in Internet security is critical because the greatest fear for government agencies is that terrorists will breach the networks of critical industries and have significant negative spillovers on the economy as a whole. Given this, the key issue is whether these businesses will under-invest in security given that they don't internalize *all* of the benefits. Granted, they produce some cyber security as the numerous examples above illustrate. But the argument is that because of the externality, they will fail to produce the optimal amount. To remedy the problem, government often intervenes to either produce the good altogether or regulate the private production of the good attempting to overcome the market failure. We now turn to a discussion of the potential limitations of government's ability to effectively do this.

---

[27] Additionally, many of these payment applications offer insurance protection as well. For instance, PayPal has a "Seller Protection Policy," which protects sellers against fraudulent buyers, as well as a "Buyer Protection Program," which provides $500 of insurance coverage against fraud at no additional cost to the buyer.

[28] It is estimated that PayPal has 14 million subscribers. Source: http://www.wilsonweb.com/wct5/paypal_assess.htm.

5. THE THEORY OF GOVERNMENT FAILURE

As discussed in Section 3, the theoretical rendering of externalities concludes that the privately optimal level will fall short of the socially optimal level. Government is often called upon to make up the shortfall through intervention and regulation. Policymakers calling for government to actively play a role in the provision of cyber security illustrates this. Fundamentally, their claims are grounded in the belief that the market will either altogether fail to supply Internet security or, where it does, will undersupply security. In many cases, theoretical academic research also concludes that the market will undersupply key elements of cyber security. For instance, the research of Gordon et al. (2003) concludes that security information sharing between firms will be sub-optimal due to the free-rider problem. One possibility, they conclude, is for government to subsidize the sharing of information between firms (2003, p. 479-80). However, just as economic theory suggests that there is the potential for market failures, it also indicates that there is a potential for government failures as well. Just as it is important to understand why the market may only imperfectly provide cyber security, it is equally important to appreciate why the government may fail to supply the efficient level. Therefore, considering the potential benefits of government involvement along with the related limitations and costs is of the utmost importance for an accurate analysis.

One potential option is for government to produce the good, either in conjunction with the market or instead of the market. The difficulty with this option stems from the issue of calculation. It must be realized that goods with significant externalities, just like all other goods, are not produced in one lump, but rather in marginal units. In the market, the profit and loss mechanism serves as the guide for determining the optimal number of units to produce. Admittedly, it is true that where externalities exist, the profit and loss mechanism may not produce the same level as compared to a situation where externalities are fully internalized.

With government, however, the profit and loss mechanism is not just imperfect in the face of externalities—it is necessarily completely absent. This means that the state will never have any way of effectively determining the optimal supply of the good in question. In short, there is no way for any external party to calculate the optimal social stock of cyber security and, hence, to claim that it is over or undersupplied. To do so would require complete and perfect knowledge that one cannot possibly possess. It may be true that private businesses have difficulties calculating the exact return on investment (ROI) for security-related expenditures, but this will be even more difficult for government agents acting outside the profit and loss mechanism. Given this realization, while it is indeed possible that the government may provide more cyber security as compared to the private market, there is no reason to believe that it will provide the socially optimal amount. From an efficiency standpoint, it is not simply a question of the

total dollar value of resources invested, but rather the allocation of those resources to their most highly valued uses. Calculating the optimal level of goods is far simpler using a theoretical model with simplified assumptions than it is in reality.

Yet another option is that government can choose to regulate the market production of the good in the hopes of internalizing the externalities. In the case of cyber security, this may involve regulating the specifications of hardware and software in order to internalize the externalities in the hopes of aligning costs and benefits and achieving the socially optimal outcome. The main problem with this solution is the difficulty in gathering the relevant information necessary to effectively regulate.

For instance, the regulators must know and be able to assign the damage done by insecurities in cyberspace. Given the interconnectedness of cyberspace, these vulnerabilities may be difficult to track and assign to a specific user. Given that the regulator aims to align costs and benefits, in addition to knowing the damage done by vulnerabilities, he must also possess the relevant information regarding the costs of remedying the situation. This information will be difficult to obtain. It is in the interest of each user with vulnerabilities to convince regulators that the damage they are causing is lower than the cheapest means of correcting the problem. In other words, it is in their interest to convince regulators that the costs of prevention are greater than the benefits.

Yet another issue deals with the policy flexibility of regulators in the context of cyberspace, and more specifically with what legal scholar Michael Froomkin refers to as "regulatory arbitrage" (1997). Because cyberspace connects users across national boundaries, Froomkin argues it will become increasingly difficult for any one nation to enforce its domestic rules. In other words, users can engage in regulatory arbitrage and evade domestic laws by engaging with users outside their national borders who are not subject to the same laws.

Admittedly, government can take steps to impede the use and effectiveness of cyberspace. For instance, China has attempted to set up an Internet censorship system known as "The Great Firewall of China." While this effort has raised the cost of engaging in cyberspace, users have found ways around the barrier largely by using servers outside the firewall. In sum, one potential limitation on the government provision of cyber security deals with constraints on flexibility stemming directly from the very nature and magnitude of cyberspace.

As was illustrated by the quotes from policymakers in earlier sections of this paper, one of the criticisms of the market provision of cyber security is that there is a lack of incentive to consider the national interest. However, it is critical to realize that there are perverse incentives in the political realm as well. As Ranum describes his research on the topic of homeland security: "I came face to face with the realization that there are gigantic bureaucracies that exist primarily for the sole purpose of prolonging their

existence, that the very structure of bureaucracy rewards inefficiency and encourages territorialism and turf war" (2004, p. xv). Indeed, as public choice theory informs us, political agents face a set of incentives that are in many times misaligned with the interests of the populace.[29] The implications are clear: the presence of misaligned incentives in the market does not give one license to jump to the conclusion that government intervention is preferable. Instead, a complete consideration of potential government intervention must involve a consideration of the incentives faced by political agents and the implications of those incentives for the provision of cyber security.

A final constraint on government regulation of cyber security is the potential for limited control of the response to policies by the private market. When considering a potential regulation, due to genuine structural ignorance, only some of the potential costs, benefits and impact on incentives can be known *ex ante*. Once a regulation is passed, it creates a new set of incentives for both political and economic agents. In many cases, the outcomes that the new policy generates will not be aligned with the initial aim. This will leave government officials in a situation where they can either retract the original policy or pass additional policies to attempt to solve the unintended outcomes. This limitation may be potentially magnified in the case of cyberspace for the reasons addressed above—namely the continually changing cyber environment.

6. POLICY IMPLICATIONS: INTERNALIZING EXTERNALITIES

We have discussed the potential limitations in both the market and government spheres in the context of cyber security. Fortunately, in addition to providing insight into the limitations of the market and government, economics also provides specific guidelines for policymakers. From an economic standpoint, the market provision of goods and services is preferable to government provision. This is due to the fact that the profit/loss mechanism inherent in the market setting guides economic actors in allocating resources to their most highly valued uses. In the context of cyber security this means that policies should be aimed at taking advantage of the desirable consequences of the market. It is only through the market process that the "right" amount of cyber security can be produced. More specifically, policy should be focused on internalizing the externalities while maintaining the allocative function of the profit/loss mechanism. Recently, several alternative courses of action have been discussed that potentially serve to internalize externalities. In theory, these potential solutions allow the desirable aspects of the market to function while overcoming the potential pitfalls of direct government regulation.

---

[29] For more on the public choice research program, see Buchanan (2003).

One potential solution is the assignment of property rights. Well-established property rights result in markets incorporating the presence of externalities. Along these lines, one solution that has been proposed by Camp and Wolfram (2000) is the assignment of property rights to cyber vulnerabilities. This solution is similar to proposals for tradable pollution permits. Camp and Wolfram not only provide a taxonomy of vulnerabilities but also propose a means of assigning property rights. They propose that each machine would receive a certain number of vulnerability credits. Processing power is suggested as a measure of how many machines, and therefore how many credits, are to be received.

The authors suggest three potential governance mechanisms to oversee this process: the federal government, the creation of a corporation similar to The Internet Corporation for the Assignment of Names and Numbers (ICANN), or the licensing of companies in the business of creating processing power who would oversee the creation and distribution of credits. Users with vulnerabilities and no credits would have a specific time period to fix the exposure and would additionally have to make a payment to the entity that discovered the vulnerability. As a result, one could envision entrepreneurial users who are in the business of discovering vulnerabilities and profiting from these payments. By defining property rights, the full cost of these vulnerabilities would fall on the owners of the insecure machines.

Given this proposal, one must recognize that there are some potential information problems on the part of regulators, as discussed in Section 5, regarding the specifics of the permits. For instance, regulators will not know the right amount of vulnerability credits to assign in order to get the optimal level of vulnerability. Further, there is the potential for bureaucratic barriers to establishing and maintaining the credit system, especially if it is governed by a government agency. This may limit the effectiveness of this remedy.

Another potential market solution is the continued growth of the already existing cyber insurance market. In addition to traditional insurance coverage, an increasing number of insurance companies are offering coverage for cyber breaches.[30] These insurance policies include coverage against damage related to hack attacks, viruses, network downtime, identity theft and the misuse of proprietary data and information. Cyber insurance is potentially beneficial on several fronts.

For one, there is an internal pressure on companies to maintain a level of security that minimizes their premiums. Insurance companies will develop standards that firms are required to meet. Given that this is a relatively new market, there is no reason to expect that it will not continue to

---

[30] The Insurance Information Institute estimates that cyber insurance could generate $2.5 billion in annual premiums by 2005. Source: Samuel Greengard, "The Real Cost of Cybersecurity," *Business Finance*, April 2003, pp. 52-55. Available at: http://www.businessfinancemag.com/magazine/archives/article.html?articleID=13957&pg=1 (last accessed 6/8/04).

grow as better actuarial data is collected and insurance companies gain a better understanding of how IT systems operate.

There is currently debate about what role the government should take in the cyber insurance market. Some argue that the market should be left to its own devices with market-determined premiums accurately reflecting the risks. Others argue that the government should guarantee cyber insurance and/or put a cap on the insurance policies.[31] Although we avoid engaging in an analysis of this issue, the economic principles discussed in previous sections, specifically issues of economic calculation, can add much insight into this debate regarding the ability of government to effectively regulate this market.

Closely connected to the subject of cyber insurance, yet another potential means of internalizing externalities is extending liability to software authors and/or system operators. In the absence of being held liable, it is argued that these parties have a weak incentive to provide security because they do not incur the full costs of their failure to do so. Fisk (2002) concludes that it would be more effective to extend product liability to system operators as compared to software developers. One reason for this conclusion is that the existence and importance of open source software poses problems for making developers liable. Those that contribute open source software receive no income to offset potential liabilities. Purchasing cyber insurance would be one way of protecting against liability, but would also raise the cost of contributing open source code, so we would expect a decrease in the amount of open source software produced.

Fisk concludes that holding system owners liable is more reasonable and advocates an insurance system where liability for cyber accidents is "expected and accepted without stigma" (2002, p. 4). Similar to the automobile industry, system operators would be required to carry insurance against unexpected events. Fisk contends that the insurance industry would have similar beneficial effects on cyber security to those discussed above. He also envisions the creation of an Underwriters Laboratory that would certify software as secure and create an environment that encouraged effective cyber security.

We have not provided an exhaustive list of all possible courses of action. Instead, our aim here has been to highlight several potential courses of action for policymakers to consider. It is not our goal to endorse any one of these alternatives as being better than the others. Instead, our purpose is to emphasize that whatever course of action policymakers choose, their focus should be on ensuring that the desirable aspects of the market are able to function effectively.

---

[31] The Terrorism Risk Insurance Act, signed in November of 2002, created a three-year federal program that backs insurance companies in addition to guaranteeing that certain terrorist-related claims will be paid.

7. CONCLUSION

Without a doubt, the issue of cyber security will remain an important policy issue in the future. We have offered some insight into this issue from an economic perspective. In addition to the policy implications discussed above, we can put forth several general guiding principles:

1. Economics is a critical aspect of cyber security—Our main argument is that economics has been neglected in the policy debate regarding the most effective means of securing cyberspace. The basic concepts discussed in this paper can offer key insights into the best course of action. Admittedly, obtaining the necessary information to utilize these concepts will not always be easy. Nonetheless it is clear that neglecting the economic aspects of the issue will lead to incomplete and incorrect analyses.
2. National cyber security must be "demystified"—A key aspect of the cyber security issue is understanding the interconnectedness of the cyber environment. Given the interconnected nature of cyber space, the term "national security," in the context of cyber space, is simply the aggregate of individual Internet users whether for personal or business use. One must be careful not to think of "national security" as something that would fail to exist in the absence of government. As Schneier points out, we need to "demystify" Internet security (2003, p. 271). Security is all around us in our daily lives in a multitude of ways and individuals take steps to secure their property, information and transactions. Cyber space is no different
3. Cyber security policy should rely on the market to the greatest extent possible—Economic analysis provides key insights into limitations in both the market and government settings. Given that the market provision of goods and services is preferable to government provision, from an economic standpoint, policy should aim to internalize externalities while maintaining the effectiveness of the profit/loss mechanism in efficiently allocating resources.

REFERENCES

Anderson, Ross. 2001. Why Information Security is Hard: An Economic Perspective. Proceedings of the 17th Annual Computer Security Applications Conference, 358 - 365.

Anderson, Ross. 2002. Maybe we spend too much? Workshop of Economics and Information Security, University of California, Berkeley, May 16-17. http://www.cl.cam.ac.uk/users/rja14/econws/37.txt.

Buchanan, James M. 2003. Public Choice: The Origins and Development of a Research Program. Center for the Study of Public Choice, George Mason University, Fairfax, VA.

Camp, L. Jean, and Stephen Lewis, eds. 2004. *Economics of Information Security*. Kluwer Academic Publishers.

Camp, L. Jean, and Catherine Wolfram. 2000. Pricing Security. *Proceedings of the CERT Information Survivability Workshop,* Boston, MA, 31-39.

Cheswick, William R., and Steven M. Bellowing. 1994. *Firewalls and Internet Security: Repelling the Wily Hacker*. Reading, MA: Addison Wesley.

