The Hacker's Dilemma: Applying Game Theory to the Hacker-Enterprise Relationship

Dan Stuart, SI 510

Abstract

In the article "Immunizing the Internet, OR: How I Learned to Stop Worrying and Love the Worm" (Harvard Law Review Volume 119, June 2006, Number 8), the authors argue that certain kinds of cybercrime should be treated differently from other crimes because of cultural features of the hacking community, and because of the net benefit to companies from minimally-harmful hacking. From the perspective of game theory, the relationship between hackers and enterprises is analogous to the classic Prisoner's Dilemma game; analysis in terms of this scenario leads to strategic recommendations for policy-makers and enterprises to increase "benevolent" hacking. The results of this analysis support both the traditional law-enforcement solutions and the more lenient approach suggested by "Love the Worm".

The Prisoner's Dilemma

One of the classic strategic situations in game theory is called the Prisoner's Dilemma. A short story motivates the game:

Two criminals, Player 1 and Player 2, are accomplices in a bank robbery, and are later arrested. The police place them in separate rooms and offer them both the same plea-bargain deal, which is as follows. The police don't have enough evidence to prosecute either prisoner for the bank robbery, but they can convict each of them on a lesser charge. However, if one prisoner will testify against the other, he will receive a reduced sentence, and his accomplice will be convicted on the greater charge. If both testify against each other, both will be convicted, though with a reduced sentence. Each player has to choose, independently,

whether to defect from his accomplice or to cooperate. The possible outcomes of the game (in the "normal form") are shown below.

	C	D
C	$-5, -5$	$-20, 0$
D	$0, -20$	$-15, -15$

The strategies for Player 1 are the rows, and Player 2 plays the columns, and the payoffs for each player are shown in the corresponding square. So, for example, if Player 1 cooperates and Player 2 defects, this corresponds to row C, column D, and Player 1 gets a payoff of -20 (i.e. 20 years in prison), while Player 2 gets a payoff of 0 (he goes free). The actual payoffs, in general, are not as important as which outcomes each player prefers over the others[1].

The interesting feature of the Prisoner's Dilemma game is that each player can do better for himself by defecting, no matter what the other player does: if Player 2 cooperates and Player 1 defects (D,C), Player 1's sentence is reduced from 5 years to 0 compared to cooperating (C,C); but if Player 2 defects, Player 1 can reduce his sentence from 20 years (C,D) to 15 by defecting as well (D,D). Clearly, each player's best individual strategy is to defect, but if both players defect, they both get their second-worst result! Furthermore, if they both cooperate (C,C), they both do better than when they both defect. This conflict between individual and group interests is the heart of the Prisoner's Dilemma, and is one of the reasons it is a useful model for a large number of real-world strategic situations such as tariff-setting. (For a full discussion, see any introductory game theory text, such as *Strategy: An Introduction to Game Theory* by Joel Watson.)

The Hacker-Enterprise Relationship

The Prisoner's Dilemma model is applicable to many situations in which individual and group interests conflict in this way, including the strategic relationship between the hacker community and the enterprise community. For the purposes of brevity, "hacker" will herein refer to a skilled programmer who is able and motivated to breach computer security systems, legally or not, and "enterprise" will refer to companies, corporations, governments, and in general any organization that maintains and tries to protect its electronic resources. For simplicity,

[1]Player 1 prefers (D,C) over all other outcomes, then (C,C), then (D,D), and last (C,D). Player 2 prefers (C,D) over all other outcomes, then (C,C), then (D,D), and last (D,C).

the community of hackers and the community of enterprises will be treated as individual players, and it will be shown that the strategic interaction between the two can be modeled by the iterated (repeated) Prisoner's Dilemma.

First, consider the hacker player's choices. On discovering a security vulnerability in an enterprise's systems, he can choose between informing someone at the enterprise who can fix the flaw, or exploiting the flaw himself for personal goals (one or more of fame, fun, profit, or ideology). The hacker may also tell others who would exploit the flaw for their own purposes, or he may perhaps tell no one. In the abstract, the hacker in this position has two options: he can choose to do a large amount of harm to the enterprise (by exploiting the flaw, telling someone else who may do so, or even by telling no one, as the next hacker to discover it may not be so prudent); or he can do a small amount of harm by informing the enterprise of the flaw (perhaps by email or by executing a small proof-of-concept attack) or even writing a patch himself.

Next, consider the enterprise's options. Because no security system is perfect, and because of the large number of hackers in the world, it is inevitable that any large system will be targeted by hackers eventually, so responsible enterprises should have a policy in place for dealing with attacks. In the abstract, an enterprise can either aggressively prosecute any unauthorized access (via legislation, lawsuits, police action, or other actions as appropriate to the specific nature of the enterprise) in order to discourage hacking and preserve its reputation; or it can work with cooperative hackers to fix security breaches.

