HACKING BACK: OPTIMAL USE OF SELF-DEFENSE IN CYBERSPACE

Jay P. Kesan* and Ruperto P. Majuca**

*College of Law
**Department of Economics
University of Illinois at Urbana-Champaign

1. INTRODUCTION

Internet security is a big problem. Several approaches have been suggested to deal with the problem, ranging from technological, to law-based, to economics-based solutions.[1] One approach emerging is the notion of self-help – using reasonable force in self-defense against hackers.[2] Hitherto, the law has not taken a clear position on whether or not counterstrike should be allowed in cyberspace. Here, we try to understand the optimality of hackback and articulate what the law on self-defense in cyberspace ought to be. In particular, we seek to answer the following questions: Should society permit hackback? How should the law on self-defense in cyberspace be designed? Which among the tools of combating cybercrimes – law enforcement, court litigation, hacking back the hacker – should be used to most effectively address cybercrimes? What optimal mix of these alternatives should be used to combat cyber-attacks? What role does technology play?

One major argument for hackback is that traditional law enforcement schemes simply do not work in cyberspace because of the speed by which attacks cause great damage to

[1] See generally Kesan and Majuca (2005) for a survey of these various approaches to Internet security.
[2] In real space, various instances of self-help have been recognized by the law, ranging from the use of reasonable force in self-defense or in defense of property in criminal law (see American Law Institute [1985], secs. 3.04 and 3.06), to recovery of property and summary abatement of nuisance in tort law, to repossession and commercial arbitration in commercial law, to the right of restraint and self-help eviction remedies in landlord-tenant relations (see Brandon et al. 1984), and even to such areas as the first amendment, trade secret law, copyright law, and patent law (see Lichtman 2005).

e-commerce sites and also because hackers can stage attacks from multiple jurisdictions with varying cybercrime laws and procedures for prosecuting Internet crimes. As Smith (2005), for example, point outs, while forensic investigation takes time, a virus or worm spreads quickly, underscoring the need to act right away in order to mitigate the grave damage that security incidents can cause. Epstein (2005) believes that even when legal remedies are available, self-help still plays a role because of the numerous instances when the judicial remedy is inadequate or too slow. Thus, Lichtman (2005) points out the importance of bringing together, and capitalizing on the interchangeability between, public and private means, particularly when legal remedies respond slowly to technological risks.

However, some commentators also point out the potential dangers associated with hackback. Kerr (2005), for example, is concerned about counter-strikers hitting innocent third parties rather than the hacker since, in his view, it is easy to conceal the real source of the attack in the Internet.[3] Katyal (2005), on the other hand, argues that private self-help methods not only raise distributional issues (since the rich would be more able to afford themselves of the private measures than the poor), but also fragment the community spirit by weakening the connectivity between people.[4] He thus proposes methods toward community action against cybercrimes.

[3] See also Himma (2004) ("For example, a set of zombies network could … have the direct effect of impairing the performance of the life-support system and hence could result in death of any number of innocent bystanders"). If such argument is correct, would driving a car be morally wrong because of a remote probability that the driver can fatally hit somebody?

[4] In Katyal's (2005) view, individual self-help can cripple interconnectivity and destroy reciprocity. In our view, however, the community spirit is in many circumstances already fragmented by other factors (such as, in the case of the Internet, the anonymity of the actors), and although individual self-help may contribute to the fragmentation, it may be not be entirely fair to withhold such option to individuals since it may be the rational response to opt for individual self-help remedies rather than to wait for the community to address the problem, especially when the community action is not forthcoming.

Those aforementioned papers articulate well the arguments for and against hackback. We believe, however, that there is room for the middle position that articulates in what situations hackback would be a good remedy and when it would not be. Thus, we differ from the more polar views of Kerr (2005) and Katyal (2005) and Smith (2005). Our position is closest to Epstein (2005) in that with him we think that neither blanket permission nor total prohibition of hackback is the right solution, and Lichtman (2005) in that we deem it important to use public means and private methods like self-help to combat cybercrimes. Unlike Epstein (2005) and Lichtman (2005) who do not lay down the criteria for the valid exercise of hackback however, here we actually formulate what are the requirements for a valid exercise of self-help in cyberspace, in the same manner that Posner (1971) proposed the conditions for the use of deadly force in real space. However, unlike Posner (1971), we employ formal modeling to generate our criteria.

Hence, in this paper, we employ formal game theory to model the strategic interaction between the firm and the hacker. This allows us to study the behavior of the hacker given the effectiveness of law enforcement and the potential counter-actions of the firm, and vice versa, and also capture the interaction between law enforcement, court remedies, and self-help remedies. From the Nash equilibria that flow from the model, we observe that the firm will find police enforcement works best in certain instances, while in some cases, resort to the courts based on civil liability litigation will be the better approach, and in

---

The building of community spirit is potentially a fruitful endeavor and thus needs to be pursued more, with the nuts and bolts of the proposal further tightened. (For instance, how does one deal with free riders and shirkers? How are groups formed? How are the responsibilities and costs allocated among the group members? How does one implement and enforce the obligation of each individual in the community?) We agree with Professor Katyal that community-based solutions are probably fruitful pursuits, but this need not entail that individual self-help should be banned outright. In fact, the two may very well go hand-in-hand, particularly in cyberspace where the quickness of the attacks, for instance, may entail that individuals should defend themselves until the community is able to act.

still other situations, self-defense and self-help will best address the cybercrimes problem.

Furthermore, from the social planner's perspective, we show that social welfare is higher when hackback is permitted in society versus when it is not. Also, by identifying the divergence between the private and the socially-optimal solutions, we are able to formulate regulations that are needed in order to bring the private solution closer to the socially-optimal outcome. Thus, explicit modeling enables us to develop litmus tests and criteria that determine if hackback is the proper remedy in certain cases, as well as formulate regulations governing proper conduct during hackback.

The model results generate the following criteria for the valid exercise of self-defense in cyberspace: (1) accounting for trace back costs, the damage to the attacked firm's (that is, the entity that is hacking back) systems that can be potentially mitigated outweigh the potential damage to third parties; (2) there is a relatively high chance of hitting the hacker, instead of innocent third parties; and (3) recourse to police enforcement or civil-action based litigation is either ineffective or impractical. The results also underscore the importance of using good technology (that is, intrusion detection systems (IDS), and trace back technology) in order for hackback to be effective as a deterrent against cyber-attacks.

When such criteria are satisfied, resort to hackback would be justified, and the rules governing proper conduct during counterstrike would come into play: (i) counter-strikers should not cause undue damage to the hacker's computer systems and use only reasonable and proportionate means to defend themselves; and (ii) counter-strikers would be held liable for whatever damages may be suffered by innocent third parties caught in

the crossfire. These added regulations are necessary in order to move the firm's Nash equilibrium outcome towards the socially optimal result. For example, making firms liable for third-party damages will cause them to internalize in their decision-making the potential damage to others and behave closer to the socially-optimal outcome.

As it turns out, these conditions resemble the traditional formulation of the "just war" doctrine,[5] which requires the following necessary elements for a valid counterstrike: (1) there is grave damage (greater than the damage that might result from the action) that will be inflicted to the defender unless it counter-strikes, (2) there is a serious prospect of success, and (3) other means for stopping the evil are either impractical or ineffective (see United States Catholic Conference 1997, ¶ 2309). Interestingly, our requirement that counterstrikers should not wantonly damage the hacker's system and use only necessary force echoes the classical authors' position that war must not be waged for "revengeful cruelty" (Augustine 400, ¶ 74) and that only necessary and proportionate force ought to be used (Grotius 1625).

Since our reasonableness conditions were generated from the social planner's optimization of social welfare, they are consistent with the economic approach to tort law which balances the rights of firms seeking to mitigate damages to their systems and of third parties not being forced to suffer economic harm.

[5] Aurelius Agustinus (354-430), generally acknowledged as the first to have articulated the "just war" doctrine, points out that war must be exercised by the sovereign (¶ 75), and must be waged in order to achieve peace and not for "love of violence, revengeful cruelty, fierce and implacable enmity, wild resistance, and the lust of power, and such like" (¶ 74) (Augustine 400, XXII, ¶¶ 73-79; see also Augustine 423, XIX, chap. 7). Aquinas (c.1271, II, II, Q.40, Art. 1) contributed to the discussion by identifying the three necessary elements for a war to be just: authority of the sovereign waging the war, just cause, and rightful intention. Hugo Grotius, generally known as the father of modern international law, articulated that a just war must contain these basic elements: immediate danger to the nation, necessity of the force employed used is necessary to adequately defend the nation's interests, and proportionality of the force employed to the threatened danger (DeForrest 1997, citing Grotius 1625).

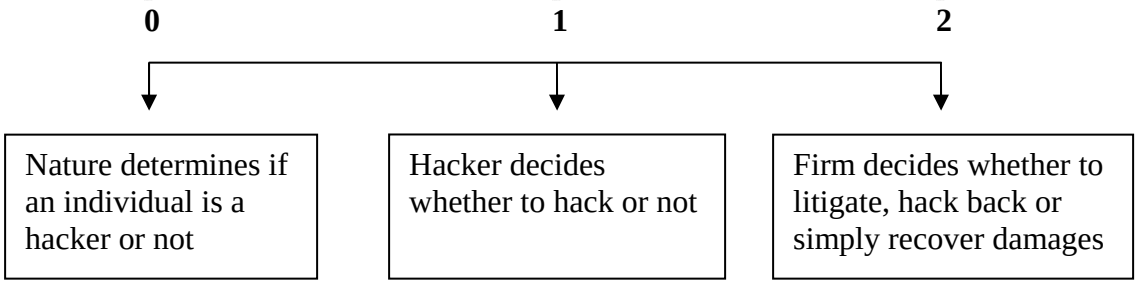
Section 2 presents the basic model. In Section 3, we introduce intrusion detection system (IDS) into the model and examine the role of technology in deterring the hacker and the effectiveness of hackback. Section 4 considers the social planner's perspective and analyzes the divergence between the private and social motive to engage in hackback. Section 5 discusses the proper liability rule for damages to innocent third parties. Based on the model results, Section 6 summarizes what the law of self-help in cyberspace should be. Section 7 concludes our discussions together with some final comments.

2. THE BASIC MODEL

2.1. The Model Set-up

In this paper, we model the interaction between several measures – IDS and traceback (technology), criminal law enforcement and liability-based court litigation (legal remedies), and costs/benefits associated with hackback (economic incentives) – in order to determine how to optimally mix these methods to best address the cybercrimes problem. We start with the basic model of hacker and firm interaction when IDS is not available, and in the next section, we consider the role of IDS technology.

The timing of the game is as follows:



The solution to this game was calculated by specifying the pay-off functions of the parties, solving the first-order conditions, and then calculating the Nash equilibria (see the Appendix for the game tree, the pay-offs and the Nash equilibrium calculations.)

2.2. Equilibrium When Police Enforcement is Effective

Lemma 1. If the probability of catching the hacker times the magnitude of the penalty is bigger than what the hacker gains from hacking, the hacker will not hack (and there is no need for the firm to hack back or to litigate).[6] (This corresponds to area A of Figures 1 and 2 below.)