Computer Security Institute and Federal Bureau of Investigation. 2004. *CSI/FBI Computer Crime and Security Survey*. http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2004.pdf.

Fisk, Mike. 2002. Causes & Remedies for Social Acceptance of Network Insecurity. Workshop of Economics and Information Security, University of California, Berkeley, May 16-17. http://www.cl.cam.ac.uk/users/rja14/econws/35.pdf.

Froomkin, Michael. 1997. The Internet as a Source of Regulatory Arbitrage. In *Borders in Cyberspace*, edited by Brian Kahin and Charles Nesson. Massachusetts: MIT Press, 129-163.

Gordon, Lawrence A., Martin P. Loeb, and William Lucyshyn. 2003. Sharing Information on computer systems security: An Economic Analysis. *Journal of Accounting and Public Policy* 22: 461-485.

Johnson, David R., and David G. Post. 1996a. And how shall the Net be governed? A meditation on the relative virtues of decentralized, emergent law. http://www.cli.org/emdraft.html.

Johnson, David R., and David G. Post. 1996b. Law and borders—the rise of law in cyberspace. *Stanford Law Review* 48: 1367-1405.

Lessig, Lawrence. 1999. *Code and other laws of cyberspace*. New York: Basic Books.

McMillan, John. 2002. *Reinventing the Bazaar*. New York: W.W. Norton and Company.

Mitnick, Kevin D., and William L. Simon. 2002. *The Art of Deception: Controlling the Human Element of Security*. Indiana: Wiley Publishing, Inc.

Mockus, Audris, Roy T. Fielding, and James Herbsleb. 2000. A Case Study of Open Source Software Development: The Apache Server. *Proceedings of the 22nd international conference on Software engineering (ICSE2000)*, 263-272.

Pollitt, Mark M. 1997. Cyberterrorism: Fact or Fancy? *Proceedings of the 20th National Information Systems Security Conference*. October: 285-89. http://www.cs.georgetown.edu/~denning/infosec/pollitt.html.

PricewaterhouseCoopers. 2000. *Security Benchmarking Service/InformationWeek's 2000 Global Information Security Survey.* Summary available at: http://www.pwcglobal.com/extweb/ncpressrelease.nsf/docid/7ABBA8E73B1E901D8525693500548A34.

Ranum, Marcus J. 2004. *The Myth of Homeland Security.* Indianapolis: Wiley Publishing, Inc.

Reidenberg, Joel R. 1996. Governing networks and cyberspace rule-making. *Emory Law Journal* 45: 911-926.

Schneier, Bruce. 2002. Computer Security: It's the Economics, Stupid. Workshop of Economics and Information Security, University of California, Berkeley, May 16-17. http://www.cl.cam.ac.uk/users/rjal4/econws/18.doc.

Schneier, Bruce. 2003. *Beyond Fear: Thinking Sensibly About Security in an Uncertain World.* New York: Copernicus Books.

SecurityStats.com. 2004. *Virus Statistics*, January 16, 2004. http://www.securitystats.com.

Verton, Dan. 2003. *Black Ice: The Invisible Threat of Cyber-Terrorism.* New York: McGraw-Hill.

IS CYBERSECURITY A PUBLIC GOOD? EVIDENCE FROM THE FINANCIAL SERVICES INDUSTRY

Benjamin Powell, Ph.D.[*]

The September 11, 2001, terrorist attacks on the United States heightened concerns about vulnerabilities to future attacks. One new area of concern is cyberterrorism: the possibility of terrorists using computers to attack our critical infrastructure electronically. The government has made efforts to better secure its own computer networks against terrorist hacking in the Pentagon, FBI, and other government agencies. Increasingly, however, the government has been concerned that the private sector is vulnerable to cyberterrorism. The private sector owns approximately 85 percent of the critical infrastructure in the U.S. (Deloitte 2004, p. 15). The government is concerned that a cyber attack on dams, trains, electrical grids, pipeline pumps, communications networks, or the financial services industry could cause significant physical or economic damage to the U.S. The policy question being asked is whether private businesses, when left to their own devices, provide enough cybersecurity or if some form of government involvement is justified.

Some policy makers are skeptical of the market's ability to provide enough cybersecurity. In a speech to the National Academy Conference on "Partnering Against Terrorism," Congressman Boehlert said, "Here is a case in which the government can't carry out its most basic mission—providing security—without the cooperation of the private sector. And here is a case in which the private sector will quickly need a range of products on which the market has never before put a premium—the classic market failure that calls out for government involvement" (Boehlert 2002). Similarly, in a February 2004 speech, Richard Clarke, the former counterterrorism czar for Bill Clinton and George W. Bush, said, "Last year was a market failure in cybersecurity, and 2004 doesn't look much better. In general, Internet Service Providers (ISPs) do nothing about security. The market

* Assistant Professor of Economics, San Jose State University; Director, Center for Entrepreneurial Innovation at The Independent Institute. E-mail: Benjamin.Powell@sjsu.edu. The author thanks Chris Cardiff, Anthony Gregory, and David Skarbek for excellent research assistance. Helpful comments from the participants at George Mason's Critical Infrastructure Project are gratefully acknowledged. The usual disclaimer applies.

isn't forcing the ISPs to do anything about security" (Ricadela 2004). Calls for government regulation of cybersecurity have accompanied these proclamations of "market failure." In 2003 the federal government published The National Strategy to Secure Cyberspace. The plan's three main goals are to prevent cyber attacks against America's critical infrastructure, reduce national vulnerability to cyber attacks, and minimize damage and recovery time from cyber attacks that do occur. The government needs to better consider the economics of cybersecurity before moving forward with any policies. Specifically, the government needs to examine if the market truly "fails" to provide the correct amount of cybersecurity. The government should also consider if it will be able to improve the situation or if "government failure" could be as pervasive as "market failure."

This paper proceeds by first examining the economics of cybersecurity and its applicability to the defense against cyberterrorism. The financial services industry is regarded as one area of critical infrastructure requiring protection from cyberterrorism; it is therefore examined as a case study in section II to determine if the market is indeed failing. Section III considers the problems confronting government cybersecurity policy while focusing on the financial services industry and examines the potential for government failure. Section IV concludes.

I. ECONOMICS OF CYBERSECURITY

Economists generally assume markets are relatively efficient. In the realm of cybersecurity, however, markets are often assumed to fail. At least one researcher (Anderson 2001) has pointed out that the incentives of the so-called "experts" in the area may cause this. Producers of information security technology may benefit financially if they can scare more people into purchasing security products. Similarly, professors competing for the latest homeland security grants may face incentives to overstate the problem. Despite these potential biases, simple economic models highlighting potential market failures in the provision of cybersecurity are worth considering.

The security of the entire Internet is affected by the security measures used by all individual Internet users (Anderson 2001). Because of this, cybersecurity is often assumed to be a "public good" that will be underprovided or fail to be provided at all in the private market. When firms or individuals have a greater level of cybersecurity, their computers are less likely to be hacked into and used to launch spam or other denial of services attacks (DOS). The security of one computer owner benefits other computer users by reducing the probability that they will be attacked through the first owner's computer. However, since individuals are not generally liable for the damage caused when a hacker takes over their computer, they

do not benefit personally from the increased security.[1] Since the user with the ability to provide the security does not benefit, they will fail to provide it. Other computer owners with access to the Internet face the same incentives, and everybody is worse off than they would be if everyone provided the security that had spillover benefits for everyone else. The incentives confronting an individual user could be modeled like the prisoner's dilemma game in Figure 1.

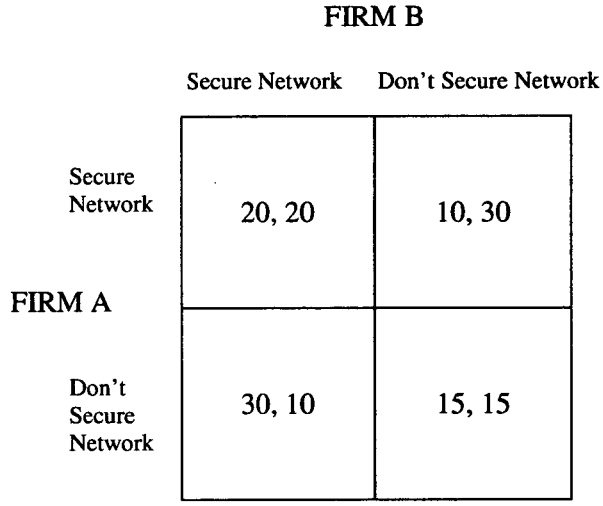


Figure 1

In this figure, "secure network" should be interpreted as a firm taking steps to prevent its computers from being used to launch attacks on other firms' computers. Thus, when one firm secures its network, the other firm receives the benefit. Since there is also some positive cost to securing their networks, neither firm has an incentive to do so. If both firms secured their networks, they would both be better off, in this case receiving a utility of "20." However, each firm only controls its own decision whether to secure its network or not. Firm B compares whether it would be better off securing its network or not depending on what A does. If firm A secures its network, B would receive 20 if it secured its own as well, but 30 if it did not, because it would still receive the benefit provided by A securing its network but would not bear the cost of securing its own. Similarly, if A does not secure its network, B would receive only 10 if it secured its own, because it would not be receiving the benefit of A's security but would be bearing the cost of securing its network. If B too did not secure its network, it would receive a higher utility of 15. Regardless of whether A does or does not

[1] Varian (2000) examines incentives under differing liability rules.

secure its network, B is better off not securing its own. The payoffs are symmetrical, so the same incentives confront firm A. The Nash equilibrium is for neither to secure its own network. This leaves them both with a utility of only 15. Both firms would clearly be better off if they could have coordinated and both secured their networks and received a utility of 20, but neither has an individual incentive to do this. Of course, with only two firms the transaction cost of bargaining to achieve the efficient outcome is fairly low, so the Coase theorem should hold and allow them to reach the efficient outcome (Coase 1960). However, in the real world these incentives face many firms and individuals. The transaction costs of bargaining between all computer users are likely high, so we would be stuck in the inefficient Nash outcome of 15, 15.

In the above analysis, all of the benefits of cybersecurity were external to the person providing the security. In reality, many of the benefits of cybersecurity accrue to the user of the security. Often the same security techniques that will secure your own private information, prevent your files from being destroyed by a virus, and prevent private financial loss are the same security techniques that benefit other computer users. Most forms of computer security create both private and public benefits. The above model highlighted why the market might fail to provide cybersecurity, but the empirical question that needs to be examined is whether the private benefits are great enough to cause individual firms and computer users to provide enough cybersecurity. If the costs of the security are high, the private benefits low, and the public benefits high, firms will underprovide cybersecurity on the market. If the costs are low and private benefits are high, firms will generally provide close to efficient levels of cybersecurity despite some positive externalities.

A word of caution is in order. In a predetermined model in which all private and public costs are known and specified in advance, it is trivial to solve the problem of finding the "optimal" level of cybersecurity and then compare what the private market provides to the theoretic optimal amount. However, it is impossible to know all the private and social costs and benefits in the real world. We know that 100 percent security is not likely to be the efficient outcome given the costs of achieving it. To observe any privately provided level of security and then deem it "market failure" because it does not conform to a predetermined optimum is unjustified. Instead, we must look at whether firms are providing security, and if so, how much, as well as whether they are increasing or decreasing their level of security.

The economic literature documents a second potential market failure in cybersecurity: the problem of information sharing and free riding. A number of papers explore this. Anderson (2001) looks at the incentives facing information sharers; Varian (2002) models the free rider problem and system reliability; Gordon et al. (2002) looks at information sharing by SB/ISOs; Gordon, Loeb, and Lucyshyn (2003) study the welfare implications of information sharing and the conditions necessary for information

sharing to increase computer security; and Schechter and Smith (2003) examine the benefits of sharing information to prevent security breaches.

The potential market failure in information sharing is a result of the incentive to free ride. The literature recognizes that if firms share information about security breaches and defenses against attacks, they can lower their security expenditures while maintaining or increasing their level of security. This sharing creates two potential problems. The first is that when a firm reports a security breach, it provides a benefit to other firms but may receive no reward itself. Thus, individual firms may fail to report breaches that would benefit others. The second potential market failure comes from the possibility of free riding on other firms' security innovations. If firms share security innovations and confront a common problem, individual firms may fail to deal with the problem because they hope they will get the benefit when another firm creates a security innovation to solve it. Because of this incentive to free ride, firms may not innovate as quickly as they should.

The key to potential market failures in information sharing is that the firm sharing the information does not benefit from sharing. This problem can be solved or at least reduced with appropriate incentive devices. Many information-sharing groups are private and can exclude non-members. Incentives for sharing would improve with the ability to kick out members suspected of holding back information (Tullock 1985). Other positive monetary incentives for sharing could also be offered. While the potential for free riding and underprovision of information sharing exists, there are benefits to be had by private groups if they can create the right incentive structure. As long as these groups are left private with the ability to make their own rules and exclude non-members, they will likely experiment to find ways to minimize the free rider problem.

Although a number of theoretic "market failures" are possible in the provision of cybersecurity, the market process may also work to solve these failures. In the next section we examine the financial services industry for evidence of market failure or success in the provision of cybersecurity.

II. FINANCIAL SERVICES INDUSTRY CASE STUDY

A cyberterrorist attack on the financial services industry, part of the "critical infrastructure" of our economy, could ripple through the entire economy. Banks, investment firms, and insurance companies all store vast amounts of important data electronically, so the economic damage that could result from a cyber attack is high.[2] We use the Deloitte Touche Toh-

[2] Policy makers sometimes seem to draw a qualitative difference between general cybersecurity and cybersecurity of "critical infrastructure." From the economic point of view, only a quantitative difference exists. Some infrastructure may be deemed "critical" because if security failed, the dollar

matsu 2003 and 2004 Global Security Surveys of the financial services industry to examine how businesses in the financial services industry are protecting themselves from cyberterrorism. The survey respondents were information security executives at major banking, insurance, and financial services firms. Of the largest one hundred firms in each sector, more than 30 percent of the largest financial services firms, 20 percent of the largest banks, and 20 percent of the largest insurers responded to the survey.