Lemma 1 states that when the expected punishment exceeds the expected benefit to the hacker, cyberintrusions will be completely eliminated, and there is no need for the firm to resort to hackback or litigation. Effective cybercrime laws and police enforcement therefore act as a broad deterrent against cybercrimes.

However, in the Internet where hackers can situate themselves in different and several jurisdictions with varying computer crime laws of several jurisdictions, the costs associated with the discovery and prosecution of hackers can be prohibitive, and hence, traditional law enforcement measures cannot be entirely relied upon to address cybercrimes. Thus, firms can provide additional deterrence by resorting to litigation and/or counterstrike. In Propositions 1 to 5, we tackle the optimal behavior of the firm when law enforcement is inadequate.

[6] Proofs are presented in the Appendix.

2.3. Equilibrium When Litigation is Not Beneficial

Proposition 1. When litigation is not beneficial, the following Nash equilibria obtain:

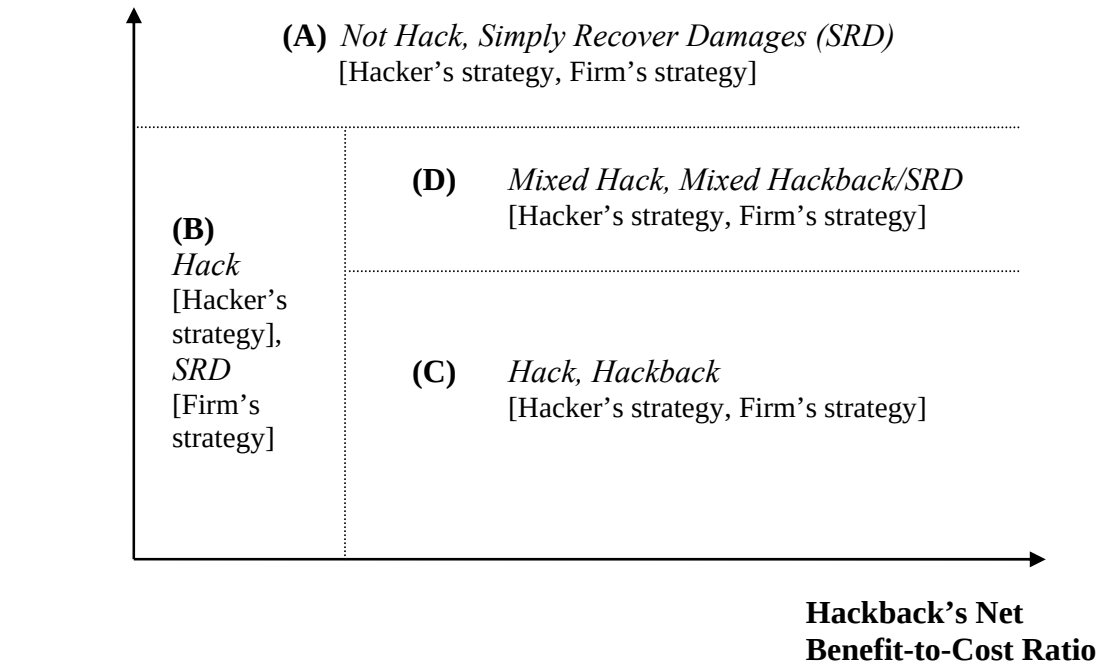


Figure 1. Nash equilibria when litigation is not beneficial
(No IDS available)

When the net payoff from going to court (times the proportion of hackers) is lower than the trace costs, going to court would not be beneficial for the firm. In such a situation, when both police enforcement and litigation are ineffective or impractical, self-help can be useful in mitigating the damage to the firm's e-assets.

The firm's first-order (optimality) condition (see Appendix) shows that the firm's decision to hack back will crucially depend on the damage that can be mitigated to firm's systems relative to those that can potentially be caused to third parties. The firm will find it advantageous to counter the attack only when it calculates that the damage that it can

mitigate by counter-striking would be considerably greater than the potential liability. Moreover, because of the discipline induced by the liability rule, the firm's objective in hacking back will be limited to mitigating damages to its systems and due care will be exercised in order to lessen the damages inflicted to others. That is, the liability rule induces the firm to internalize the potential damages that it may cause others.

Also, the net pay-off from hackback depends crucially on the probability of hitting the right person instead of innocent ones. Thus, from the firm's perspective, the optimality of hackback depends on the available traceback technology: hackback will make sense only if the probability of successfully tracing and hitting the hacker exceed a certain threshold level.[7] If the probability is below this level, then active defense is not optimal. The firm will thus hack back only if there is "serious prospect of success".

In effect, our model shows that the concern of other authors about innocent parties getting hit (see, e.g., Himma 2004, Kerr 2004) is alleviated by liability rules. Liability for damages to innocent third parties causes the firm to strike back only if the probability of hitting the hacker (instead of innocent third parties) is large enough relative to the amounts of damages involved. The fact that the firm would be liable makes it cautious in calculating its chances of (and benefits from) success as compared to the potential liability.

Figure 1 also shows that the propensity to hack back decreases with the effectiveness of law enforcement. Thus, increasing either the probability or the magnitude of the fine has two important effects. First, it reduces the hacker's intrusion rate (compare, for example, the hacker's equilibrium strategy for regions (C), (D) and (A)). Hence, an

[7] This threshold level of probability depends on ratio of third-party damages to the sum of the mitigated damages and third-party damages.

investment in more police resources and international coordination of enforcement efforts, for example, can reduce hacking activity. Second, better police enforcement also reduces the firm's propensity to hack back, as the firm perceives a higher level of protection that law enforcement affords (compare, for example, the firm's strategy for regions (C), (D) and (A)). This emphasizes the substitutability between self-help and law enforcement. This also reduces the force of the argument that since hackers are anonymous in the Internet, hackback should be prohibited (see, for example, Kerr 2005). True, the fact that hackers can attack anonymously in the Internet could mean that innocent third parties may be caught in the crossfire.[8] But that fact, as well as the fact that hackers can situate themselves in different jurisdictions, also mean that cybercriminals are harder for the police to pin down thus lowering the efficacy of traditional police enforcement measures, and increasing the need for self-help measures to substitute in for the slack by providing additional remedy and deterrence against intrusions.

Thus, equilibrium D illustrates the deterrent effect of hackback when law enforcement per se is not sufficient to completely eliminate cybercrimes. When effective police enforcement lies in region D, the hacker adopts a mixed hack strategy in response to the firm's adopting a mixed hackback strategy. In contrast, in the same region of police enforcement, if the firm does not hack back, the hacker will definitely hack (region B). Thus, when the probability times the magnitude of the fine falls below the gain from hacking, self-help can supplement law enforcement because the hacker incorporates in his decision-making the potential damage a counterstrike can inflict on its system.

[8] This concern though, as mentioned, could be handled by liability rules holding firms responsible for the consequences of their actions.

2.4. Equilibrium When Litigation is Beneficial

Proposition 2. When litigation is beneficial, the following Nash equilibria obtain:

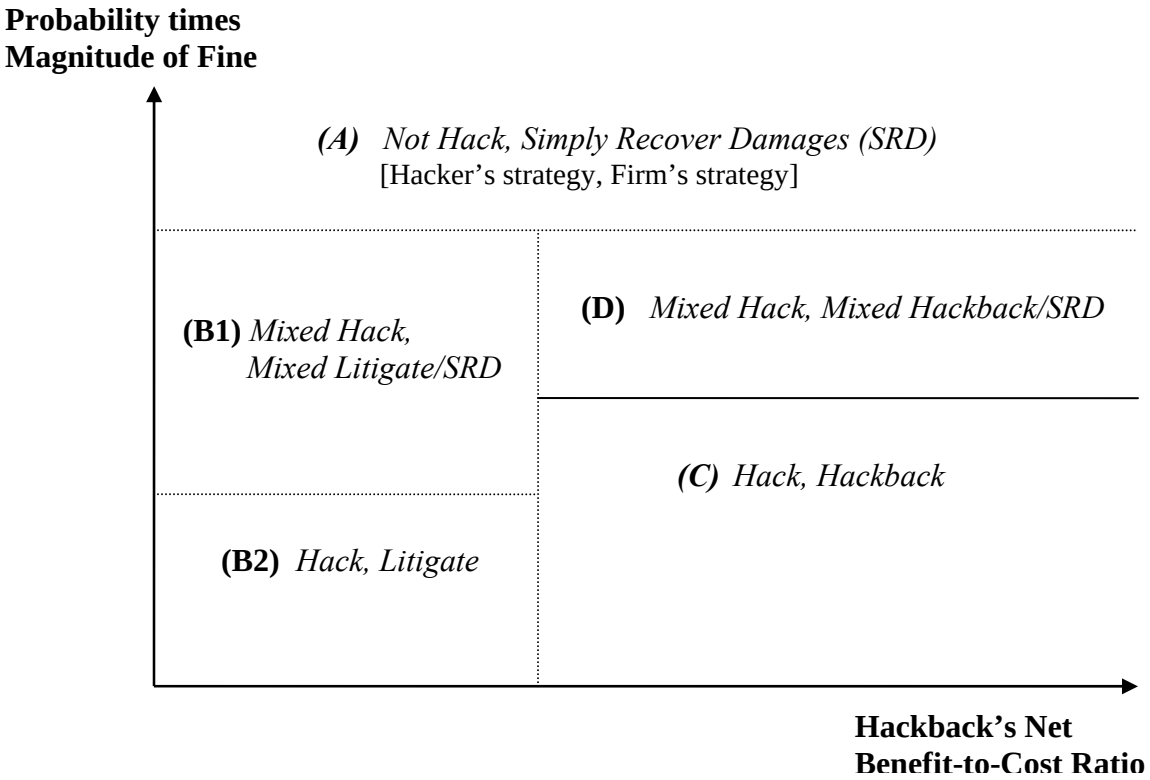


Figure 2. Nash equilibria when litigation is beneficial
(No IDS available)

When the proportion of hackers times the net pay-off from litigation exceeds the trace costs, litigation is a beneficial option for the firm. In this case, the firm will choose between litigation or hackback, depending on whether the net pay-off from litigation exceeds that of hackback's (equilibria B1 and B2 on the left side of Figure 2) or vice versa (equilibria C and D on the right).

The firm will prefer to rely on active defense if its net pay-off exceeds that of litigation. Thus, if active defense will allow the firm to save further damage, or if there

are high transaction costs of going court, self-help will be the more effective remedy. When it is more beneficial for the firm to engage in self-help remedies (equilibrium C and D), the law should perhaps not compel the firm to go to court.

However, self-help is not the better remedy in all situations. Equilibria B1 and B2 illustrate cases where the benefits from self-help are small compared to the more effective relief afforded by the courts. In these cases, self-help measures are not cost-effective, and the courts should be provided as an alternative that the firm can resort to. This illustrates why there would still be a need for the legal system to provide formal legal protections since in certain instances, self-help remedies do not provide complete assistance (see Epstein 2005; Lichtman 2005). The law should thus permit hackback as an option, but not force it as a requirement (see Brandon et al. 1984, p. 870 et seq. for examples of judicially required self-help).

3. THE ROLE OF IDS TECHNOLOGY

In this section, we study the effect of introducing IDS into the model and analyze the role that technology such as IDS can play against cyber-attacks.

If the firm installs an IDS in its security architecture, the timing of the game is modified (see Figure A4 of the Appendix for the modified game tree). With this new IDS set-up, the firm has to consider strategy for two cases: (a) when the IDS signals an intrusion; and (b) when the IDS does not signal an intrusion. Proposition 3 below

presents the case where litigation is not beneficial regardless of whether the IDS signals an intrusion.[9]

Proposition 3. When litigation is not beneficial irrespective of whether or not the IDS signals an intrusion, the following Bayesian Nash equilibria obtain:

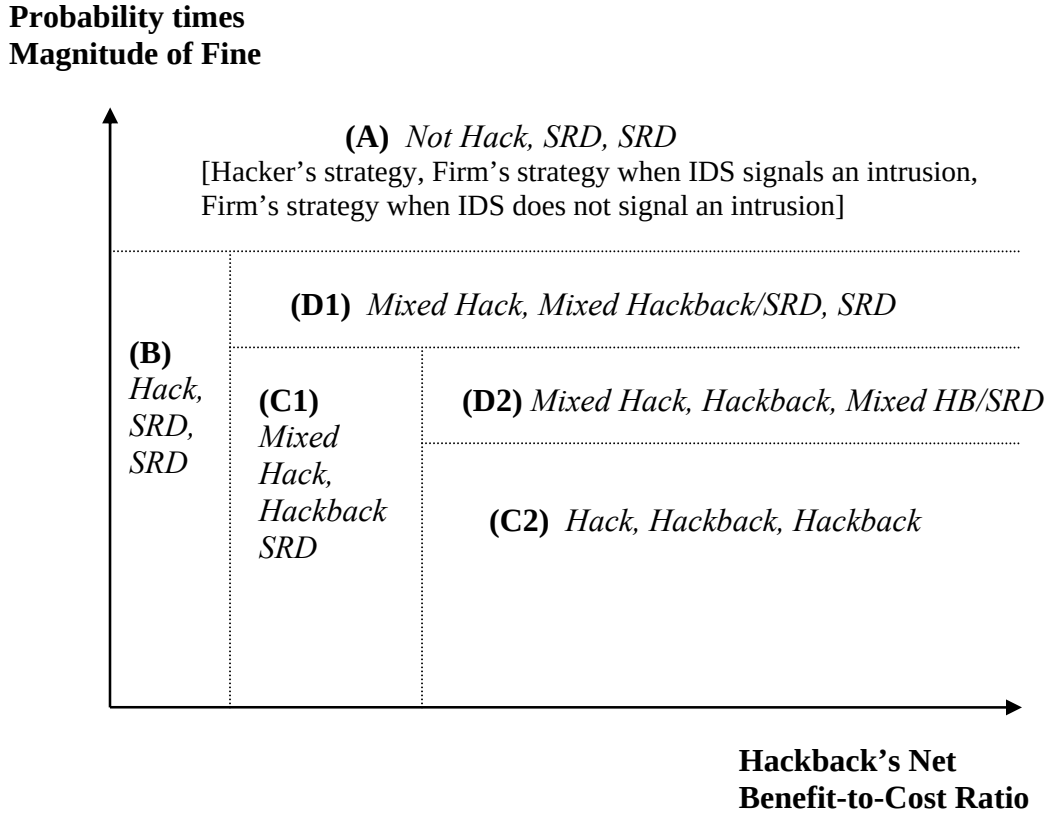


Figure 3. Nash equilibria when litigation is not beneficial (IDS available)

In general, the firm will counter-strike with a higher probability when the IDS signals an intrusion than when it does not. Thus, in region C1 of Figure 3, the firm hacks back when the IDS sends a signal and does not hack back when the IDS does not signal an

[9] Proposition 4 (which covers situations where litigation is beneficial in both the signal and non-signal states) and Proposition 5 (which contemplates cases where litigation is beneficial when the IDS signals an intrusion, but not otherwise) are presented in the Appendix.

intrusion. Also, as Region D1, for example, illustrates, the probability of hacking back is greater when the IDS signals an intrusion than in the previous case when IDS was not available. So too, the firm hacks back with less frequency when the IDS does not signal an intrusion compared to the previous no-IDS case.

The intuition behind these results is that with the IDS, the firm's information as to the probability of intrusion is updated using Bayes' rule (see the Appendix). The IDS thus enables the firm to have better information as to whether or not its systems are under attack. Better information in turn enables an organization to better identify an imminent danger so as to determine if the proper self-help response is a defensive one or a more pro-active one. The IDS thus allows the firm to fine-tune its strategy and be more efficient with its hackback/litigation response (see also Cavusoglu, Mishra, and Raghunathan 2005). The better the IDS configuration (that is, the lower the false positives and false negatives), the more efficient the firm's hackback/litigation strategy will be, and hackback's effectiveness as a cybercrime countermeasure increases.

4. SOCIALLY-OPTIMAL SOLUTION

4.1 The Social Planner's Problem

In Sections 3 and 4, we considered the (firm's) private solution to the hackback game. Since the private and socially-optimal solutions can diverge, we need to consider the perspective of the social planner in order to analyze whether permitting hackback would be optimal for society or not.

In contrast to the firm, the social planner takes into account the total pay-off to all members of society (the hacker, the firm, and the third party) (see equation A10 of the

Appendix for the social planner's pay-off). From the social planner's first-order conditions, we know that the social planner will never litigate because from the social planner's perspective, litigation will result in court costs without any corresponding net social gain (since the amount awarded by the court constitute a mere transfer from the hacker to the firm). Consequently, for the social planner, litigation is never beneficial. Hence, its Nash equilibria would be:

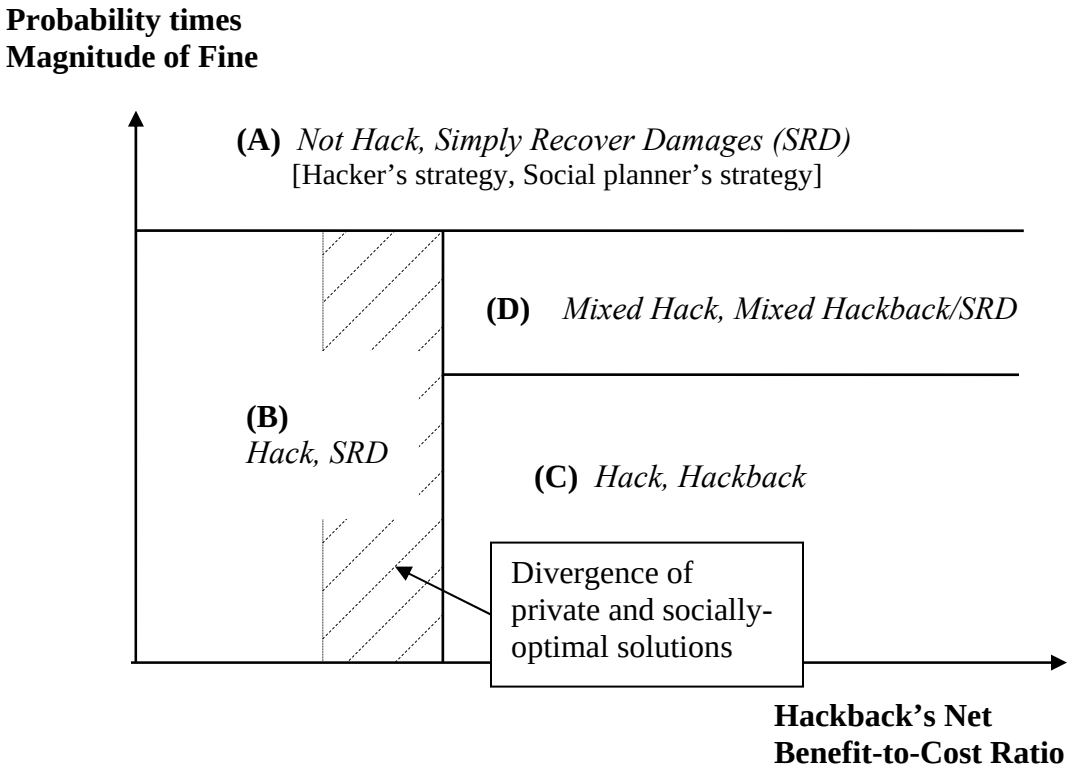


Figure 4. Nash equilibria of the social planner's problem

By comparing the Nash equilibria of the social planner's problem with that of the firm's solution, we can see the divergence between the private incentive to hack back and the socially-optimal level of hackback. Thus, from the shaded region above, we can see that the firm has an incentive to engage in excessive hackback. This is because the firm does

not take into consideration the damage to the hacker's systems in its decision to hack back while the social planner views such damages as losses to society.

We thus think that hackback must be regulated in order to steer the firm's behavior closer towards the socially-optimal solution. The law, for one, should require that in conducting hackback, firms must exert efforts not to wantonly destroy the digital assets of the hacker. In Section 6, we discuss how the law on self-defense in cyberspace should be written.

4.2. The Optimality of Hackback

If hackback were to be for the good of society, it must be the case that the overall social welfare is higher when hackback is allowed, compared to the case when it is not. Hence, here we compare the societal welfare under both regimes by calculating the total societal pay-off across the different regions of the Nash equilibria under both regimes.

Figure 5 below summarizes the social welfare comparisons between the two cases (see the Appendix for the details of the calculations). In region A, the two cases – (a) hackback available, and (b) hackback not available – have similar pay-offs. In region B, cases (a) and (b) again have similar pay-offs. In region C, societal pay-off is higher if hackback is available. Hence, hackback is "good" for society in region C. Finally, in region D, the net society's pay-off is higher when hackback is available provided that the expected net social waste from hacking is positive. We thus conclude that, in general, the availability of hackback is beneficial to society.