If there were a massive market failure to provide cybersecurity in the financial services industry, we would expect little investment in cybersecurity, lack of industry concern in providing it, and little use of security products. If, however, cybersecurity provides large private benefits, we might observe the opposite. If the financial services industry responds to heightened threats by increasing security staffing, increasing budgets, and using new technology, then there is reason to believe that the private benefits to security induce firms to provide it despite some publicness characteristics.

The point of this section is not to prove the "optimality" of the current level of security provided in the financial services industry, but only to emphasize the widespread investment in and use of technology to highlight the market's ability to provide security despite publicness characteristics. Testing for "market failure" by examining current provision compared to an optimum is impossible. To know the optimal level of security (or any other good in the market), we would have to know all the costs and benefits to all market participants under ever changing conditions at every point in time. But this information is not available because it is decentralized, subjective, and often tacit or inarticulate knowledge that cannot be made available to analysts (Hayek 1945). This is why the market process is a discovery procedure in which the optimal pattern of resource use is constantly evolving and being discovered anew (Hayek 1978).

In examining financial services companies, we find most do make large investments in cybersecurity. In the U.S., financial services companies spend between 6 and 7 percent of their entire information technology budgets on security (Deloitte 2004: 20). Most firms have an upper level executive dedicated to cybersecurity. Sixty one percent of respondents to the Deloitte survey in 2003 had a Chief Security Officer or a Chief Information Security Officer (p. 10). The survey summarized the financial services industry view of cybersecurity as follows: "Executives rank security as a high priority and security initiatives are seen as a good investment. Security is a business issue driven by shareholder value, customers' perception, brand and reputation protections, legal and regulatory compliance, vulnerability and sustainability" (2004, p. 15).

---

value of damages would be much greater than the losses from damages in non-"critical" infrastructure. From an economic efficiency standpoint, holding probability of breach constant, we should hope to see greater security in industries and firms where the expected dollar value of damages is higher.

Financial companies' investment in cybersecurity has translated into widespread use of many security technologies. The percent of financial firms that have fully deployed or are piloting various defenses are: 85 percent use intrusion detection/prevention systems, nearly 100 percent use anti-virus software, 40 percent use smart cards, and 20 percent use biometrics (Deloitte 2004, p. 24). These are all increases over the percent of firms using them in 2003. Other widely used technologies include public key infrastructure (30 percent), virtual private networks (70 percent), content filtering/monitoring (60 percent), and single sign-on (30 percent).

With significant budgets, widespread use of technology, and upper level executives devoted to cybersecurity in the financial services industry, we would expect that if businesses found themselves not earning enough of a private return on these investments because of publicness characteristics, they would shrink their investments and staffing. However, we observe financial firms increasing or maintaining their cybersecurity budgets. The Deloitte survey found that from 2003 to 2004, fewer than 10 percent of firms had reduced their security budget, while 25 percent of firms retained their previous security budget. More than 63 percent of firms reported a security budget increase. Of these, more than 20 percent experienced a 0 to 5 percent budget increase, just under 15 percent experienced a 5 to 10 percent budget increase, just under 15 percent of firms had a 10 to 20 percent increase, and about 13 percent of firms had a security budget increase greater than 20 percent (2004, p. 20). U.S. firms experienced the greatest budget growth (2004, p. 20). Increasing budgets have also translated into larger security staffs. Forty-seven percent of respondents reported that their IT security staffing levels had increased in the last year, while 29 percent remained unchanged and only 19 percent reported decreases (2003, p. 14).

Security executives seem confident that their cybersecurity spending is appropriate. Only 10 percent of respondents felt their organizations' spending on security was "inadequate," and when asked to characterize their organizations' investments in security, 84 percent classified it as on plan or catching up while only 8 percent felt it was falling behind (2003, p. 14).

Budgets, technology, and employees are all allocated to cybersecurity, as are upper level planning resources. Eighty-one percent of firms report that risk management is part of strategic planning, and 16 percent report that risk management is informally considered. Only 3 percent report that they have no strategy in place around risk, and no firms report that it is not considered at all (2004, p. 23). The Deloitte survey also found that "[i]n terms of respondents who have a comprehensive IT disaster recovery/business continuity plan in place the survey highlighted the following: 91 percent of respondents say that their organizations have one, 54 percent characterize themselves as 'very confident' that their backups either work or are being stored off site in accordance with policy" (2004, p. 25). Both of these numbers were increases from 2003.

All of the investment in cybersecurity translates into confidence among many in the financial services industry that they are up to the task of providing cybersecurity. When asked about their organizations' readiness to face cybersecurity threats, 50 percent of respondents stated that their organization was either well skilled and had the competency to respond or that staff supplementation or outsourcing was being used to gain that competency. Another 30 percent recognized that they were missing some skills but said they were adequately closing the gap. Fewer than 3 percent of organizations felt that they were missing skills and had large gaps (2004, p. 19). The survey concluded, "The majority of respondents are confident that their networks are protected from cyber attacks (e.g. DOS attack, malicious code, sabotage, etc.)" (2004, p. 23).

Non-survey evidence also suggests that businesses are receiving private benefits and possibly even overproviding cybersecurity. Campbell et al. (2003) found that security breaches can decrease the market value of firms by a statistically significant level. Soo Woo (2000) estimates that firms actually overinvest in cybersecurity. He finds a return on investment in security of around 20 percent, which is lower than the 30 percent return on investment required for most information technology investments at the time of his study.

A final concern of market failure that includes but is not limited to the financial services industry is that knowledge of damaging viruses is not communicated to those at risk of attack. However, since many people could benefit from an advanced warning, a major private incentive exists to provide it. In fact, the media, through traditional print, television, and radio stories, as well as Internet news sites, frequently warn about viruses. Since viewers and listeners value the information, providing the information increases ratings or circulation. A cursory search of Lexis Nexis for articles in major U.S. news sources from October of 2003 to October of 2004 for the words "computer virus, computer hacker, Mydoom, Sasser" turned up 755 articles. Surely many other articles were published in other forms of media and smaller circulation papers.

If cybersecurity were a purely public good, we would not see the private sector devoting so many dollars, employees, and planning resources or employing so many technologies to provide cybersecurity. There must be enough of a private return to cybersecurity to cause firms to invest so much in it. If the publicness characteristics of cybersecurity were very troubling, we would not likely see the industry continue to devote more resources to security. In general, firms do not appear to be free riding or holding off for other companies to innovate. In fact, the Deloitte survey reports that "US respondents felt that their competitors had no relevance to the way they operated or spent their money" (2004, p. 10).

The market is providing cybersecurity in the financial services industry; a complete "market failure" has not occurred. The policy question of

whether we should expect government failure if it tried to provide cybersecurity remains.

III. GOVERNMENT FAILURE AND THE FINANCIAL INDUSTRY

If there are public benefits that firms do not take into account, then the possibility remains that the government could be needed to provide the difference between the optimal level of cybersecurity and the level the private sector voluntarily provides.

It is impossible to determine the optimal level of cybersecurity and then compare it to what the private market has provided, because public goods are not bought and sold on the market. Clearly cybersecurity has costs, and making cyberspace 100 percent secure is unlikely to be optimal. As one survey respondent put it, "There is no such thing as 100 percent security. Security is not only a technology issue but a management issue as well" (2003, p. 17). Governments are simply not in a position to calculate the optimal number of resources that should be devoted to cybersecurity.

An inefficient level of cybersecurity can occur with either under- or overprovision. The market is often accused of underproviding security, but overprovision, in which security spending exceeds the expected value of losses from breaches, is likely to occur when government regulators determine the level of security. Even with the efficient level of cybersecurity, some costly breaches will occur. These will cause public relations problems for the bureaucrats in charge of regulating security, so they will likely seek to minimize breaches even if it means forcing firms to overspend on security. The costs of preventing breaches will be slower innovation rates and more expensive products and services, but these costs are largely unnoticed by the public, while costly breaches are not. The incentive problem is much the same as the one facing regulators at the FDA, which has been long recognized in the economics literature (Peltzman 1973, 1974).

Government regulators will also lack the information necessary for error correction that market participants possess. When an individual firm provides too much security, it gets feedback from the market in its profit and loss statements. It can compare itself to other firms making different decisions and see that security provisions should be reduced. When regulators mandate security provision levels for all firms, this type of error correction information is not available, since the innovative and competitive process of different companies providing different levels of security is stifled. Of course, some feedback is available if regulations are far too costly; an entire industry could go bankrupt, for instance. This informational feedback is not as precise as what a decentralized, competitive market provides, so overprovision is less likely to be detected.

The problem of calculating the efficient level of security, the incentive to overregulate, and the lack of precise feedback all make government failure likely. But even if we look at specific instances of how the market

might fail to provide enough security in the financial services industry, we find that U.S. policy is unlikely to be able to fix these problems.

Former homeland security czar Tom Ridge stated the problem by saying, "Anywhere there is a computer . . . whether in a corporate building, a home office or a dorm room . . . if that computer isn't secure, it represents a weak link. Because it only takes one vulnerable system to start a chain reaction that can lead to devastating results" (Ridge 2003). If his statement is true and literally any unsecured computer poses a threat, then U.S. policymakers cannot correct the public good problem of cybersecurity. For U.S. policy to be effective, the externality would have to be external to individual firms and users but internal to the United States. However, the Internet spans national boundaries and there are millions of computer users overseas. Computers in foreign countries can be used to launch attacks on U.S. critical infrastructure as easily as computers within the U.S. Since it is neither practical nor desirable to cut off all U.S. computer users from the world's Internet, U.S. policy could not possibly hope to secure cyberspace in the U.S. if an externality between all computer users exists.

When we observe the activities of major financial firms worldwide, we find that U.S. firms are already providing greater levels of cybersecurity than foreign firms. The Deloitte survey found that

> [w]ith the largest security staff and the greatest number of financial institutions with security strategies, it is not surprising that the U.S. reported that they were likely spending more on security than any other part of the world, given the events of the last few years. They also felt that they were prepared to take higher risks and be the leaders in adopting new forms of technology. This is a similar finding to last year, when US respondents felt that their competitors had no relevance to the way they operated or spent their money (2004, p. 10).

In 2003, the survey similarly found that U.S. firms are

> early adopters of technology, and characterize the level of risk that their organizations strive to achieve as "effective and efficient." Respondents from the United States show the highest level of BCP/DRP development, maintenance and testing over the past 12 months, which comes as no surprise given the events of September 11, 2001 (2003, p. 9).[3]

Any U.S. policy requiring greater cybersecurity from financial firms in the U.S. would likely have little impact on the industry's protection from cyberterrorism launched through third-party computers. Since U.S. firms are already providing higher levels of cybersecurity than foreign firms, any cyber attack launched on the financial services industry that first requires breaching an individual firm's security before being launched on other firms would likely come from outside of U.S. borders. Protection for individual

---

[3] BCP stands for Business Community Plan; DRP stands for Disaster Recovery Plan.

firms against such an attack once launched from inside the financial services industry is almost certainly a private good already being provided.

Even if most of the relevant externalities between firms were caused by low cybersecurity in the U.S., it is not clear that government policy could fix one of the major sources of weakness. Many breaches that threaten companies come not from technical problems or lack of investment but from simple human carelessness. As one survey respondent said, "The behavioral aspects are as worrying as the technical aspects. Everyone has to understand that it is their personal responsibility to manage risk and assets" (2003, p. 18). More specifically, another respondent even claimed that humans are *the* weakest link: "We feel that the biggest threat to us is security awareness, or lack of it. One person who opens a virus-laden attachment can cause a lot of damage. People are the weakest link. Technology can only help reduce risks to a point" (2003, p. 12). In 2004 yet another respondent claimed, "Lack of internal security awareness is still one of our biggest threats. Technology can reduce risks to a point but it is people who are the weakest link" (2004, p. 17). Direct regulation is unlikely to be able to address lax behavior.

Although the focus of this paper has been on the market's ability to provide cybersecurity and the potential that government failure could be worse than market failure, other reforms with less potential for government failure can be considered. If spillover costs of lax security between computer users are significant, legal reform to internalize the externalities could be pursued. This could involve reforming negligence standards so that computer owners could be held liable when their computer is hacked into and used to attack others. This type of reform would better address lax behavior by creating proper incentives. It would leave the market's discovery procedure in place. It would not require government to calculate the optimal level of security or give officials the incentive to over-regulate. Interventions other than legal reforms to internalize externalities would make government failure likely.

Even if the market underprovides cybersecurity, direct government regulations are unlikely to help achieve the optimal level. Government regulators have no way to know the optimal level of security. Given their incentives, they would likely force companies to invest too much and would lack the feedback mechanism to force them to revise their judgments. Most of the relevant externalities that need to be corrected exist outside of the U.S. When attempting to eliminate security breaches, policy is unlikely to directly impact one of the greatest sources of risk, lax individual behavior.

IV. CONCLUSION

Cyberterrorism against private critical infrastructure is not a problem that should be considered separately from ordinary private cybersecurity.

As Green (2002) wrote, "There is no such thing as cyberterrorism—no instance of anyone ever having been killed by a terrorist (or anyone else) using a computer. Nor is there compelling evidence that al Qaeda or any other terrorist organization has resorted to computers for any sort of serious destructive activity." Even Richard Clarke, Bush's Cybersecurity Czar, admitted, "To date, we've never seen any of the officially designated terrorist groups engage in a cyberattack against us." (Green 2002).

Green observes that this "is not to say that cybersecurity isn't a serious problem—it's just not one that involves terrorists . . . the real danger is from the criminals and other hackers who did $15 billion in damage to the global economy last year using viruses, worms, and other readily available tools" (2002). This is consistent with how the financial services industry perceived cybersecurity. Fewer than five percent of respondents ranked cyberterrorism as a high threat, but viruses and worms were overwhelmingly ranked the greatest threat with more than 70 percent of respondents giving them the highest threat rating (2004, p. 22).