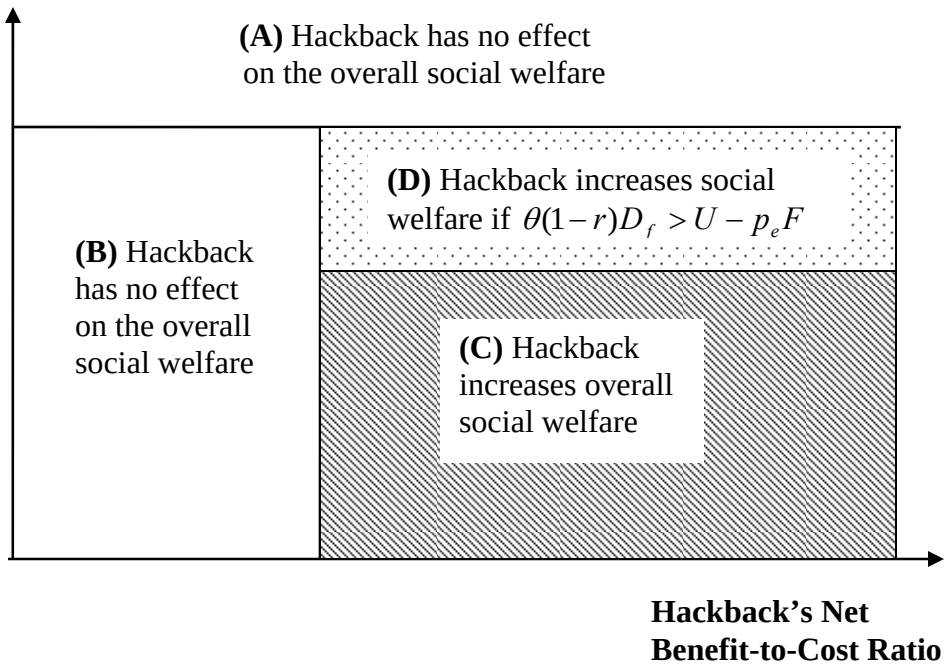


Figure 5. Hackback vs. no hackback social welfare comparisons

5. PROPER LIABILITY RULE FOR DAMAGES TO INNOCENT THIRD PARTIES

We assumed in our model in Section 2 that counter-strikers are liable for third-party damages. If, instead, they are not, the Nash equilibria would be different, as shown below:

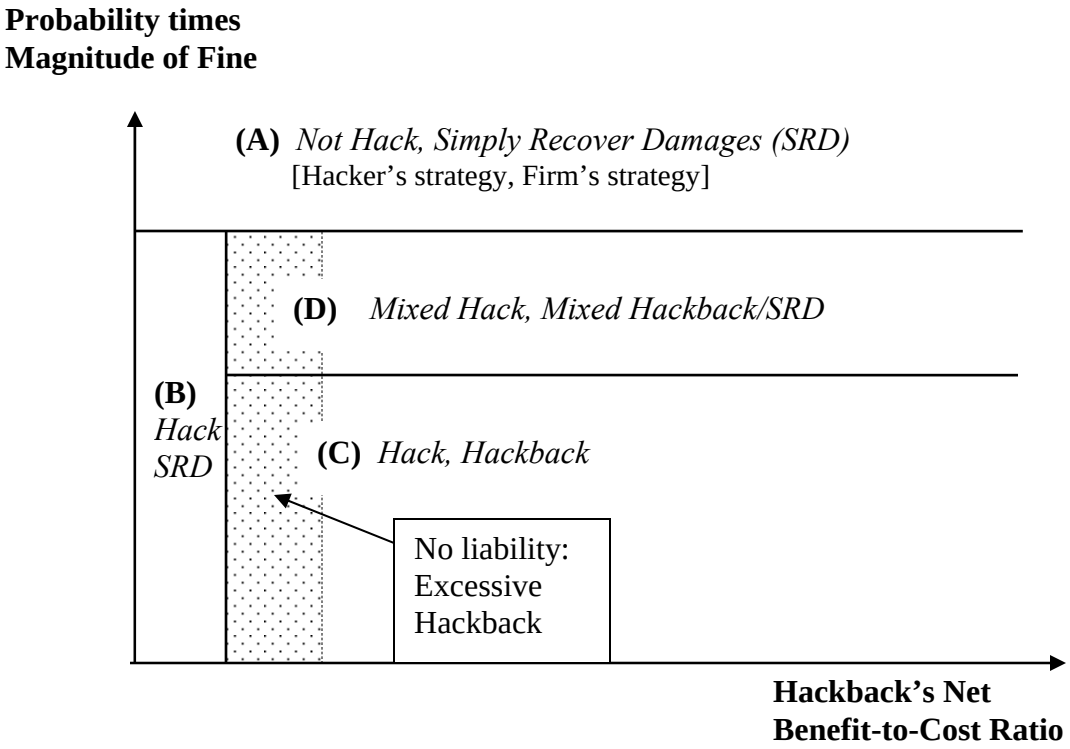


Figure 6. Nash equilibria of the firm's problem
(No liability rule)

Figure 6 shows that, *not* holding counterstrikers liable for third party damages would cause a distortion that compounds the wedge already caused by the firm not incorporating the damage to the hacker's systems into its cost-benefit calculations (Figure 4). With both these distortions, the overall divergence of the private and socially-optimal solutions would be larger, as depicted in Figure 7.

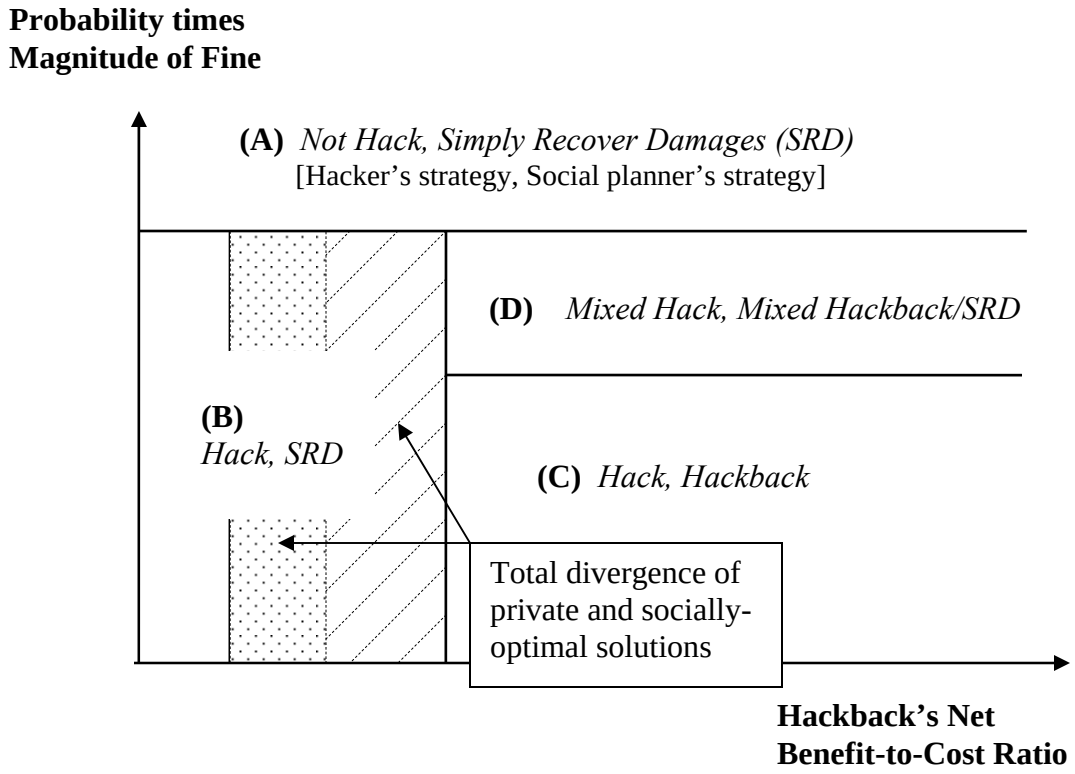


Figure 7. Divergence of private and socially-optimal equilibria

Thus, not holding firms liable for third-party damages would exacerbate the inefficiency of the *laissez-faire* hackback regime. Hence, there will now be two distortions that cause the private solution to diverge away from the socially-optimal solution: (a) the fact that the damage to the hacker's system are social losses not considered by the firm (striped lines), and (b) the fact that the damage to third parties are social losses not internalized by active defenders (dotted lines). Thus, our model shows that liability rules for third-party damages function like an "invisible hand" guiding the private solution closer to the socially-optimal solution.

On top of liability rules, however, regulations are still needed to handle the other distortion. Besides, even if third-party liability rules are present, several "frictions" could cause a wedge between the efficient amount of hackback and the actual amount hackback

(see Figure A10 in the Appendix for the Nash equilibria when these frictions are present). For example, several of those caught in the line of crossfire are likely to not have enough computer sophistication to even detect that their systems have been hit.[10] Also, others may decide not to sue given the transaction costs involved with going to court.[11] Because of these concerns, liability rules for damage caused to third parties, in and of themselves, may not be enough to generate the optimal outcome. Consequently, on top of having liability rules, regulations establishing criteria and guidelines for valid hackback should be set in order to constrain the parties closer to the socially-optimal outcome.

6. WHAT THE LAW ON SELF-DEFENSE IN CYBERSPACE SHOULD BE

Given the results of the model, we are now ready to formulate what the law on self-defense in cyberspace ought to be.

First and foremost, we do not see any overriding reason why the law should outright deny firms the right to exercise self-defense in cyberspace (see, for example, equilibria C and D of Figure 5 where the availability of hackback increases the social welfare). With the time and expense associated with court-administered remedies, the availability of self-help could provide an equitable solution (Brandon et al. 1984, pp. 869-70). Moreover, by deterring criminals, active defense can supplement law enforcement: once precedents of hackers experiencing damage from counterstrike are established, a large number of script kiddies would vanish (see discussions in Section 2.3 on the deterrent

[10] We thank George Deltas for bringing up this point.
[11] Also, the damage may be *de minimis*, in which case resort to courts may not be available (see Epstein 2005).

effect of hackback). In sum, we think that absent the showing of widespread misuse, active defense should not be outlawed at the outset (see also Epstein 2005).

Secondly, counterstrikes can, however, function as a wrong against innocent third parties and what passes as self-defense may in reality be another wrong. Similarly, self-help though originally justified can bring about several harmful results.[12] For these reasons, reasonableness standards must be instituted and resort to legal remedies prescribed when the planned counter-actions fall outside their boundaries (Epstein 2005). Thus, in our view, the law should in some instances allow (though not require) resort to self-help remedies, yet at the same time regulate the exercise of the privilege so as to check against its potential abuse.

This is where we differ from Kerr (2005) and Katyal (2005) in that while we recognize that excesses and abuses can potentially occur, for us, this does not necessarily mean that the privilege of self-defense should be denied outright. Given the potential benefits self-help can generate when used responsibly, we think that regulating the exercise of the privilege is the best way to deal with these potential excesses.

Based on our model results, the governing regulation on self-defense in cyberspace should have the following features:

(1) Attacked firms and individuals can hack back if, and only if, the following requirements are satisfied:

(a) Hackback does not result in greater harm to innocent parties compared to the damage to the defender's systems that is sought to be mitigated. Furthermore, due care

[12] The standard litany of criticisms of hackback include misidentification problems, use of automated program by counter-strikers, shooting matches between trigger-happy defenders and intruders, self-proclaimed "white hats" releasing worm patches with good intention but with terrible results, etc.. (see, e.g., Katyal 2005).

should be exercised to avoid or minimize damage to third parties and the purpose of the hackback should be limited to the prevention of damage to the firm's information technology infrastructure.

(b) Recourse to other alternatives is either ineffective or impractical. In particular, this occurs when:

(i) police enforcement is ineffective. Lemmas 1 and 3 show that effective criminal law enforcement provides wide-ranging deterrence against cybercrimes and does away with the need for counterstrikes or civil litigation.

(ii) litigation is impractical (see discussions in Sections 2.3 and 2.4).

(iii) a more defensive strategy, such as simply recovering damages or simply dropping incoming packets, would not deter the hacker.