Cyberterrorism against private critical infrastructure is not a problem that requires special government attention. According to the evidence examined here, the government should not be concerned with any general market failure in the provision of cybersecurity. While some aspects of cybersecurity have certain "publicness characteristics," we find many ways in which private orderings in the market provide security despite theoretical problems. Examining the financial services industry, part of the critical infrastructure of our economy, we find no evidence of a pervasive market failure to provide cybersecurity. Instead, we find widespread use of many technologies, increasing budgets, and innovation in adopting new technology. When compared to firms in other countries, financial firms in the U.S. are early adopters and generally better prepared for cyber attacks than foreign competitors. Since any externality created by unsecured computers is not limited by national boundaries, it is unlikely that U.S. policy could correct for such an externality anyway. Cybersecurity is being provided in the private sector, and it is best left free of cumbersome government regulations that may prevent private voluntary orderings from continuing to innovate to secure cyberspace.

REFERENCES

Anderson, Ross (2001) "Unsettling Parallels Between Security and the Environment." http://www.sims.berkeley.edu/resources/affiliates/workshops/econsecurity/econws/37.txt.

Anderson, Ross (2001) "Why Information Security is Hard—An Economic Perspective." In: Proceedings of the 17th Annual Computer Security Applications Conference, New Orleans, LA.

Boehlert, Sherwood (2002) Speech to National Academy Conference on Partnering Against Terrorism. October 3. Accessed Sept. 8, 2004. http://www.house.gov/boehlert/nationalacademyspeech1003a.htm.

Campbell, K; Gordon, Lawrence; and Loeb, Martin (2003) "The Economic Cost of Publicly Announced Information Security Breaches: Empirical Evidence from the Stock Market." *Journal of Computer Security*. Vol. 11, No. 3, pp. 431–448.

Coase, Ronald (1960) "The Problem of Social Cost." *Journal of Law and Economics*. Vol. 3 pp.1–44.

Deloitte, Touche, and Tohmatsu (2003) 2003 *Global Security Survey*. http://www.deloitte.com/gfsi.

Deloitte, Touche and Tohmatsu (2004) 2004 *Global Security Survey*. http://www.deloitte.com/gfsi.

Gordon, Lawrence; Loeb, Martin; and Lucyshyn, William (2002) "An Economics Perspective on the Sharing of Information Related to Security Breaches: Concepts and Empirical Evidence." In: Proceedings of the First Workshop on Economics and Information Security, May 16–17, University of California, Berkeley.

Gordon, Lawrence, Martin Loeb, and William Lucyshyn. 2003. Sharing Information on Computer Systems Security: An Economic Analysis. *Journal of Accounting and Public Policy* 22:461–485.

Green, Joshua. 2002. "The Myth of Cyberterrorism." *Washington Monthly*, November. http://www.washingtonmonthly.com/features/2001/0211.green.html.

Hayek, Frederic. 1945. "The Use of Knowledge in Society." *American Economic Review*. 35:519–530.

Hayek, Frederic. 1978. "Competition as a Discovery Procedure." In *New Studies in Philosophy, Politics, Economics and the History of Ideas*. Chicago, IL: University of Chicago Press.

Peltzman, Samuel. 1973. "An Evaluation of the Consumer Protection Legislation: The 1962 Drug Amendments." *Journal of Political Economy*. 81:1049–91.

Peltzman, Samuel. 1974. *Regulation of Pharmaceutical Innovation: The 1962 Amendments*. Washington, American Enterprise Institute for Public Policy Research.

Ricadela, Aaron. 2004. Market Failure is to Blame for Security Problems. *Security Pipeline*. Feb 27. http://www.securitypipeline.com/trends/showArticle.jhtml?articleId=18201051&printableArticle=true.

Ridge, Tom. 2003. Speech at the National Cyber Security Summit. December 3. http://www.us-cert.gov/press_room/detail/RidgeSummitSpeech.html.

Schechter, S. and Smith, C. 2003. "How Much Security is Enough to Stop a Thief? The Economics of Outsider Theft via Computer Systems Networks." In: Proceedings of the Financial Cryptography Conference, January 27-30, Cosier, Guadeloupe.

Tullock, Gordon. 1985. "Adam Smith and the Prisoner's Dilemma." Quarterly Journal of Economics. 402: 1073-81.

Varian, Hal. 2002. "System Reliability and Free Riding." In: Proceedings of the First Workshop on Economics and Information Security. May 16-17. University of California, Berkeley.

THE ECONOMICS OF COMPUTER HACKING

Peter T. Leeson, Ph.D.[*] *& Christopher J. Coyne, Ph.D.*[**]

ABSTRACT

This paper considers various classes of computer hackers, with a special emphasis on fame-driven versus profit-driven hackers. We use simple economic analysis to examine how each of these hacking "markets" work. The resulting framework is employed to evaluate current U.S. policy aimed at reducing the threat of computer hacking and shows that this policy is largely effective. We consider policy adjustments consistent with the insights of the framework provided as a means of strengthening cyber security.

1. INTRODUCTION

In the digital age cyber security is perhaps the most important form of security with which individuals must be concerned. Banks, schools, hospitals, businesses, governments, and virtually every other modern institution you can think of stores and organizes its information electronically. This means that all of your most sensitive information—from credit card numbers and checking accounts to medical records and phone bills—is accessible for viewing, stealing, or manipulating to anyone with a PC, an Internet connection, and some computer know-how. The increasingly computer-based world is increasingly vulnerable to malevolent computer hackers.

While we know little about these shadowy hackers, we have a very clear picture of the damage they do. In 2003, hacker-created computer viruses alone cost businesses $55 billion—nearly double the damage they inflicted in 2002 (SecurityStats.com 2004). In 2000 the total cost of all hack attacks to the world economy was estimated at a staggering $1.5 trillion (PricewaterhouseCoopers 2000). In a 2004 survey of American companies and government agencies conducted by the Computer Security Institute, over half of respondents indicated a computer security breach in the

[*] Department of Economics, West Virginia University. Email: pete.leeson@mail.wvu.edu.

[**] Department of Economics, Hampden-Sydney College. Email: ccoyne@hsc.edu. We thank Peter Boettke, Tony Carilli and Tyler Cowen for helpful comments and suggestions. The financial support of the Critical Infrastructure Project, the Earhart Foundation and the Oloffson Weaver Fellowship is also gratefully acknowledged.

past 12 months and 100 percent of respondents indicated a Web site-related incident over the same period (CSI 2004).

If anything these figures probably understate the volume of hacker-related security breaches. Firms, especially financial institutions, are extremely reluctant to report hacker-related break-ins for fear of how this may affect customers' and stockholders' impressions of their security. In the survey of American businesses conducted jointly by CSI and the FBI, nearly 50 percent of firms that experienced system intrusion over the last year stated that they did not report this intrusion to anyone. The primary reason cited for this was the perceived negative impact on company image or stock (CSI 2004, pp. 13-14), and similar findings have been corroborated by others (see for instance, United Nations 1994; Schell and Dodge 2002, p. 40). What can we say about the enigmatic community of computer hackers and what can we do about the cost these hackers impose?

This paper uses simple economic analysis to try and better understand the phenomenon of hacking. In particular we are interested in creating a framework for analyzing hacking that is policy relevant. Towards this end we divide the community of hackers into three classes separated by motivation. The first class consists of "good" hackers. These hackers illegally break into computer systems but voluntarily share security weaknesses with those in charge of these systems. The second class of hackers is fame-driven. This class constitutes a dangerous subculture of unethical hacking in which members seek infamy and the accolades of their cohorts by breaking into the electronically stored information of vulnerable parties and wreaking havoc. The third group of hackers is "greedy." These hackers are not motivated by considerations of fame but are instead driven by profits. Profit-driven hackers can be "good" or "bad" depending upon which type of behavior yields the greatest monetary return.

An economic analysis of these distinct hacker categories yields important insights for policy aimed at reducing the security threat posed by computer hacking. In Section 2 we offer a brief history of hacking. Section 3 discusses good hackers. Section 4 examines fame-driven hackers. Section 5 considers profit-driven hackers. Section 6 turns to the policy implications of our analysis, and Section 7 concludes.

2. A BRIEF HISTORY OF HACKING

The history of hacking can be traced to 1960s America where members of the Tech Model Railroad Club at MIT "hacked" the control systems of model trains to make them run faster, more effectively, or differently than they were designed to run. Around the same time MIT introduces its Artificial Intelligence Lab where some of the first large mainframe computers are located. With an innate curiosity for how things work, several club members are drawn to MIT's AI lab. These computers—called PDP-1's—are large, slow, and extremely expensive to operate. To overcome

some of these problems the more clever programmers created "hacks"—system shortcuts that make performing certain operations faster and easier.

MIT is not the only locus of hacking activities. Computing think tanks, like Bell Labs, are at it too. In one of history's most important hacks, in 1969 two AT&T Bell Lab workers, Dennis Ritchie and Ken Thompson, create the forerunner of the open source operating system, which they name UNIX. UNIX quickly becomes the standard language of computing. In its first stages hacking has nothing to do with illicit activities or cyber-crimes. On the contrary, access is consensual, and hackers improve systems rather than defacing them.

In the 1970s, however, things begin to change. Hackers start to realize the potential of hacking for personal benefit. In particular, hacking activities are increasingly directed at the telephone—an activity called "phreaking." In the early 1970s a Vietnam veteran named John Draper discovers that the free plastic whistle that comes in boxes of Captain Crunch cereal identically reproduces the 2600 Hz tone required to make long distance phone calls. By blowing the whistle into the phone at the appropriate time AT&T's switching system believes that legitimate access has been granted to make a long distance call and the caller is granted the ability to do so without paying.

After his discovery Draper takes on the pseudonym "Cap'n Crunch" and quickly generates an underground following among hackers and phreakers for his creativity with long distance calling. Other hackers build on Draper's innovation by constructing "blue boxes" designed to aid in the long distance phone fraud process. Notable hackers engaged in such phreaking at the time include Steve Wozniak and Steve Jobs—the future founders of Apple Computers. In 1978, two hackers from Chicago start a computer to computer bulletin board, creating the first virtual meeting place for the growing hacker community where members can share tips, stolen credit card numbers, and other information going into or coming out of their hacking activities.

Partly spurred by the publicity given to hackers in the 1983 film *War Games*, partly spurred by the new affordability of personal computers, and partly spurred by the increasing presence of the online world (ARPANET during this time is becoming the Internet), the prevalence of computer hacking rises yet again in the 1980s. Among the most important hacking developments of this decade is the emergence of hacker "gangs" like the Milwaukee area's "414" gang that consist of hacker die-hards who live to gain unauthorized access to outside computer systems and wreak havoc. The 414 gang is among the first to be apprehended and punished by the law for their cyber-crimes, which include illegally accessing the computer system at Los Alamos National Laboratory where nuclear weapons are developed and breaking into the system at Sloan Kettering Cancer Center in New

York. The 414's are not alone in the new world of hacker crime. The "Legion of Doom" and the "Masters of Deception"[1]—two leading, rival hacker gangs—are also born in the 80s. In response to the growing number of hacker-related crimes, in 1984 the U.S. government makes it a crime to gain unauthorized access to computer systems.

But hacker activity is not limited to breaking into computer systems. In 1988 the world witnesses the first of a new type of hacker act—the Internet worm, which is inadvertently spread by its creator Robert Morris of Cornell University. Morris is identified, fined $10,000, and sentenced to three years probation. The late 80s also see the first cases of hacker action directed at government. Several members of the West German hacker gang, the "Computer Chaos Club," steal electronically stored information from the U.S. government and sell it to the Soviet KGB.[2]

In the 1990s the growing trend of hacker activity prompts the U.S. government to perform surprise raids on the locations of suspected hacker outfits in 14 cities across the nation ("Operation Sundevil"). Although arrests are made, and many inside the hacking community turn on their cohorts in exchange for immunity, hacker activity continues. No longer is hacking mostly about the pranksterish behavior of teenage boys or petty crime. Now hackers turn their talents to much larger deals. In 1995 two Russian hackers steal $10 million from Citibank. In response to more serious hacker activities like this one, in 1998 the U.S. government unveils its National Information Infrastructure Protection Center, designed to protect America's telecommunications, transportation, and technological systems from hacker attacks.

In the new millennium, hacking—an activity once largely restricted to Americans and Western Europeans—is a worldwide phenomenon. The seriousness of the crimes perpetrated by hackers increases again as well. Hackers design "denial of service" hacks that crash the networks of companies like Yahoo!, eBay, Amazon, and others, costing them millions in lost business. The potency and prevalence of damaging viruses also continue to grow, culminating in May of 2000 with the "I LOVE YOU" virus, which is estimated to have cost the global economy close to $9 billion, and is the most harmful hacker-created virus to date (CEI 2002).

As its history indicates, "hacking" refers to multiple activities. It includes, for instance, breaking passwords; creating "logic bombs;" e-mail bombs; denial of service attacks; writing and releasing viruses and worms; viewing restricted, electronically-stored information owned by others; URL redirection; adulterating Web sites; or any other behavior that involves accessing a computing system without appropriate authorization. Furthermore, although for the most part hacking is restricted to computers, it need not be and may be extended to fraudulent activities relating to telephones

---

[1] For a detailed account of the Masters of Deception see Slatalla and Quittner (1996).

[2] For a detailed account of this story see Stoll (1989).

(e.g., tricking phones into authorizing free long distance calls, so-called "phreaking"), credit cards (for instance, creating gadgets to "steal" the magnetic code stored on credit cards and copy it on to others), subway passes (for example, adulterating passes or pass readers to enable unlimited free rides), parking meters (rigging parking meters to allow unlimited free parking) or virtually any other item with electronic components. We restrict our discussion primarily to computer hacking, although the basic principles we elucidate may be applied to other forms of hacking as well.