In short, active defense is an extraordinary remedy, available only when other alternatives are ineffective or impractical.

(c) There is a serious prospect of success. There must be a relatively high chance of hitting the hacker, instead of hitting innocent persons. Thus, reasonable effort must be exerted to employ state-of-the-art traceback technology.

(d) Reasonable effort must be exerted to employ good IDS technology. This helps the firm to more carefully ascertain the existence or the imminence of the attack/danger; it also decreases the error of hitting innocent persons; and enhances the deterrent effect of hackback (see Section 3).

If a firm hacks back without these conditions being present, it oversteps the bounds for reasonable exercise of self-defense in cyberspace. The law can hold those who exercise self-help not in a legally permissible manner, liable for penalties.

(2) Even if those preconditions are present – and thus the exercise of the privilege is justified – the conduct *during* hackback must also be regulated by the law:

(a) In order to internalize the damage to third parties, active defenders should be held liable to third parties caught in the crossfire. Not holding active defenders responsible for the consequences of their action will result in externalities and excessive amount of hackback activity (see Figure 6).

(b) Counter-strikers must also use only "proportionate force", that is, they must not wantonly damage the hackers' digital systems out of retaliation, but rather, only use force that is necessary to avoid damage to their own systems (see Figure 4).

In sum, the law needs to layer liability rules on top of the reasonableness conditions.

7. CONCLUSION

We believe that self-defense springs from the natural instinct for self-preservation. Hence, hackback should not be banned outright – it is generally accepted that one has the right to defend one's self and one's property and, toward this end, use reasonable force (see Aquinas c.1271, II, II, Q.64, Art. 7). The fact that the exercise of this right can be abused does not necessarily mean that the right should be denied at the outset; it does, however, mean that the exercise of the privilege should be regulated.

In this paper, we formulated criteria and guidelines that articulate under what circumstances self-defense is proper in cyberspace, and in what situations should we instead rely on the police or resort to the courts. Using a game-theoretic model of the interaction between the defender and the hacker, we were able to capture the interplay

between legal remedies (police enforcement and court litigation), technology (IDS and traceback), and economic incentives (cost and benefits of self-help remedies), and thus develop specific rules or tests for resolving whether resort to hackback is justified vel non. Based on the results from the model, the criteria for valid resort to hackback are: (1) other alternatives, such as police enforcement and resort to courts, are either ineffective or ineffectual; (2) there is a genuine prospect of hitting the hacker instead of innocent third parties; and (3) the damage that can be mitigated to the defender's systems outweigh the potential damage to third parties. Additionally, when hackback is justified, the following rules govern conduct during hackback: (4) defenders must not use excessive force, that is, they must only use force necessary to defend their property and not needlessly destroy the hacker's digital assets; and (5) counter-strikers would be held liable for damage to other third parties. Thus, liability rules should be set in place so that firms will internalize the damage to third parties, thereby bringing the private incentive to hackback closer to the socially-optimal outcome. In sum, the law should layer the third-party liability rules on top of the reasonableness conditions mentioned above.

APPENDIX

Summary of Notations

U = hacker's utility from hacking
D_f = firm's damage due to hacking
D_h= damage to the hacker if the firm hacks back
F = fine the hacker pays if caught by law enforcers
W = amount awarded the firm if it wins the litigation
K = cost of going to court
r = percentage of damage recovered if the firm decides to "simply recover damage"
h = percentage of damage mitigated by resorting to hackback
d = damage incurred by third parties (as percentage of the counter-striker's damage)
p_h = probability of hitting the hacker if the firm hacks back
p_w = probability of winning the case if the firm litigates
θ = proportion of hackers in the population
π = fraction of affected third parties who forgo suit against the counterstriker
q_1= probability of getting an IDS signal given that there is an intrusion
q_2= probability of not getting an IDS signal given that there is no intrusion
η_1 = probability of intrusion given the IDS signals an intrusion
η_2 = probability of intrusion given the IDS does not signal an intrusion
δ = probability that the hacker hacks
σ_1 = probability that the firm hacks back in the no IDS case
σ_2 = probability that the firm litigates in the no IDS case
α_1= probability that the firm hacks back if the IDS does not signal an intrusion
α_2= probability that the firm litigates if the IDS does not signal an intrusion
β_1= probability that the firm hacks back if the IDS signals an intrusion
β_2= probability that the firm litigates if the IDS signals an intrusion

The Model without Intrusion Detection System

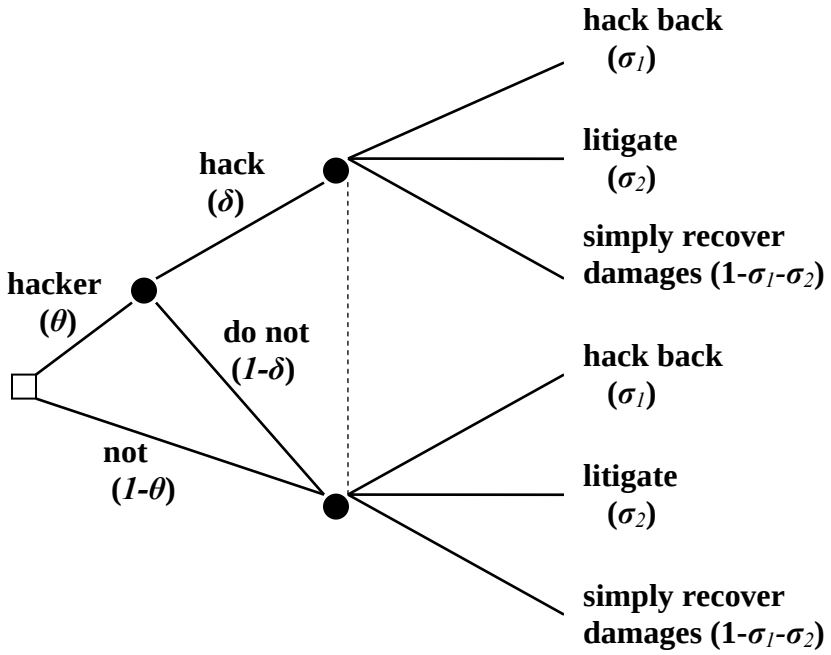


Figure A1. Game tree, no IDS case

The total expected pay-off for the firm and the hacker, are respectively:

$$F(\sigma_1,\sigma_2,\delta) \equiv -C_t(\sigma_1+\sigma_2) - \theta\delta\begin{Bmatrix}\sigma_1[p_h(1-r-h)D_f + (1-p_h)(1-r+d)D_f] \\ +\sigma_2[(1-r)D_f + K - p_w W] \\ +(1-\sigma_1-\sigma_2)(1-r)D_f\end{Bmatrix} \quad \text{(A1)}$$

$$H(\sigma_1,\sigma_2,\delta) \equiv \delta(U - p_e F - \sigma_1 p_h D_h - \sigma_2 p_w W) \quad \text{(A2)}$$

The first-order conditions are:

$$\frac{\partial F}{\partial \sigma_1}(\delta) = -C_t + \theta\delta[p_h h - (1-p_h)d]D_f \quad \text{(A3)}$$

$$\frac{\partial F}{\partial \sigma_2}(\delta) = -C_t + \theta\delta[p_w W - K] \quad \text{(A4)}$$

$$\frac{\partial H}{\partial \delta}(\sigma_1,\sigma_2) = U - p_e F - \sigma_1 p_h D_h - \sigma_2 p_w W\,. \quad \text{(A5)}$$

Proof of Lemma 1. (A5) $\Rightarrow \frac{\partial H}{\partial \delta} < 0 \Rightarrow \delta = 0$. Substituting, $\frac{\partial F}{\partial \sigma_1}(\delta = 0) = -C_t < 0$ and $\frac{\partial F}{\partial \sigma_2}(\delta = 0) = -C_t < 0 \Rightarrow \sigma_1 = \sigma_2 = 0$.

Proof of Proposition 1.

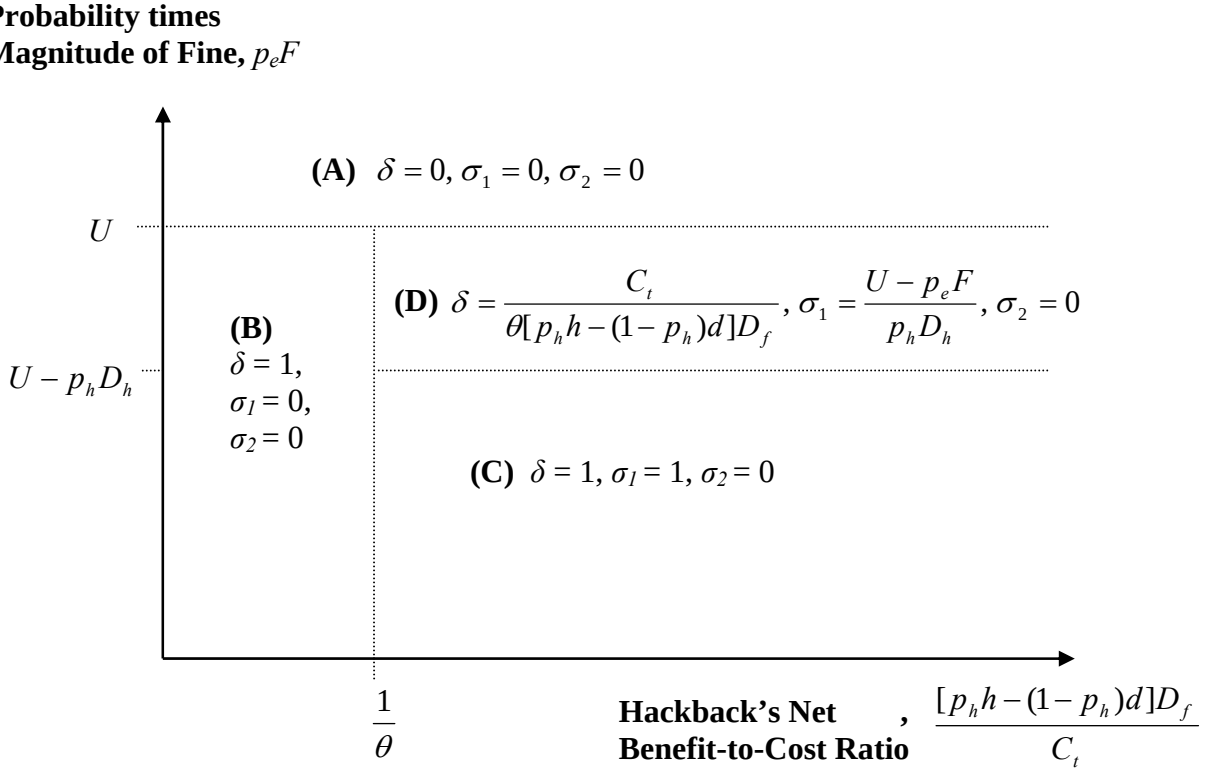


Figure A2. Nash equilibria when litigation is not beneficial (No IDS available)

$\theta[p_w W - K] < C_t \Rightarrow \frac{\partial F}{\partial \sigma_2} = -\frac{C_t}{\delta} + \theta[p_w W - K] < 0 \Rightarrow \sigma_2 = 0$ in all equilibria here.