Some hackers object to calling many of the destructive activities mentioned above "hacking" and their perpetrators "hackers." These terms, they insist, should be reserved to the harmless (albeit often illegal) activities of computer enthusiasts who break into systems, look around to learn how things work and leave things undisturbed. According to this view the name "cracker" should be applied to the malicious "cracking" behaviors enumerated above that are all too frequently conflated with harmless hacking. While we recognize this difference, we nonetheless opt to refer exclusively to hackers and hacking throughout our discussion. On the one hand, in most cases, both hacking and "cracking" involve unauthorized access and so constitute security threats whether or not the individual breaking in uses her illicitly gained access to do harm. Second, for better or worse, in the parlance of our day "hacking" refers to the activities that we describe and the general public does not have the nuanced appreciation of illegal computer activity that members of the hacking community do to merit the terminological distinction implored by some members of this community.[3]

3. GOOD HACKERS

While the psychology of hacking is still in its nascent stages, initial research seems to have come to some consensus regarding what motivates hackers to hack. Individual hackers and hacker gangs operate in the context of a larger underground social network or community consisting of similar individuals. The best empirically grounded work that examines the hacker mind therefore draws primarily on interviews and surveys administered to members of this underground community. We will briefly overview some recent findings of this small literature below. Before doing so, however, we should point out that members of the hacking community are notorious for lying to journalists, researchers, and others who approach them for information about how they and their associates work. Many hackers seem to "get a kick" out of misleading scientists or generally giving others a false im-

[3] As Dann and Dozois put it: "just about everyone knows what a hacker is, at least in the most commonly accepted sense: someone who illicitly intrudes into computer systems by stealth and manipulates those systems to his own ends, for his own purposes (*Hackers* 1996, p. xii).

pression about their reasons for hacking (Platt 1997, p. 53).[4] Of course, this fact must be kept in mind when considering the results of research aimed at identifying hacker motives. Nevertheless, this data is the best we have to date so we must make use of it unless we are to avoid empirical investigations of the subject altogether.

The most current and comprehensive data regarding hackers' demographics, motives, lifestyles, etc. is that collected by Schell et al. (2002). These researchers surveyed over 200 hackers who attended two of America's largest hacker conventions (yes, there are annual hacker conventions in which hackers from across the globe get together to share tips ranging from the latest computer hardware to how to steal credit card numbers stored electronically) in July of 2000. These conventions included the H2K convention in New York and the DefCon 8 convention in Las Vegas. In addition to administering anonymous surveys, researchers randomly interviewed some hackers with in-depth questions (again on the condition of anonymity) when hackers would agree to do so.

The total size of the hacking community is unclear, though by most accounts it is fairly small. According to Sterling, "some professional informants . . . have estimated the size of the hacker population as high as fifty thousand." However, "This is likely highly inflated . . . . My best guess is about five thousand people" (Sterling 1992, p. 77). While we know little about the total size of the hacking community we have a very good idea about its gender proportions. Consistent with figures from others which suggest the population of hackers is overwhelmingly male, only 9 percent of those surveyed by Schell et al. (2000) were female (see for instance, Taylor 1999; Gilboa 1996). Also consistent with older findings, most hackers surveyed were under the age of 30, with a mean age of about 27, a mode of 24 and a median of 25.

The motivation for hacking varies but a significant proportion of hackers surveyed indicated innocuous reasons for their behavior. Thirty-six percent said they hack to "advance network, software, and computer capabilities," 34 percent claimed they hack "to solve puzzles or challenges," and 5 percent said they hack to "make society a better place to live." If we can believe these numbers the overwhelming majority of hackers are harmless. It is true, in gaining unauthorized access to computer systems they pose potential security threats, but they do not themselves cause damage. Of course, to the extent that they share security holes with other less responsible members of the hacking community they indirectly jeopardize computer users; but it is unclear to what extent "good" hackers do this.[5]

---

[4] Taylor suggests that hacker manipulation of the media is partly in order to "revel in the subsequent notoriety" that stigmatizing themselves creates (1999, p. xiii).

[5] In the early 1980s an elite group of hackers calling themselves the "Inner Circle," formed to pass new information gleaned from their hacking activities between one another without making this information available to unethical hackers who would abuse it.

Among these good hackers there is some part of the population that performs a questionably valuable service to computer users. Some of these hackers report security holes to programmers and systems operators of computer systems where they find security weaknesses. This information can then be used to patch holes or strengthen vulnerabilities, preventing intrusion by less benevolent hackers.

Nevertheless, we say questionable here because the advice of these hackers (as well as the hack itself) is unsolicited. According to one popular hacking analogy, it is a bit as if someone broke into your house, didn't steal anything, but left you a note telling you that your alarm system is weak and your windows unprotected, so you should look into having that fixed. While in one sense you are better off because of it, in another sense you may be justifiably outraged.

Unfortunately, data on what proportion of the good hackers are benevolent in this way is not available.[6] We do know that some such hackers exist because insiders at some companies have hinted that certain patches they have released are in response to "good hacker" tips like these. Complicating the issue of good hackers is the fact that some good hackers are far more adamant that vulnerable programmers and systems operators respond to their advice than others. Some good hackers not only inform organizations of security weaknesses but also threaten to release the hole they've found unless action is taken to correct the problem. This is as if someone broke into your house and told you that if you don't buy a better alarm they will inform the criminal community about how it may plunder you.

Good hackers appear to be the most complicated to deal with because they are not motivated by "base" human desires like money or fame. Fortunately, because they pose the weakest threat and are likely responsible for the least damage to individuals and businesses among the hacking community, we lose relatively little at least in terms of felt costs by this dearth of understanding. Far more important from the standpoint of security are bad hackers—those who perform damaging acts in order to gain peer recognition and those who perform such acts for personal profit.

4. BAD HACKERS: HACKING FOR NOTORIETY

The survey conducted by Schell et al. (2000) suggests that only 11 percent of respondents are malevolently motivated. However small the proportion of bad hackers may be, they are the most important to consider because they are responsible for the costly damage inflicted by hackers each year. Contrary to other work which suggests that a substantial propor-

---

[6] Eight percent of those surveyed by Schell et al. (2000) said that they hack to "expose weaknesses in organizations or their products." It is unclear from this, however, whether the reason behind this motive of these respondents is benevolent or malevolent.

tion of hackers are motivated by fame or reputation inside the hacking community, none of those surveyed by Schell et al. noted this reason as their motivation. It is difficult to say why this is, but this result is evidently counter to other examinations of hacker motivation. Fame or peer recognition ranks among the most prominent hacker motivations cited by security experts and hackers alike, as well as in other discussions of hacker psychology (see for instance, Taylor 1999b; Blake 1994; Sterling 1991; Hannemyr 1999; Platt 1997; Thomas 2002; Verton 2002).[7]

As Denning has pointed out, "Although the stereotype image of a hacker is someone who is socially inept and avoids people in favour of computers, hackers are more likely to be in it for the social aspects. They like to interact with others on bulletin boards, through electronic mail, and in person. They share stories, gossip, opinions and information; work on projects together; teach younger hackers; and get together for conferences and socializing" (1992, p. 60).

Bigger, more difficult, more devastating, or new types of hacks bring their creators notoriety among members of their underground community.[8] Word of a hacker's exploits can be spread among community members in a number of ways. First, hackers may spread this information by their own word of mouth, repeating it to fellow hackers or rival gangs who repeat this to other community members and so on. Second, hackers may publicize their responsibility for acts of hacking on Websites, bulletin boards, or on hacker e-mail lists like "BugTraq,"[9] "rootshell," "RISKS Digest," and "VulnWatch." In these virtual spaces hackers take credit for damage done, make information or software that they have stolen available to other hackers, or share their newest methods of hacking or hacking programs they have created with other members of the community so that these individuals may consume them.

In each of these cases hackers identify themselves as the individuals behind new hacks by posting information under their "handles"—pseudonyms chosen by hackers and hacker gangs to give them identity within the hacking community and yet retain their anonymity from authorities.[10] Pseudonyms selected by hackers tend to the memorable and dra-

---

[7] Some other hacker motivations such as the "feeling of power" and "ability to share knowledge" can also be collapsed into considerations of fame. For instance, the more notorious a hacker becomes, the greater her feeling of power. Similarly, her ability to share knowledge will increase with the amount of new information she collects and disseminates, which will also increase her fame.

[8] We should also note that the general public's fascination with the mysterious hacking underworld has helped to fuel fame for members of the hacking community as a whole. Numerous popular movies, for instance, glorify hacking, contributing to this phenomenon. *War Games*, *The Net*, *Hackers*, *Sneakers* and others all provide cases in point.

[9] Interestingly, BugTraq was recently purchased by the computer security firm Symantec for $75 million.

[10] Not all hackers identify themselves by their handles all of the time. Most hackers, however, do so most of the time. The survey conducted by Schell et al. (2000), for instance, indicates 63 percent of

matic, for instance, "Dark Dante" (aka Kevin Poulsen), "Captain Zap" (aka Ian Murphy), "The Nightstalker" (a leading member of the influential hacker group the "Cult of Dead Cows"), etc.—a factor that aids hackers' ability to generate notoriety within the community when they post new information. The same is true of names selected by hacking gangs, for example, "World of Hell," "Bad Ass Mother F*ckers," "Circle of Death," "Farmers of Doom," and so on.[11] The fame-based motivation of many bad hackers helps to explain why profane, absurd, and overstated gang names and handles pervade the hacking underground.

Hackers and hacker gangs that generate celebrity status for their hacks can also set trends inside the hacking community. For instance, two of hacking history's most famous hacker gangs, the Legion of Doom and the Masters of Deception, sparked a trend whereby subsequent hackers and gangs created handles based on comic book characters. Similarly, the 414 gang—one of the first hacker gangs raided by authorities—set the trend of creating handles based on numbers (Schell et al. 2000, p. 58).

The underground world of hackers also has its own popular media that publishes hacking-related books, newspapers, and magazines or e-zines. Some examples of the latter include *2600: The Hacker Quarterly*, *Black Hacker Magazine*, *Computer Underground Digest*, *Phrack Magazine*, *Hack-Tic Magazine*, *The Hackademy Journal*, *Hacker Zine*, *H.A.C.K.*, *Bootlegger Magazine* and *Binary Revolution* to name a few. Inside these outlets hackers publish "how to" articles (e.g., how to defraud an ATM machine) and share new information they have gleaned from their most recent hacking exploits. Articles and books are published under the author's handle and give well-published hackers access to large audiences who thus come to know certain hackers as the "best" in their area, increasing the author's fame inside the community. One of the largest of these publications—*Phrack*—even contains a section called "Pro-Philes" in which famous hackers, retired legends, or rising stars in the hacking community are profiled and interviewed for readers, with special highlights on their biographies and most impressive hacks. In this way, outlets like *Phrack* "served as the means to legitimate hackers for the underground . . . presenting them as celebrated heroes to the readers that made up the underground" (Thomas 2002, p. 140).

Becoming famous through these channels has its benefits for hackers who can generate stardom in the digital underground. Some subcommunities within the hacking underworld will only allow relatively well-known hackers into the community. On the one hand, this gives famous hackers who are admitted greater exposure inside the hacking community,

---

respondents typically use their handles when hacking. This finding is also corroborated by Meyer (1989). Obviously, to some extent the use of handles will depend upon the illegality of the activity. Bad hackers, it is safe to assume, rely upon their handles more than good hackers do.

[11] For examples of other hacker gang names see Platt (1997).

and on the other hand, it gives them access to additional information that may only be shared within the group. Peer recognition also enables hackers to enter elite hacker gangs that are well known and highly respected by other members of the community. As one hacker put it: "Peer recognition was very important, when you were recognized you had access to more . . . many people hacked for fame as well as the rush. Anyone who gets an informative article in a magazine (i.e., *Phrack*, *NIA*, etc.) can be admitted to bulletin boards."[12]

When done right, celebrity in the hacker underground can evolve into outright cult star status as other hackers seek to imitate a notorious hacker's methods or view him as a leader within their community. Such was the case, for instance, with Cap'n Crunch, whose name is forever linked to the practice of phreaking and whose big discovery has led to, among other things, one of the largest hacker publications—*2600*—which is named after his discovery.

"Condor," aka Kevin Mitnick, obtained similar superstar status inside the hacking underground and generated a cult-like following of his own. Mitnick, arrested numerous times for his hacking activities, not only gained notoriety within the hacking sub-community, but became well known to the outside world as well. His picture and story appeared throughout the country in newspapers and magazines, and Mitnick told his story on television's *60 Minutes*. In addition to serving as the basis for numerous books, Mitnick's hacking helped inspire use of the term "Cyberpunk" in popular culture, which was famously used partly in reference to Mitnick by authors/journalists Katie Hafner and John Markoff (1991).[13] Following Mitnick's last arrest in 1995, a group of his hacker community followers protested his trial in the late 1990s. This group of hackers, which had organized itself into a gang called "Hacking for Girlies," broke into the *New York Times* Web page and created a message the *Times* could not remove, exonerating Mitnick for all the site's readers.

Select hackers get the reputation among their cohorts as "elite"—the cream of the underground. These individuals are often gang leaders like "Lex Luther" (former head of the Legion of Doom), or "Phiber Optik" (a former leader of the Masters of Deception), who was even heralded by *New York Magazine* as one of the city's "smartest 100 people." These hackers are the most innovative in the underground and are responsible for making hacking programs publicly available to the hacking community at large. Hacking programs can be downloaded from hacker bulletin boards, for in-

---

[12] Quote from a hacker's email interview with Taylor (1999, p. 59).

[13] William Gibson, credited with coining the term "cyberspace," helped spawn the science fiction genre now called "cyberpunk" in the 1980s (see for instance, Gibson 1984). Some believe that this genre contributed significantly to the shape of hacking culture by glorifying cyber anti-heroes (see for instance, Thomas 2002).

stance, and used with minimal knowledge and effort to hack various systems.

Most hackers, of course, do not reach this level of fame. Their inferior programming skills prevent them from creating effective hacking programs, and instead, most of their energies are devoted to finding and reporting relatively small or already known security holes to fellow hackers, or simply downloading information and prefabricated programs like "Trin00," "Tribal Flood Network," or "Stacheldraht," which were developed by superior hackers and using these to attack systems.[14] These "script kiddies," as they are called, are unlikely to gain fame in the larger hacker community for their hacking skills, but some may gain notoriety for the damage they cause using the programs and information created by more elite hackers. It requires little hacking prowess to crash Amazon.com, for instance, as was demonstrated by "Mafiaboy," the 15 year-old script kiddie whose hacking antics cost some of the Internet's largest vendors $1.7 billion in February of 2000.