Equilibrium B: (A3) $\Rightarrow \frac{\partial F}{\partial \sigma_1} = \frac{-C_t}{\delta} + \theta[p_h h - (1-p_h)d]D_f < 0 \Rightarrow \sigma_1 = 0$. Given $\sigma_1 = \sigma_2 = 0$, $H(\sigma_1 = \sigma_2 = 0) = \delta(U - p_e F)$. $\frac{\partial F}{\partial \sigma_1} > 0 \Rightarrow \delta = 1$. *Equilibrium C:* (A3) $\Rightarrow \frac{\partial F}{\partial \sigma_1} > 0 \Rightarrow \sigma_1 = 1$. Given $\sigma_1 = 1$ and $\sigma_2 = 0$, $\frac{\partial H}{\partial \delta} > 0 \Rightarrow \delta = 1$. *Equilibrium*

D: We know by Nash (1950) that if $U \geq p_e F > U - p_h D_h$ and $\theta[p_h h - (1-p_h)d]D_f > C_t$, an equilibrium in mixed strategies exists. Setting $\frac{\partial F}{\partial \sigma_1} = 0 \Rightarrow \ \delta = \frac{C_t}{\theta[p_h h - (1-p_h)]D_f}$. Given $\sigma_2 = 0$, $H(\sigma_2 = 0) = \delta(U - p_e F - \sigma_1 p_h D_h)$. Hence, $\frac{\partial H}{\partial \delta} = 0 \Rightarrow \sigma_1 = \frac{U - p_e F}{p_h D_h}$.

Proposition 2.

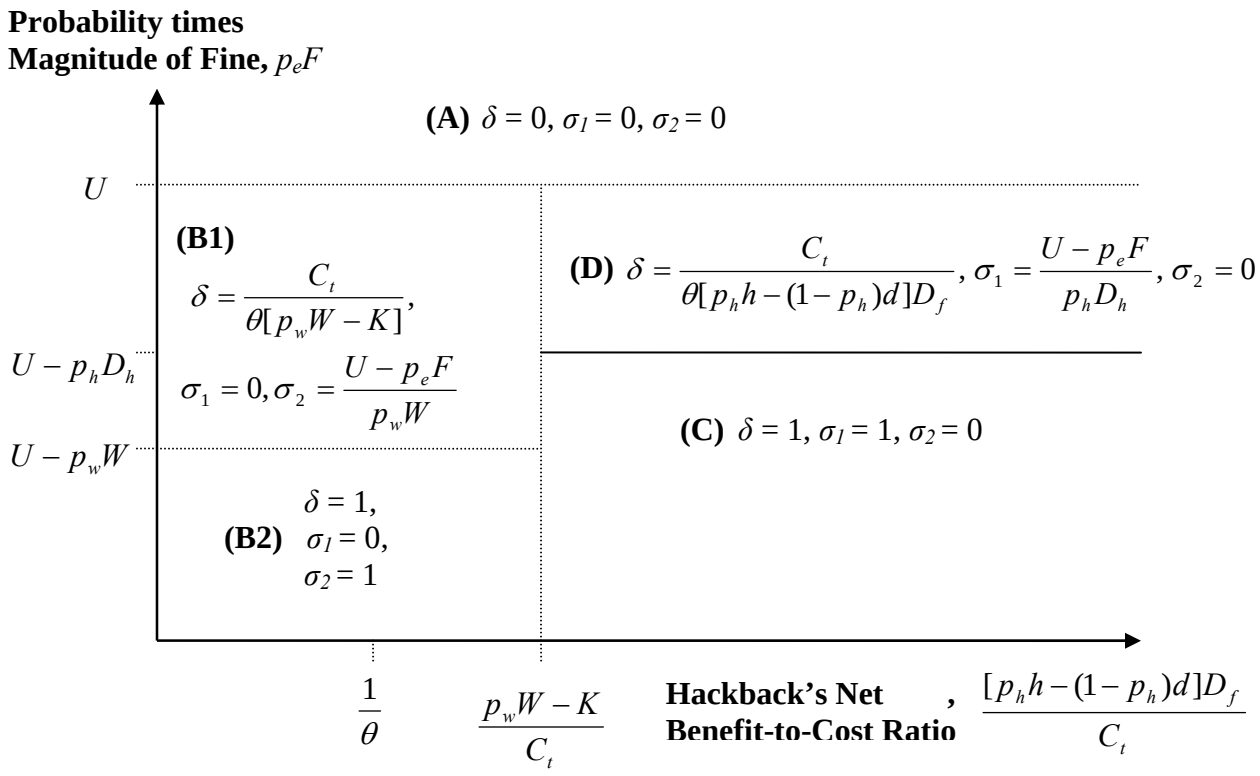


Figure A3. Nash equilibria when litigation is beneficial
(No IDS available)

The proof, which is omitted due to space constraints, is available from the authors upon request.

Model with Intrusion Detection System

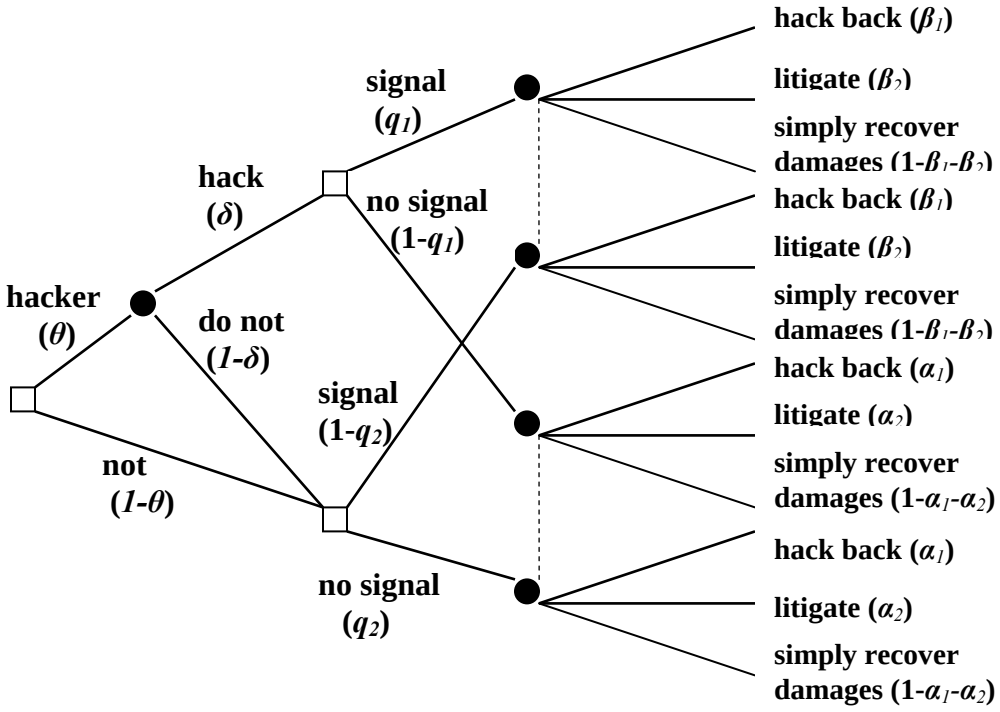


Figure A4. Game tree, IDS case

Firm's expected pay-off in the signal states:

$$F_s(\beta_1,\beta_2,\delta) \equiv -(\beta_1+\beta_2)C_t - \eta_1\beta_1[p_h(1-r-h)D_f + (1-p_h)(1-r+d)D_f] \\ -\eta_1\beta_2[(1-r)D_f + K - p_w W] - \eta_1(1-\beta_1-\beta_2)(1-r)D_f \quad \text{(A6)}$$

where η_1 = Pr(intrusion|signal) $= \dfrac{q_1\theta\delta}{q_1\theta\delta + (1-q_2)(1-\theta\delta)}$.

Firm's expected pay-off in the non-signal states:

$$F_n(\alpha_1,\alpha_2,\delta) \equiv -(\alpha_1+\alpha_2)C_t - \eta_2\alpha_1[p_h(1-r-h)D_f + (1-p_h)(1-r+d)D_f] \\ -\eta_2\alpha_2[(1-r)D_f + K - p_w W] - \eta_2(1-\alpha_1-\alpha_2)(1-r)D_f \quad \text{(A7)}$$

where η_2 = Pr(intrusion| no signal) $= \dfrac{(1-q_1)\theta\delta}{(1-q_1)\theta\delta + q_2(1-\theta\delta)}$.

Firm's overall expected pay-off:

$$F(\alpha_1, \alpha_2, \beta_1, \beta_2, \delta) \equiv [q_1\theta\delta + (1-q_2)(1-\theta\delta)]F_s + [(1-q_1)\theta\delta + q_2(1-\theta\delta)]F_n . \quad \text{(A8)}$$

Hacker's expected pay-off:

$$H(\alpha_1, \alpha_2, \beta_1, \beta_2, \delta) = \delta[U - p_e F - (1-q_1)(\alpha_1 p_h D_h + \alpha_2 p_w W) - q_1(\beta_1 p_h D_h + \beta_2 p_w W)]. \quad \text{(A9)}$$

[Note: Subsequent first-order conditions and proofs are omitted due to space limitations.]

Lemma 2. (a) $\beta_2 \geq \alpha_2$ and (b) $\beta_1 \geq \alpha_1$.

Lemma 3. $U < p_e F \Rightarrow \delta = 0, \beta_1 = \beta_2 = 0$.