Most fame-driven hackers explicitly eschew monetary gain as part of their hacking expeditions. They have contempt for profit-driven hackers who operate or work for computer security companies, or other large computer-related corporations, as though these individuals were beneath them. Fame-driven hackers even have a special, derisive name for these hackers—they call them "Microserfs." This negative reaction to profit-driven hacking has much to do with the cultural norms of the fame-driven hacking community, which in large part believes that big businesses are unscrupulous and views such entities as subordinating the creative skills of the hacker to the greedy corporate world.

4.1 The Economics of Fame-Driven Hacking

The fame-based drive of many hackers has particular implications for how this segment of the "hacker market" looks. The "coin of the realm" for fame-driven hacking is, of course, fame. How we model this "market," therefore, differs from traditional markets in which money drives production and price adjusts to equilibrate suppliers and demanders. The fame-driven hacking "market" considers the relationship between fame and the quantity of hacking. It maps supply and "demand" (which as we will see

[14] Other examples of programs created by hackers that can be downloaded and used by virtually anyone to hack systems include "Black Orifice" created by the Cult of Dead Cows and "L0phtCrack" created by L0pht, and "WinNuke"—all used to hack Microsoft Windows. A similar program called "AOHell" can be used to hack AOL. In 1995, Dan Farmer and Wieste Venema released their "Security Administrator Tool for Analyzing Networks," aka SATAN, an automated program to be used by systems administrators to find flaws in their security. This program could also be used, however, by low-level hackers to hack vulnerable systems, and thus there was great concern it would lead to many problems. To date, it has not caused the harm expected by many.

below is not demand in the conventional sense) in fame/quantity of hacking space.

On one side of this "market" are the producers of hacks who desire fame. The supply schedule for these hackers has the conventional positively sloped shape. When hackers stand to become more famous or better known within the hacker community for hacking, they supply a greater quantity of hacking (which may be expressed in terms of the inventiveness of hacks, the severity of hacks, etc.). When they stand to receive less fame or notoriety for hacking, they are willing to supply less.

The position of this supply curve is determined largely by the cost of hacking. Hackers face a moderate initial fixed cost of hacking, which in most cases comes down a computer, a telephone line (or cable), and a modem. For more sophisticated attacks fixed costs may also include training in basic programming and computer languages, though many kinds of devastating hacks require little specialized training at all. Hackers' variable costs consist primarily of the cost of electricity.

The other primary determinant of the supply curve's position is the number of hackers in the industry. This population is constrained significantly by the number of people who desire fame in the hacker underground (your sister, for instance, is probably capable of hacking but does not desire to be famous among hackers and so does not), which is relatively small. This factor—the population of individuals who desire to enter the "Hacker Hall of Fame"—ends up being the limiting factor determining the position of the supply curve for hacking. Thus, although virtually anyone can cause a lot of damage as a hacker because it is so cheap, very few do so because very few desire the reward it offers—fame among hackers.

The other side of this "market" is unusual in that it does not consist of demanders in the usual sense. When hackers supply more hacks the rest of the hacking community becomes happier. This may be because it gives them access to new information, new hacking methods, and software, which they may value for the purposes of undertaking their own hacking activities or because they view these things as goods in and of themselves. Members of the hacking community may view acts of hacking as expressive of their stand against corporate entities or their belief that all information ought to be publicly available and "free."[15] Others may simply be malicious and enjoy seeing the security of big corporations, for instance, jeopardized, or they may view hack attacks as indirectly serving their political ends.[16]

---

[15] A core component of the hacker "code" ascribed to by so many hackers is that access to computers and all information should be unlimited and free. For a more detailed description of this code see Levy (1994).

[16] Many hackers tend to be strongly left leaning and are adamantly against "commodifying" information. This partly stems from their roots in the "Yippie" movement of the 1960s and 1970s,

In the fame-driven case the hacking community does not pay for more hacking with a higher price. The producers of hacks do not seek money and, as we noted previously, often explicitly reject monetary reward. They seek fame. This, in conjunction with the fact other members of the hacking community value additional hacking, leads them to cheer more, so to speak, when additional hacking occurs. Additional cheering is translated into additional fame for the suppliers of hacks. Rather than demanding the output of suppliers in the usual sense, the other side of the fame-driven "hacker market" consists of individuals (the hacking community) who respond to the supply of hacking with greater or lesser applause. In the language of economists, the hacking community has a reaction function, which specifies how this community reacts with fame to various quantities of hacking that are supplied by hackers. More hacking is rewarded with more applause and less with less applause. The hacking community's reaction function is therefore positively sloped like the supply of hacking itself. The interaction of the supply curve for hacking and the hacking community's reaction function creates two possibilities, depicted in Figure 1 and Figure 2.

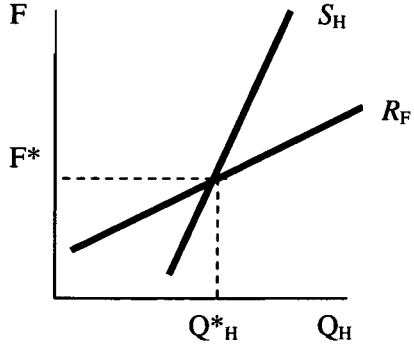


Figure 1.

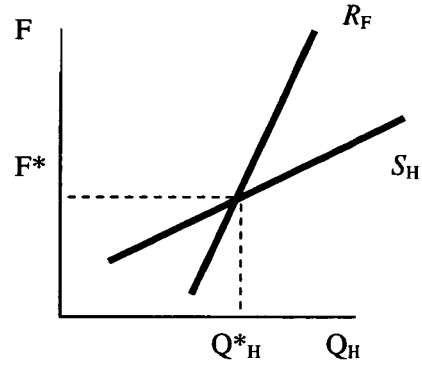


Figure 2.

In Figure 1 hackers' supply curve is less elastic than the hacking community's fame reaction function. In Figure 2 the reverse is true. This means that in Figure 1 the producers of hacks are more responsive (sensitive) to changes in fame than the community of reacting hackers, and in Figure 2 the community of reacting hackers is more responsive to changes in fame than are producers of hacks. These two possibilities have very different (and in fact, contradictory) implications for policy aimed at reducing the quantity of hacking in the fame-driven hacking industry. It is therefore very important to carefully consider the impact of existing policy in each

which in addition to advocating phreaking was largely anchored in the leftist political environment among young people of this time (see for instance, Sterling 1992).

case and, if possible, identify which case is more likely to prevail. We address these issues in Section 6.

5. GREEDY HACKERS: HACKING FOR PROFIT

A third class of hackers is driven by the profit potential of hacking activity. These hackers are concerned with dollars not fame and may come from either pool of hackers, good or bad. From the bad pool are hackers who engage in activities such as credit card fraud, stealing from banks, selling sensitive information stolen from one company to another, or those who are hired by other criminals to do their bidding for a fee.

From the good pool are hackers who work for or operate computer security firms. In 2001 this was a $1.8 billion industry in the United States alone (Wingfield 2002). These hackers sell their skills at finding security weaknesses in computer systems and programs to governmental institutions and private businesses that want to strengthen their security. These organizations hire security firm employees to engage in simulated hacker attacks on their systems and then report vulnerabilities so that they may be corrected. Some of the security experts employed by or running these firms are reformed hackers—individuals who used to hack illegally and either gave it up voluntarily or were caught and punished for their former crimes and so turned to legitimate hacking. Some examples of this include the now defunct, Comsec Data Security operated by four former members of the Legion of Doom, and Crossbar Security operated by Mark Abene (aka Phiber Optik), a former leader of the Masters of Deception. Successful examples of reformed hacker-run security firms include, for instance, ShopIP, run by John Draper (aka Cap'n Crunch) which now has made available a new firewall it calls the "Crunchbox," and Ian Murphy's (aka Captain Zap) IAM Secure Data Systems, Inc.[17]

Out of mistrust, many businesses are reluctant to hire reformed hackers to improve their security. This was ultimately responsible for why Comsec went out of business. Many other organizations, however, are especially drawn to this feature of some security firms because these firms provide the most realistic hack attacks on their systems. Hackers are said to possess a unique way of thinking that leads them to find inventive ways into systems that normal hired hands could not. Major corporations such as American Express, Dun & Bradstreet, and Monsanto, have all hired so-called "tiger teams" to test their systems for vulnerabilities (Roush 1995, p. 39).

The markets for both good and bad profit-motivated hackers look conventional. Since producers seek money, the supply and demand for hacking

---

[17] Former notorious hacker Kevin Poulsen (aka Dark Dante) is now an editorial director for *Security Focus*, an on-line information network for computer security.

are expressed in traditional price/quantity space and price equilibrates the behavior of suppliers and demanders. Both markets exhibit positively sloping supply curves and negatively sloped demand curves. In both cases hackers will provide a larger quantity of hacking if they are paid more and less if they are paid less. Similarly, both criminals and legitimate businesses that hire profit-driven hackers for their purposes demand smaller quantities of hacking when hackers charge more and demand greater quantities when hackers charge less.

The price elasticities of these curves are determined by the standard factors and there is no reason to think that they will be extreme for either the supply of or demand for hacking. Similarly, the position of these curves is determined by the typical elements in each case, with the exception of the fact that the cost of hacking for bad hackers is higher than it is for good hackers because the former involves the possibility of legal punishment while the latter does not. It is therefore reasonable to think that the equilibrium price of hacking in the market for bad profit-driven hacking will be higher than it is in the market for good profit-driven hacking. To the extent that for-profit hackers are willing to supply their services to the highest bidder, the rates of return on bad versus good profit-driven hacking will determine the flow of hackers between these two industries that compete for their labor.

This can be a good thing or a bad thing from the perspective of computer security. If good for-profit hacking is more profitable than bad for-profit hacking, society wins on two fronts from the standpoint of security. The number of bad hackers shrinks endogenously and exogenously. On the one hand more hackers will be employed in activities that do not involve illegally breaking into others' systems, thus reducing the number of potentially harmful hackers out there. Not only this, but the supply of profit-driven hackers no longer employed in harmful hacking is actually employed in fighting the attempts of bad hackers attempting to cause trouble. If, however, bad for-profit hacking is more lucrative, the opposite is true. The supply of hacker threats rises as the best and brightest for-profit hackers are recruited to the dark side.

6. POLICY IMPLICATIONS

The primary federal law in the United States designed to deal with computer hackers is the Computer Fraud and Abuse Act, originally created in 1984 but modified in 1996 by the National Information Infrastructure Protection Act. Originally this law applied only to government computers but it has subsequently been extended to include any computer involved in interstate commerce. This act prohibits under penalty of law: accessing a protected computer without authorization (or exceeding authorized access); accessing a protected computer without authorization and acquiring information; transmitting a program, information, code or command, and as a

result of that conduct, intentionally causing damage to a computer system without authorization (computer viruses); trafficking in computer passwords or other such information through which a computer may be accessed without authorization; and interstate threats for the purposes of extortion to cause damage to a protected computer (Raysman and Brown 2000). The act also prohibits accessing a protected computer without authorization with the intent to defraud where as a result of such action the hacker causes damage in excess of $5,000 over a one-year period.

Most violations of this law can result in up to five years in prison and $250,000 in fines for the first offense and up to ten years in prison and $500,000 in fines for the second offense. Any violation of this law results in a sentence of at least six months. The Computer Fraud and Abuse Act also allows any person who suffers damage as a result of its violation to bring civil charges against the perpetrator for damages. Additionally, since some hacks involve the violation of copyrighted materials, the Digital Millennium Copyright Act punishes those who attempt to disable encryption devices protecting copyrighted work.

In a nutshell, the present law punishes computer hackers, be they good or bad, with stiff fines and jail sentences. It is hoped that through these punishments, hackers will be deterred from hacking. What can our analysis say about this policy?

6.1 Policy and Profit-Driven Hacking

In the case of profit-driven hackers, present policy achieves its desired end. By increasing the cost of bad for-profit hacking through making this behavior criminal, current policy reduces the supply of bad for-profit hacking. The effect of this legislation is two-fold. First, it raises the equilibrium wage of producers who remain in the bad for-profit hacking industry, and second it reduces the quantity of bad for-profit hacking supplied. These effects of current legislation are depicted in Figure 3.

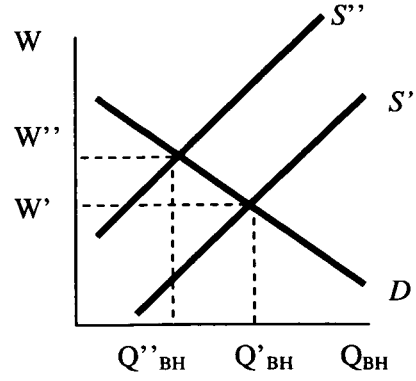


Figure 3.

Although present policy that criminalizes bad profit-driven hacking effectively reduces the quantity of this hacking, this is not all that policy can do towards this end. As we noted earlier, the relative rates of return on working as a bad versus a good for-profit hacker determine which of these markets will garner the best and largest number of profit-driven hackers in general. If it becomes more profitable to be a good profit-driven hacker who owns or works for a legitimate firm, profit-driven hackers currently employed in bad for-profit hacking will be lured out of this industry and into the good profit-driven hacking industry. As we already noted, this has two positive effects on computer security. First, it reduces the number of bad profit-driven hackers, and second, it recruits them to the "good side" in the fight against bad hackers.

One way of making good for-profit hacking look relatively more attractive to for-profit hackers is to raise the cost of bad for-profit hacking, which existing legislation prohibiting this activity does. Another way to increase the competitiveness of good profit-driven hacking, however, is to increase its return vis-à-vis bad profit-driven hacking. To do this, government could subsidize laborers and businesses in the good for-profit hacking industry via outright transfers or through tax breaks and other preferential treatments that result in raising the incomes of those in this industry. The effects of this policy are depicted in Figure 4.