Proposition 3. When $\frac{p_w W - K}{C_t} < a_1 \equiv \frac{q_1\theta + (1-q_2)(1-\theta)}{q_1\theta}$, the following Bayesian Nash equilibria obtain:

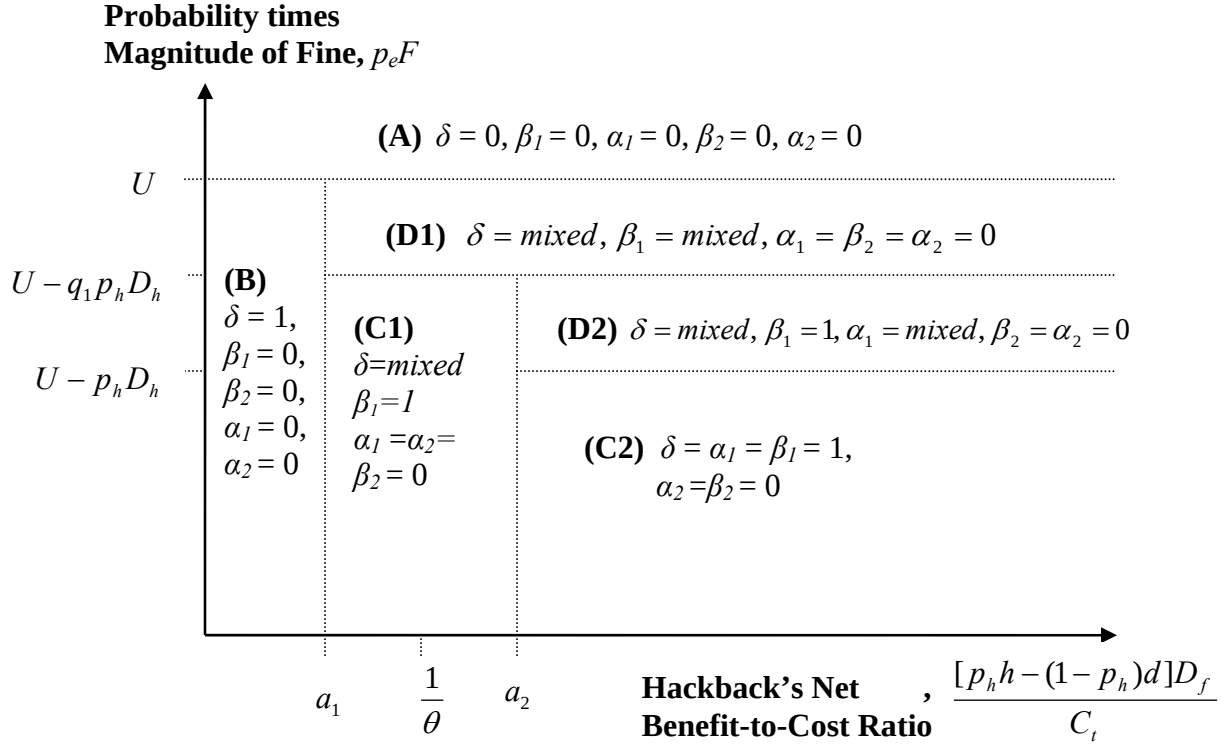


Figure A5. Nash equilibria when litigation is not beneficial (IDS available)

Proposition 4. When $a_1 < a_2 \equiv \dfrac{(1-q_1)\theta + q_2(1-\theta)}{(1-q_1)\theta} < \dfrac{p_w W - K}{C_t}$, the following

Bayesian Nash equilibria obtain:

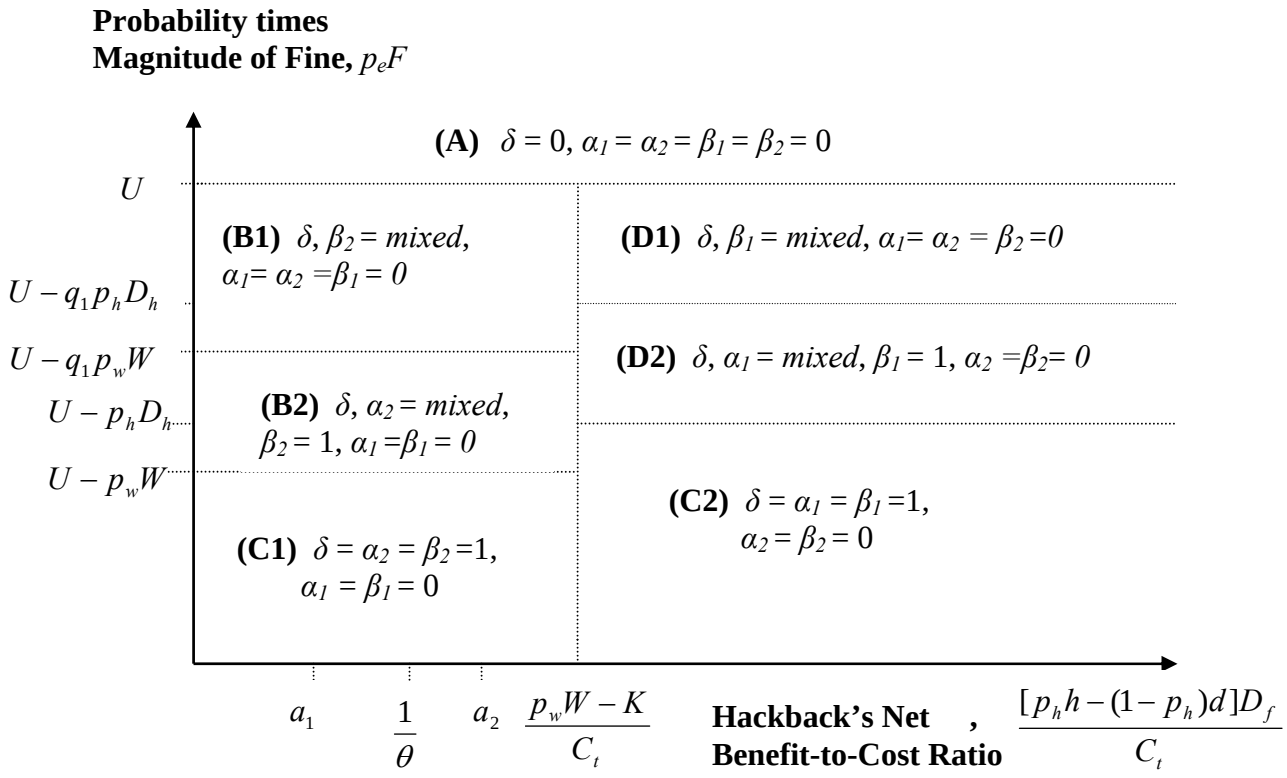


Figure A6. Nash equilibria when litigation is beneficial (IDS available)

Lemma 4. $\eta_1 > \delta$ (given $q_1, q_2 > 0.5$)

Proposition 5. When $a_1 < \dfrac{p_w W - K}{C_t} < a_2$, the following Bayesian Nash equilibria

obtain:

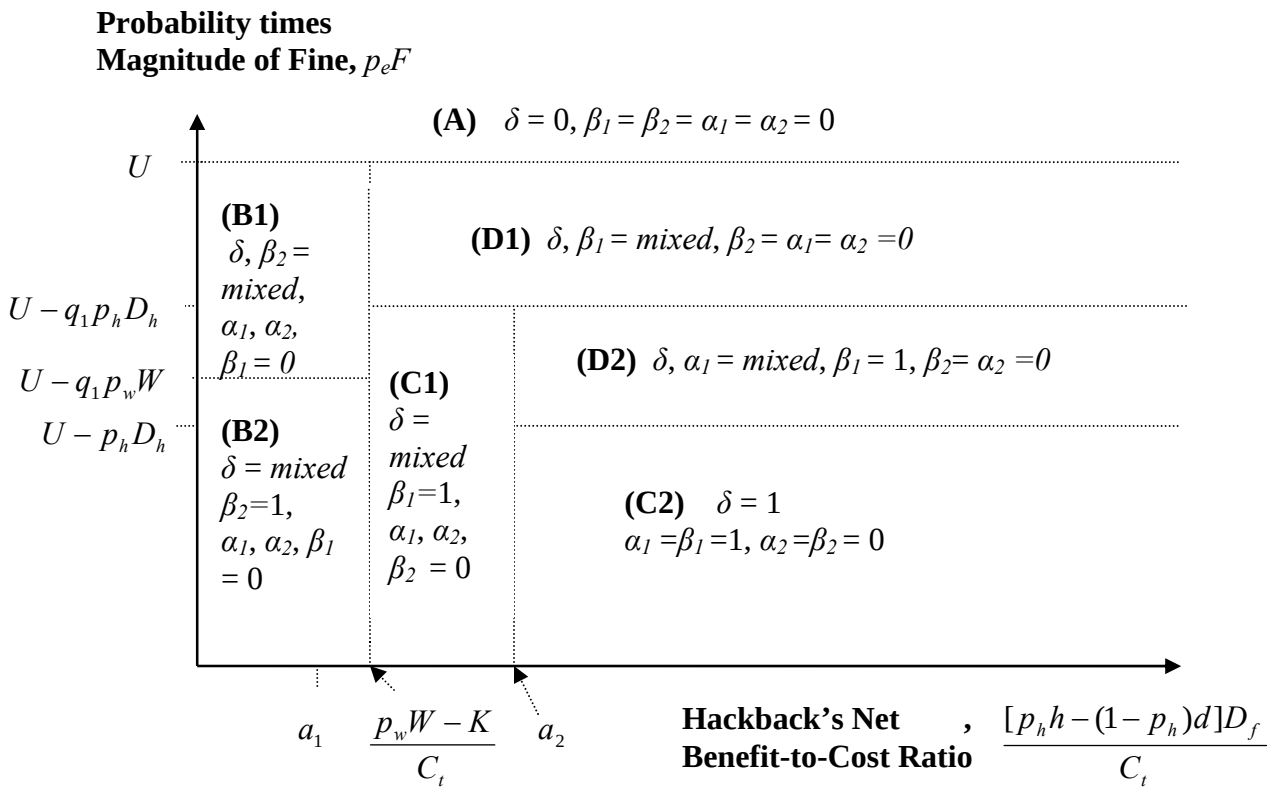


Figure A7. Nash equilibria when litigation is beneficial when the IDS signals an intrusion, but not otherwise

Socially-Optimal Solution

<u>If Hackback is Available</u>

Social planner's problem:

$$Max \quad S = -C_t(\sigma_1 + \sigma_2) - \theta\delta \begin{Bmatrix} \sigma_1[p_h(1-r-h)D_f + (1-p_h)(1-r+d)D_f] \\ +\sigma_2[(1-r)D_f + K - p_w W] \\ (1-\sigma_1-\sigma_2)(1-r)D_f \end{Bmatrix} \qquad \text{(A10)}$$
$$+ \quad \delta[U - p_e F - \sigma_1 p_h D_h - \sigma_2 p_w W]$$

Hacker's problem:

$$Max \quad \delta[U - p_e F - \sigma_1 p_h D_h - \sigma_2 p_w W] \qquad \text{(A11)}$$

Proposition 6: Nash equilibria:

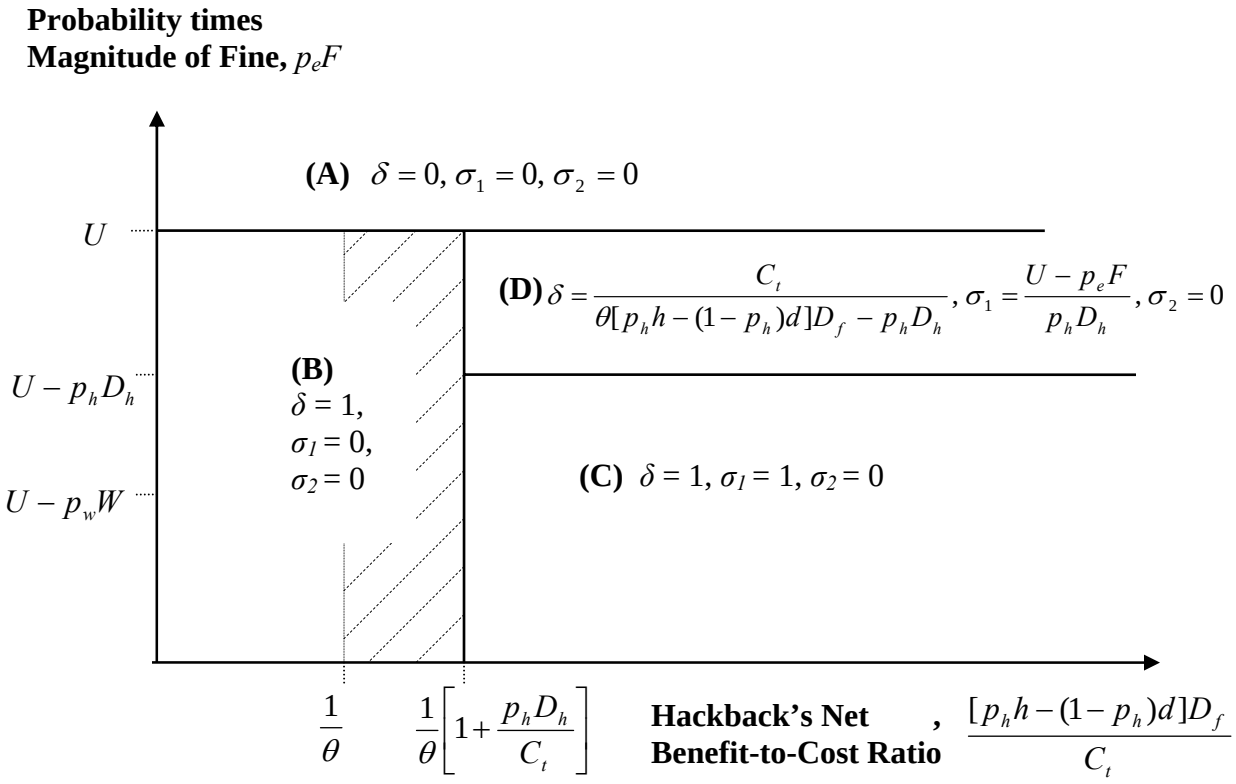


Figure A8. Nash equilibria of the social planner's problem, (No IDS available)

<u>If Hackback is Not Available</u>

Social planner's problem:

$$Max \quad S = -C_t\sigma_2 - \theta\delta\begin{Bmatrix}\sigma_2[(1-r)D_f + K - p_w W] \\ +(1-\sigma_2)(1-r)D_f\end{Bmatrix} + \quad \delta[U - p_e F - \sigma_2 p_w W] \qquad \text{(A12)}$$

Hacker's problem:

$$Max \quad \delta[U - p_e F - \sigma_2 p_w W] \qquad \text{(A13)}$$

Nash equilibria has simply 2 regions.

Social Welfare Comparisons

(omitted due to space limitations)

If the Firm is Not Held Liable for Third Party Damages

Firm's pay-off:

$$F = -C_t(\sigma_1 + \sigma_2) - \theta\delta\begin{Bmatrix}\sigma_1[p_h(1-r)D_f - p_h hD_f] \\ +\sigma_2[(1-r)D_f + K - p_w W] \\ +(1-\sigma_1-\sigma_2)(1-r)D_f\end{Bmatrix}. \qquad \text{(A14)}$$

Nash equilibria:

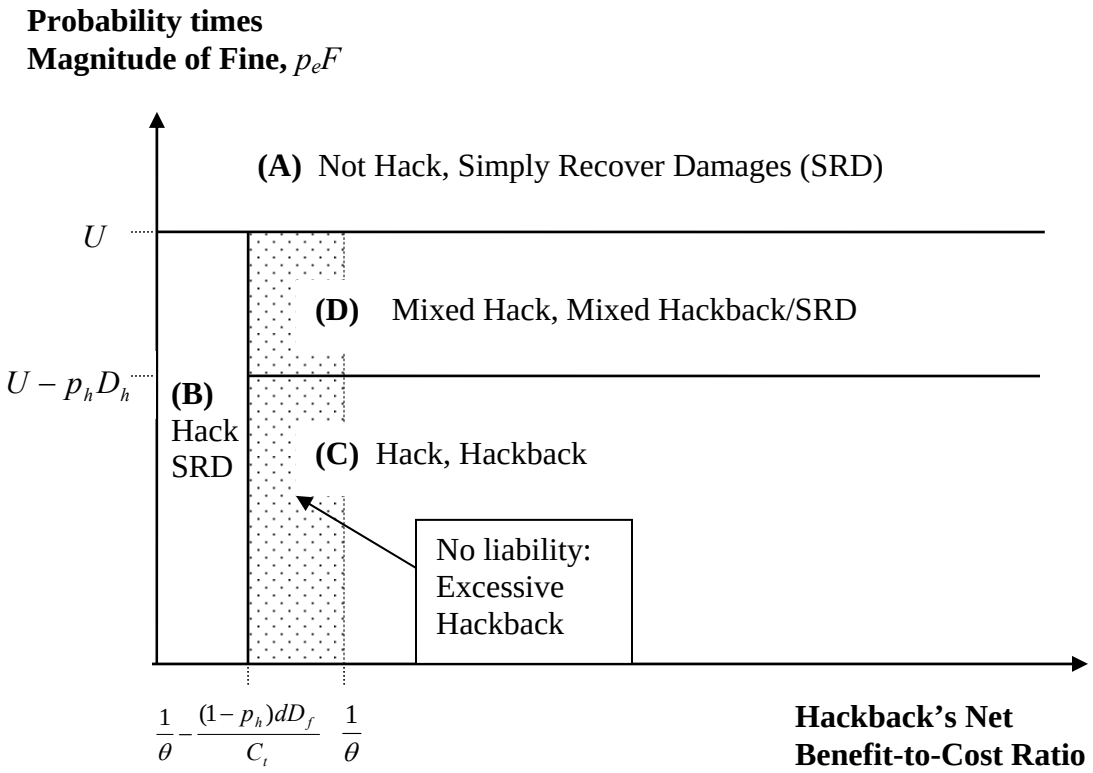


Figure A9. Nash equilibria of the firm's problem
(No liability rule)

Frictions in the Application of the Third Party Liability Rule

Due to frictions, only a fraction $1-\pi$ goes to court.

Nash equilibria:

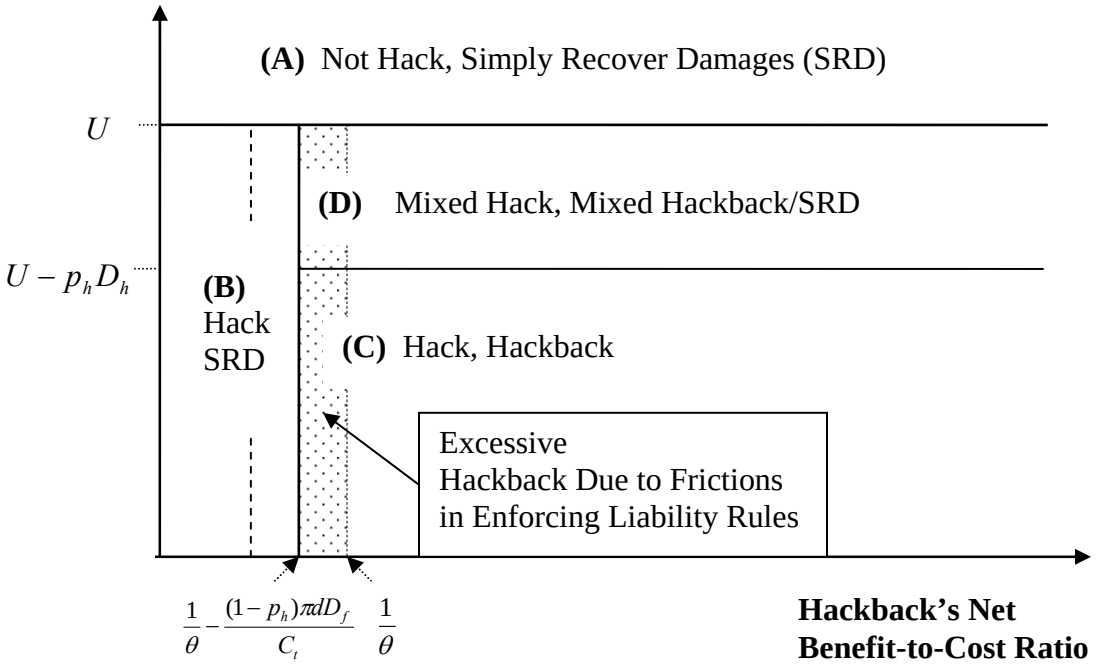


Figure A10. Nash equilibria of the firm's problem
(Frictions in enforcing liability rules)

REFERENCES

Aquinas, Thomas. c.1271. *Summa Theologica.* http://www.newadvent.org/summa/

American Law Institute. 1965. *Restatement (Second) of Torts.* Philadelphia: American Law Institute.

American Law Institute. 1985. *Model Penal Code and Commentaries (Official Draft and Revised Comments)*. Philadelphia: American Law Institute.

Augustinus, Aurelius**.** 400. *Contra Faustum Manichaeum.* http://www.newadvent.org/fathers/1406.htm.

Augustinus, Aurelius**.** c.423. *De Civitate Dei.* http://www.newadvent.org/fathers/120119.htm.

Brandon, Douglas Ivor, Melinda Lee Cooper, Jeremy H. Greshin, Alvin Louis Harris, James M. Head Jr., Keith R. Jacques, and Lea Wiggins. 1984. Special Project, Self-help: Extrajudicial Rights, Privileges and Remedies in Contemporary American Society. *Vanderbilt Law Review* 37:845-1040.

Cavusoglu, Huseyin, Birenda Mishra, and Srinivasan Raghunathan. 2005. The Value of Intrusion Detection Systems (IDSs) in Information Technology (IT) Security. *Information Systems Research*, forthcoming, March.

DeForrest, Mark Edward. 1997. Just War Theory and the Recent U.S. Air Strikes Against Iraq. Gonzaga University, School of Law. http://law.gonzaga.edu/borders/documents/deforres.htm

Epstein, Richard A. 2005. Self-help: From Stone Age to the Internet Age. *Journal of Law, Economics, and Policy* 1.

Grotius, Hugo. 1625. *Jure Belli ac Pacis ("On the Law of War and Peace").* http://www.geocities.com/Athens/Thebes/8098/

Himma, Kenneth Einar. 2004. Targeting the Innocent: Active Defense and the Moral Immunity of Innocent Persons from Aggression. *Journal of Information, Communication, and Ethics in Society* 2, no. 1.

Katyal, Neal. 2005. Community Self-help. *Journal of Law, Economics, and Policy* 1

Kesan, Jay P., and Ruperto P. Majuca. 2005. Cybercrimes and Cyber-Attacks: Technological, Economic, and Law-Based Solutions. In *Encyclopedia of Cybercrimes*, edited by Pauline C. Reich. Forthcoming: Oceania Press.

Lichtman, Douglas. 2005. How the Law Responds to Self-help. *Journal of Law, Economics, and Policy* 1.

Nash, John. 1950. Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences* 36:48-49.

Posner, Richard A. 1971. Killing or Wounding to Protect a Property Interest," *Journal of Law and Economics* 14:201-32.

Smith, Bruce P. 2005. Hacking, Poaching, and Counterattacking. *Journal of Law, Economics, and Policy* 1.

United States Catholic Conference. 1997. *Catechism of the Catholic Church*. 2d ed. Washington, DC: USCC Publishing Services.