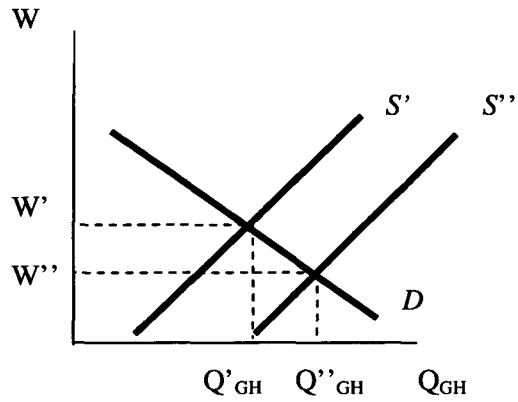


Figure 4.

6.2 Policy and Fame-Driven Hacking

Although current legislation is appropriate for profit-driven hacking, it may not be effective in reducing the quantity of hacking for fame-driven hackers. Recall from Section 4 that the fame-driven hacking industry may look one of two ways. In the first case, the supply schedule for hacking is

less elastic than the fame reaction function for hacking, and in the second case the opposite is true. We also noted in Section 4 that these differing cases have contradictory implications for the effectiveness of present policy. To see why this is so, consider Figures 5 and 6.

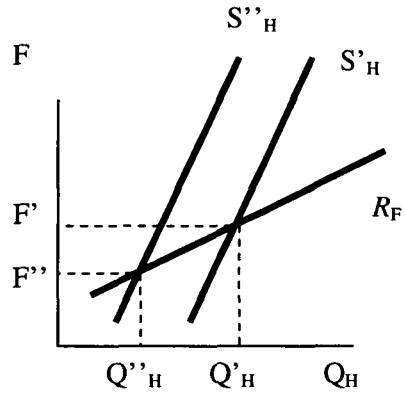


Figure 5.

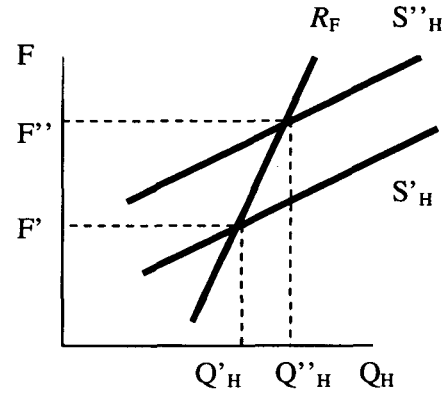


Figure 6.

As with for-profit hacking, current legislation that generically punishes hacking activity raises the cost of fame-driven hacking as well. This leads to a reduction in the supply of hacking, which in Figures 5 and 6 is illustrated by a leftward shift in the supply of hacking from S'_H to S''_H. Note the disparate impact this policy has in each case above. In Figure 5 where the supply of hacking is less elastic than the fame reaction function of the community of hackers, current policy has the desired affect—the equilibrium quantity of hacking drops from Q'_H to Q''_H. Where the supply of hacking is more elastic than the reaction function of the hacking community, the reverse is true. In Figure 6 policy has a perverse effect. Legislation that raises the cost of hacking counter-intuitively leads to more hacking, not less. Specifically the quantity of hacking rises by the amount $Q''_H - Q'_H$. Perhaps strangely, the stiffer the penalty for hacking imposed by law, the greater the increase in fame-driven hacking.

In light of policy's contradictory effects in each of these cases the important question thus emerges: Which of them most likely characterizes the actual fame-driven hacking industry? The "fame elasticity of supply" depends heavily upon hackers' ability to meet increased demand for hacking with additional hacking. Because the marginal cost of hacking is positive and increases with additional output, it is reasonable to think that the supply of hacking is fairly inelastic over at least some range of output.

In contrast, the hacking community's fame reaction function is likely to be relatively elastic. The logic here is simple. The marginal cost of pro-

viding fame is extremely low, if not zero, for the hacking community. Unlike giving up money, which involves sacrificing successively more important alternatives as the price paid rises, providing fame is essentially costless. Increasing the amount of fame the hacking community will "pay" to producers of hacks is very inexpensive. As Cowen points out, "fame remains positive-sum *at its current margin.* Although fame is growing in supply, it is not close to being so plentiful as to lose its exclusive flavor and its power" (2000, p. 114). While the number of famous individuals may grow, fame is not a winner take all, negative-sum game. This is especially true as technologies progress that allow fans to monitor an increasing number of "artists." Increasing fame therefore remains a cheap way to induce more hacking. This means that fame bestowed upon hackers by other members of their community is relatively responsive to changes in the quantity of hacking supplied. Taken together with the fact that the supply of hacking is relatively inelastic, this implies that the fame-driven hacking industry we actually confront most likely corresponds to the case depicted in Figure 5, where raising the cost of hacking does not have a perverse effect. This is good news from the perspective of present policy because it suggests that current legislation is effectively decreasing the quantity of hacking in the fame-driven hacker industry rather than increasing the problem, as it would if the relative elasticities were reversed.

While it is desirable to retain current legislation—which affects the hacking industry through the supply side—demand management could also be effectively used to fight fame-driven hackers. Policies that make it more costly to make the producers of hacks famous—those that reduce the level of fame the hacking community is willing to offer producers for any given quantity of hacking—will further reduce the quantity of fame-driven hacking. Such policies shift the hacking community's reaction function rightward instead of shifting producers' supply curve leftward.

There are at least a few measures that might be taken in this direction. Unfortunately, the most obvious measures towards this end involve violations of basic civil liberties to which many will be opposed. For instance, as we discussed previously, one way by which members of the hacking community give fame to inventive hackers is by publishing them in hacker magazines and books. Prohibiting these publications would not prevent the hacking community from giving fame to hackers, but it would likely force them to find more costly avenues of applauding fame-seeking hackers. The same measures might be taken against hacking community bulletin boards and e-mail lists. Prohibiting hackers from posting hacker programs, tips, etc., it will make it more costly for members of the hacking community to award fame to innovative hackers. Again, for obvious and good reasons, steps like this one are likely to be unpopular. Still, they may remain effective means of reducing the quantity of fame-driven hacking.

7. CONCLUSION

While computer hackers constitute a major security concern for individuals, businesses and public institutions across the globe, hacking and hackers' underground culture remain much of a black box for both lawmakers and those vulnerable to hacker attacks. The mystery that surrounds much of hacking prevents us from arriving at definitive solutions to the security problem it poses; but our analysis provides at least tentative insights for dealing with this problem.

Analyzing computer hacking through the lens of economics gives rise to several suggestions in this vein. First, it is critical to recognize that there are different kinds of hackers characterized by disparate motivations. Because of this, the most effective method of reducing the risk posed by hackers in general will tailor legislation in such a way as to target different classes of hackers differentially. We looked at fame-driven and profit-driven hackers and showed how punishment appropriate for one may actually worsen the problem generated by the other. Current policy directed at reducing hacking by affecting the supply side effectively reduces the quantity of bad profit-driven hacking. Fortunately, there are also good reasons to think that this policy effectively reduces the quantity of fame-driven hacking. If, however, there were strong reasons to think that the elasticities characterized in Figure 6 prevailed over those in Figure 5, supply management that raises the cost of hacking would exacerbate instead of reduce the quantity of fame-driven hacking. We have suggested why we believe this is unlikely to be the case. Still, because of its contradictory policy implications it is important to investigate this issue further.

Our analysis has only touched upon the many and complicated issues regarding computer hacking. In particular, we have not given adequate attention to good hackers who are driven neither by fame nor money, but who voluntarily report security weaknesses to vulnerable computer operators. While the behavior of these hackers is still illegal, it may play an important role in helping to prevent the attacks of more malicious hackers.

We have also not paid sufficient attention to the potential impact that tailoring hacking-related punishments to the age group of the perpetrator may hold for reducing the security threat posed by computer hackers. We noted that most hackers are relatively young—under the age of 30. While this demographic generally cuts across fame-driven and profit-driven hacking groups, there is some evidence suggesting that a disproportionate number of profit-driven hackers are above this age threshold.

The different ages of the individuals in these two different groups suggests that punishments designed to hit each age group where it hurts will be more effective in reducing hacking than a one-size-fits-all approach that may deter the members of one group who are older, but do little to deter the other class of hackers who are younger. In other words, we may want to punish fame-driven hacking, where hackers are younger, with one kind of

punishment that deters younger individuals, and punish bad profit-driven hacking, where hackers are older, with another kind of punishment. This seems relatively simple and yet to our knowledge has not yet been addressed in policy discussions. Presumably 14 year-old script kiddies and 50 year-old men value different things, so effective deterrence will mean differential punishments.

If even after considering these issues it is decided that a uniform punishment for all types of hacking (fame or profit-driven) is desirable, it will still be wise in developing legislation for dealing with hackers to take into consideration the fact that it will inevitably apply primarily to young men. This suggests that effective punishment might be unconventional even if it is uniform across types of hacking. We leave issues like these for future research.

REFERENCES

Blake, Roger. 1994. *Hackers in the Mist.* Chicago, IL: Northwestern University.

Computer Economics Institute. 2002. http://www.computereconomics.com/ article.cfm?id=133.

Computer Security Institute. 2002. *CSI/FBI Computer Crime and Security Survey.* http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2004.pdf.

Cowen, Tyler. 2000. *What Price Fame?* Cambridge, MA: Harvard University Press.

Denning, Dorothy. 1992. Hacker Ethics. *Computing Security.* New Haven, CT: South Connecticut State University.

Gibson, William. 1984. *Neuromancer.* New York: Ace Books.

Gilboa, Netta. 1996. Elites, Lamers, Narcs and Whores: Exploring the Computer Underground, edited by Lynn Cherny and Elizabeth Reba Weise. *Wired Women: Gender and New Realities in Cyberspace.* Seattle, WA: Seal Press.

Hackers. 1996. Edited by Jack Dann and Gardener Dozois. New York: Ace Books.

Hafner, Katie, and John Markoff. 1991. *Cyberpunk: Outlaws and Hackers on the Computer Frontier.* New York: Simon and Schuster.

Hannemyr, Gisle. 1999. E-mail interview with Paul Taylor. In *Hackers: Crime in the Digital Sublime.* London: Routledge.

Levy, Steven. 1994. *Hackers: Heroes of the Computer Revolution.* New York: Penguin Books.

Meyer, Gordon. 1989. *The Social Organization of the Computer Underworld.* MA Thesis. http://project.cyberpunk.ru/idb/social_organization _of_the_computer_underground.html.

Platt, Charles. 1997. *Anarchy Online.* New York: Harper Prism.

PricewaterhouseCoopers. 2000. *Security Benchmarking Service/Information-Week's 2000 Global Information Security Survey.*

Raysman, Richard, and Peter Brown. 2000. *Computer Intrusions and the Criminal Law.* http://www.brownraysman.com/publications/techlaw/nylj0300.htm.

Roush, Wade. 1995. Hackers: Taking a Bite Out of Computer Crime. *Technology Review*, April 1995.

Schell, Bernadette, and John Dodge. 2002. *The Hacking of America: Who's Doing It, Why, and How.* Westport, CT: Quorum Books.

SecurityStats.com. 2004. *Virus Statistics*, January 16, 2004. http://www.securitystats.com.

Slatalla, Michelle, and Joshua Quittner. 1996. *Masters of Deception: The Gang that Ruled Cyberspace.* New York: Harper Perennial.

Sterling, Bruce. 1991. *Cyber View 91 Report.* http://www.eff.org/Misc/Publications/Bruce_Sterling/cyberview_91.report.

Sterling, Bruce. 1992. *The Hacker Crackdown: Law and Disorder on the Electronic Frontier.* London: Viking.

Stoll, Richard. 1989. *The Cuckoo's Egg.* New York: Doubleday.

Taylor, Paul. 1999a. *Hackers: Crime in the Digital Sublime.* London: Routledge.

Taylor, Paul. 1999b. E-mail interview with Zoetermeer. In *Hackers: Crime in the Digital Sublime.* London: Routledge.

Thomas, Douglas. 2002. *Hacker Culture.* Minneapolis: University of Minnesota Press.

United Nations. 1994. *International Review of Criminal Policy.* Available at: http://www.uncjin.org/Documents/EighthCongress.html.

Wingfield, Nick. 2002. "It Takes a Hacker," *Wall Street Journal*, March 11, 2002.

BOOK REVIEW

Adam Thierer and Clyde Wayne Crews, Jr., editors, *Who Rules the Net? Internet Governance and Jurisdiction* (CATO Institute, 2003).

Vint Cert can be legitimately described as one of the fathers of the Internet. In his foreword to *Who Rules the Net?*, he challenges the implication in the title that it can be "ruled." In the technology-agnostic, international realm of the Internet, all entities and interests must coexist. Ultimately, Cert concludes that all users may, in fact, own the Internet to some degree.

Ownership may have many connotations: the ability to use, alter, regulate, establish standards, or preclude others from doing so. The authors of the chapters in this book present compelling arguments of the constituencies having an interest in exerting some measure of control over the Internet without necessarily resolving them. The subtitle of the work is *Internet Governance and Jurisdiction.* The traditional notion of jurisdiction is the ability to exert control over a geographically-defined area. But the Internet is not bound by geography except by the terrestrial location of its component parts. Users can go virtually (i.e. not physically) to any point in the world and complete a commercial transaction, deliver or receive information, or initiate electronic or other activity (e.g. control dams or transportation nodes). Some would say that the Internet is thus not amenable to "control" in the traditional sense, while others assert that it should be treated no differently than existing rules for ordering physical society.

Should the Internet be akin to the "common heritage of mankind", to be shared equally by all without regulation; or has the world become so dependent upon it that it is essential that some minimum world order prevail? And, should the order be imposed by market forces, governments, international advisory bodies, or some combination thereof? The attempt to control on the part of national governments, international organizations, and private entities has already begun. Should control of the Internet, or portions thereof, be left to chance; or should there be some attempt at a rational distribution of authority?

The book is organized into two sections. The first contains essays that discuss some of the central themes of the competing philosophies of Internet governance. Most of the articles are written by legal scholars and educators, or practicing attorneys. In the first chapter, former US Congressman Christopher Cox (now Chairman of the Securities and Exchange Commission) sees the power of the Internet not in economic terms, but in its ability to introduce freedom of thought and democratic ideals into otherwise closed societies. As a member of the House Commerce Committee, Cox had been heavily oriented toward a free market, anti-regulation approach to the Internet and had resisted any expansion of the Federal Communications Commission's authority into new information technologies and services. Other

commentators in the first section focus on the choice of law (i.e. where the event and its effects take place, or the intended "target" audience), extraterritorial jurisdiction, and effective enforcement issues that arise in the cyberspace realm but are also regularly applied now in similar situations.

The second section highlights recent events, many of which will be familiar to a reader, to demonstrate how the broad questions presented above have been addressed in practice. The first such event is a case brought in a French court against Yahoo!® for advertising or selling Nazi memorabilia. While this may be offensive to local sensitivities, it is seen in the United States as an improper curb on First Amendment free speech. Arguments were presented on both sides as to the balance of national preferences as well as the purely technological difficulty of complying with and enforcing the French Court's edict. Recently, both Microsoft® and Yahoo!® have been forced to impose restrictions on their users by the People's Republic of China in order to continue providing Internet access in that country. Other specific issues discussed by commentators in this section include taxation of Internet commerce, enforcement of antitrust rules, and protection of private information.

The second-to-last chapter focuses on the Internet Corporation for Assigned Names and Numbers (ICANN) as an attempt to privatize and stabilize control of the Internet. The idea was to gradually move control of the principal nodes of the network (the Domain Name Server system) from the United States government into the private sector and to replace the voluntary coordination structure with contracts having clear delineations of authority and responsibility. The author concludes that, by trying to satisfy both the government regulatory and the private free market models, ICANN did not satisfy either effectively.

In fact, ICANN's future has figured prominently in international fora recently. The United Nations Working Group on Internet Governance released a report in August of this year in preparation for the second meeting of the World Summit on the Information Society in Tunis in November. Not surprisingly, the report recommended that an international body affiliated with the UN should exercise governance over the Internet to include functions now performed by ICANN. Congress responded to this trend by passing a concurrent resolution of both Houses stating that the current mechanism governing the Internet (i.e. ICANN) should continue, thereby providing the predictability required by the free market. An 11th hour deal at the Tunis summit left the United States in charge, but established multilateral talks to enhance international cooperation.

These recent events demonstrate the utility of *Who Rules The Net?* as a primer on the technical, policy and legal considerations attendant to the Internet. Its authors provide the background for, and form the debate on, questions that will likely go unresolved for many years. Ultimately, some form of "ownership" will be necessary to allow reliable use of a medium which has transformed the world in a very short period of time.

Timothy J. Nagle[*]

[*] Program Manager and former Director of Information Security, Northrop Grumman Corporation.

BOOK REVIEW

Daniel J. Solove, *The Digital person: Technology and Privacy in the Information Age*. (New York University Press, November 2004).

In 1945 F.A. Hayek characterized the market economy as a process of transmitting knowledge between individuals. His insight has been used to explain the functionality of the market system over centrally-planned systems. Through the mechanism of prices, information about individuals' tastes and preferences (as they relate to the quantities and qualities of goods, services, capital, and labor) are communicated across apparently insurmountable obstacles. Knowledge, or lack thereof, in addition to physical scarcity, stand as problems which inhibit individuals from completing their plans and satisfying their wants. Markets, armed with functioning price systems, serve as solutions to this knowledge problem. Since 1945, the role which Hayek placed upon knowledge in society has been a beneficial one; the literature which has developed out of his scholarship has been successful at defeating notions that centrally-planned (socialist) economies can overcome such knowledge problems by intense calculations.[1]

Computer technology has had an interesting part in this socialist calculation debate. Originally, planners made the claim that technology would solve the knowledge problem by providing planners with supercomputers capable of computing the long and intricate calculations of where, when, to whom, and how much goods and services to make and ship.[2] With regard to the application of computer technology in solving the calculation problems of a planned economy, the planners' hopes fell short. Further theoretical claims have been explained to present socialism as completely infeasible; with regard to the computational capacity of technology, the planners' prophecy was more accurate. We have seen the benefits of computers to facilitate the calculation and communication process. This streamlining has opened doors to the potential of trade and wealth creation, diligently noted throughout *The Digital Person*.[3] In summary, telecommunications, as

---

1 F.A. Hayek, *The Use of Knowledge in Society*, 35 Am. Econ. Review 519 (Sept. 1945).

2 Wassily Leonteif, *Input-Output Analysis*, 212 Sci. Am. 25 (Apr. 1965).

3 Solove writes:

These innovations made targeted marketing—or "database marketing" as it is often referred to today—the hottest form of marketing, growing at twice the rate of America's gross national product. In 2001, direct marketing resulted in almost $2 trillion in sales. On average, over 500 pieces of unsolicited advertisements, catalogs, and marketing mailings arrive every year at each household. Due to targeting, direct mail yields $10 in sales for every $1 in cost – a ratio double that for a television advertisement – and forecasters predict catalog sales will grow faster than retail sales. Telemarketing is a $662 billion a year industry. In a 1996 Gallup poll, 77 percent of U.S. companies used some form of direct mail, targeted email or telemarketing (Solove, 2004, p 19).

applied to marketing, production, and distribution, works—allowing companies to attain higher levels of production and profitability.

Listing all of the intricately surveyed information in *The Digital Person* would be redundant if not impossible. It has received numerous rave reviews and rightly so. It is without question that Solove is a thorough scholar and well-versed in the topics of privacy law. This work possesses a unique creativity of metaphor that helps bring the reader through what otherwise would be a tedious journey of technical language and legal precedent.

The knowledge which can be gained from the study of this text is found in the place that Solove's topic has within the broader debate surrounding the role of knowledge in society. Solove's points chime in right around the time we recognize that computers have great potential for advancing the spread and use of productive information. Computers provide tools capable of tapping into dispersed knowledge; but, we must simultaneously recognize that they are not miracle cures to be implemented from central positions of authority. The knowledge which they coordinate is valuable only in so far that it is dispersed and subjective.[4] The hazardous notions of knowledge, in the Hayekian sense, would be those which claim to be more complete and universally applicable than they actually are. When based upon such false notions of knowledge, actions stand to be erroneous, misinformed, and the cause of unintended consequences.

Solove's point is slightly different; he seems to recognize the dynamic nature of information and knowledge[5], but he stresses the unstoppable and inescapable characteristics of such a dynamic process. Solove tries his best to draw attention to the problems associated with the unhampered spread of information resulting from the Internet revolution without distinguishing between central planning and dispersed authority. By emphasizing the importance of privacy Solove plays up the sordid reality of the transmission of knowledge. Attributing the dispersion of knowledge throughout society to malevolence rather than Hayekian productivity has serious implications on the characteristics that responsive policy will take. Such policy stands capable of unintentionally limiting economic growth by stagnating information technology markets and the industries which subsequently rely on them. Thus the insight from Solove's book, placed within the context of

---

[4] This is what Hayek refers to as tacit knowledge. For more on tacit knowledge, see F.A. HAYEK, THE FATAL CONCEIT: THE ERRORS OF SOCIALISM (W.W. Bartley, III, ed., University of Chicago Press 1989).

[5] Solove writes:

> By its nature, tort law looks to isolated acts, to particular infringements and wrongs. The problem with databases doe not stem from any specific act, but is a systemic issue of power caused by the combination of relatively small actions, each of which when viewed in isolation would appear quite innocuous. Many modern privacy problems are the product of information flows, which occur between a variety of different entities. There is often no single wrongdoers; responsibility is spread among a multitude of actors, with a vast array of motives and aims, each doing different things at different times (Solove, 2004, p 61 – 62).

classical liberal constitutional political economy, is of the utmost importance.

Solove's political stance on the issue is clear. Resting upon the description of negative consequences stemming from the transmission of knowledge, Solove seeks a regulatory system akin to our financial and environmental markets in the name of protecting privacy.[6] Forming such imagery is dependent upon the application of successful metaphors. Solove uses metaphors from common literature to express the hazards associated with unhampered collection and distribution of digital data.

At this point, we could argue on the particulars of Solove's assertions, the bulk of which rest upon the notion that we hold a positive right to privacy. It is clear that Solove's interpretation of constitutional appropriateness is ideologically and methodologically different from that of the classical liberal tradition. He views the restriction on government from inhibiting individuals from free association as a spawning ground to support the claim that individuals have a right to privacy. More specifically, governments are particularly restrained from collecting membership rosters of churches and similar groups.[7] If Solove successfully makes the claim that companies or private institutions succeed in diminishing individual privacy in no distinctively different way from states, then he implies that restricting private institutions' ability to collect information in the same fashion that constitutions were used to restrain the state is justified. Thus, we see Solove's constitutional interpretation as recognizing a binding characteristic in line with classical liberal thought, but he wants to spread such binding characteristics into the realm of private companies and institutions. Solove's tendency to lump governments and businesses under the same general category stems from his desire to attribute the problems of bureaucracy to both equally. This stands as yet another point of contention which I would rather not delve into deeply; however, I make reference to the Public Choice School as successfully demonstrating that voting processes and elections contain unique paradoxes which give no epistemological explanation for why the outcomes of such processes should be considered good.

I would claim that this notion of a positive right to privacy stems from the state's monopolization of the production of legislation. Competition drives the process of product improvement. Without competition in the interpretation of legislation or, simply put, the market for judges or courts,

---

[6] Publishers Weekly, Book Review, *at* http://www.amazon.com/ (search "The Digital Person", then follow "The Digital Person" hyperlink, then see "Editorial Reviews").

[7] Solove writes:

> In addition to protecting free speech, the First Amendment safeguards the right of people to associate with one another. Freedom of association restricts the government's ability to demand organizations to disclose the names and addresses of their members to compel people to list the organizations to which they belong. As the Supreme Court reasoned, privacy is essential to the freedom to associate, for it enables people to join together without having to fear loss of employment, community shunning, and other social reprisals (Solove, 2004, p 62 - 63).

we have no certainty with which to judge the legitimacy of judicially proclaimed rights to privacy. What are the costs associated with enforcing a positive right to privacy? Given such costs, would we expect to see firms providing for the enforcement of such privacy, or would enforcement be most prevalent in the competitive market for courts and/or judges? Most likely, we may see individuals take precautions to lower their costs of enforcement; precautions like high fencing, window tinting, and less promiscuous behavior in general. Solove wants to claim this adaptive behavior is a coercive abuse of power, but I am unwilling to make such a normative imputation. Solove concedes the point that marketers are not out to "get us," except to "get us" to buy something; or if anything, to make us aware of the benefits of their products over their competitors' in meeting our needs. The process of competing sellers bidding for consumer dollars opens the potential for individuals to satisfy more complicated demands and live at what they themselves would deem conditions of higher standards of quality.

But the purpose of this review is not to refute Solove's assertion point by point, I doubt there will be many converts either from Solove's camp to classical liberalism or vice versa. I would rather attempt to learn from his position on the margin so as to recognize new applications for the constitutional political economy which has grown from insights such as the opening description of Hayek and other classical liberal positions.

We can concede that the notion of privacy is a concern and take Solove's presentation as playing the role of devil's advocate, being particularly paranoid about the negative effects of insufficient attention paid to privacy. We are still left with the question of which structural system, centrally-planned or market-based, better alleviates such paranoia? I think the market presents a degree of structural compatibility with the technological environment. Solove describes the dynamics of new technologies and the inability of legal torts to keep up.[8] Hayek's knowledge problem rings true again. How can we expect any notion of centrally-planned legislation to keep pace with the momentously changing information technology market? If we cede the point that they cannot, but that we must try anyway, do we not seal our fate to a system of costly enforcement and greater need for state investigation? How could this be? Solove's desired regulatory policy is aimed at inhibiting the breach of privacy, and yet, I assert that such a policy will increase the state's investigations into our private lives. If we recognize the incentives of profiting off of knowledge as momentous, unstoppable, or omnipresent, then the costs of enforcing the prohibition of gathering such information becomes nearly infinite.

Finally, we could respond, on empirical grounds, to assure that such paranoia is rare or even unfounded. Solove's claims that information shar-

---

[8] *See* Solove, *supra* at note 3.

ing has profound effects on the economy are described and elaborated through symbolic metaphor and explained by legal history. But how profound is profound, how big is big? Other than his subjective preference for privacy, is there any notion of economic progress or growth which is dependent upon privacy? I would concede that some markets are intrinsically related to privacy. Any of Solove's examples such as medical prescriptions, personal lifestyles, or credit histories would suffice to show that the structure of society and institutions is influenced by, and in some circumstances, dependent upon, privacy. But is that structure self-reinforcing or completely liable to information breach via technology's advancement? Just as George Mason University's Critical Infrastructure Protection Project (CIPP) papers show that markets won't come screeching to a halt from the marginal effects of cybercrime, the same can be said of the depletion of privacy. It is this link between the notion of cybercrime, information technology, and privacy that makes including a review of *The Digital Person* with the publication of the CIPP papers logical.

Solove offers a neo-Marxist commentary on the state of information technology in society. It is an interesting point, and there is something of value to be learned from it. But, his conclusions are directly dependent upon his subjective attribution of malevolence to the transmission of knowledge. Knowledge is only scary in the sense that we must recognize just how much we do not know. When forced to accept this point, entrepreneurship is encouraged and driven by its placement in a do-or-die scenario. Entrepreneurs are constantly striving to maximize the productive and profitable potential of the knowledge they have and, more specifically, the knowledge they know they have, and have correctly. We see this ring true in database marketing as Solove places the metaphor that information is the "perspiration" of technology. It is a by-product inevitably left over from an existing process but a productive resource in and of itself, which we have not fully mined.

But how do we keep the state in check? Is the answer some form of constitution? Constitutions restrict governments; if we impute a dependency upon the restriction for our own rights, then we require the active production of the restriction. If we give in to the point that governments and corporations are equal, in the sense that they are merely collectives of people and interests, we miss the real point of constitutionalism, allowing it to mutate into intervention by regulation and subjecting ourselves to quite probably a greater loss of privacy.

Daniel J. D'Amico[*]

[*] Ph.D. student, George Mason University Department of Economics.