The above considerations can be compiled into the following normal-form table:

	Lenient	*Aggressive*
Cooperate	a, α	c, γ
Harm	b, β	e, ϵ

Here, the hacker is the row player and the enterprise is the column player. The hacker's payoffs are given by Latin letters, and the enterprise's payoffs by Greek letters. As above, only the order of each player's preferences matters, and not the actual values of the payoffs, so in this table it is assumed that $b > a > e > c$ and $\gamma > \alpha > \epsilon > \beta$. In the short run, a hacker is better off exploiting a hack for all its worth, while an enterprise is better off aggressively prosecuting any hackers. If both the hacker and the enterprise take a hostile stance, they will both do poorly, but both can do fairly well if the hacker and the enterprise cooperate. So indeed, a simplified version of the hacker-enterprise

relationship has all the features of the Prisoner's Dilemma.

Cooperation

Is it possible to encourage cooperation between the enterprise and hacking communities? In the short run, it is not. While an enterprise gains a net benefit in the long term from releasing a fix for a security vulnerability, in the short term its reputation (and sales) may suffer. If a hacker can quickly steal enough credit card numbers and make enough money from them, he can try to disappear with the money, never to be heard from again. However, there is more to the situation than the short-term benefits of the players; enterprises and hackers don't simply come together, play a single game, and then go their separate ways. In the real world, the hacker and enterprise communities play this game repeatedly; this fact changes the nature of the relationship, and admits further analysis.

Though the focus of this section is on the long term, it is still the case that players may, in general, place more value on present benefit than on future gain. That is, a future payoff of x may, in the present, be worth only δx, where $0 < \delta < 1$. The factor δ is called the discount factor, by analogy with economics. Assuming that the discount factor is applied once per round in the future, and that the players (enterprise and hacker) have agreed before the game to cooperate, it is possible to find the conditions in which neither player has an incentive to alter their cooperative strategy[2]. The rational player will of course have a plan for the case that his opponent decides to deviate from the cooperative agreement. Two such plans are known as "Tit-for-tat" and "Grim Trigger". In the tit-for-tat strategy (the most forgiving strategy aside from blind cooperation), the player reciprocates each of his opponent's actions on the next round. So, if the opponent deviates from cooperation on any round, the player deviates on the next round, but then goes back to cooperating on the round after that. Under the unforgiving Grim Trigger strategy, if the opponent deviates once, the Grim Trigger player deviates on all subsequent rounds. There are many other such strategies, but these are common examples for analysis.

[2]In this section, the shorthand terms "cooperate" and "deviate" are used, partly for convenience and partly by convention, to refer to each player's friendly and aggressive strategies, respectively.

First, consider the hacker's strategy. The hacker's total present payoff for cooperating in the present, and all future rounds, assuming that the enterprise player is also cooperating, is given by:

$$p_H(C) = a + \delta a + \delta^2 a + \ldots \tag{1}$$
$$= a(1 + \delta + \delta^2 + \delta^3 + \ldots) \tag{2}$$
$$= \frac{a}{1-\delta} \quad \text{For } \delta < 1 \tag{3}$$

The moves for this sequence of games are then: $\{(C,C),(C,C),(C,C),\ldots\}$.

What about the hacker's payoff for deviating on the first round, and cooperating afterward? That depends on the enterprise player's strategy. If the enterprise player is playing tit-for-tat, denoted below by T, then:

$$p_H(D,T) = b + \delta c + \delta^2 a + \delta^3 a + \ldots \tag{4}$$
$$= b + \delta c + \delta^2 a(1 + \delta + \delta^2 + \delta^3 + \ldots) \tag{5}$$
$$= b + \delta c + \frac{\delta^2}{1-\delta} a \tag{6}$$

The sequence of moves here is $\{(D,C),(C,D),(C,C),\ldots\}$. If the enterprise player is playing Grim Trigger, denoted by G, then:

$$p_H(D,G) = b + \delta c + \delta^2 c + \ldots \tag{7}$$
$$= b + \delta c(1 + \delta + \delta^2 + \delta^3 + \ldots) \tag{8}$$
$$= b + \frac{\delta}{1-\delta} c \tag{9}$$

Using equations (3), (6), and (9), it is simple to find the conditions under which the hacker player will do better by cooperating than by deviating, by finding the value of delta for which $p_H(C) > p_H(D)$. For the case of the enterprise player playing tit-for-tat:

$$p_H(C) > p_H(D,T) \tag{10}$$
$$\frac{a}{1-\delta} > b + \delta c + \frac{\delta^2}{1-\delta} a \tag{11}$$
$$a > (1-\delta)b + \delta(1-\delta)c + \delta^2 a \tag{12}$$
$$(c-a)\delta^2 + (b-c)\delta + (a-b) > 0 \tag{13}$$

The expression on the left is equal to zero for $\delta = 1$ and $\delta = \frac{b-a}{a-c}$. Since $b > a > c$, this second value is positive, and since $c - a < 0$, the quadratic

function in equation(13) is convex down, so the function is positive in the range $\delta \epsilon [\frac{b-a}{a-c}, 1]$. This means that the hacker an has incentive to continue cooperating in the tit-for-tat case if $\delta > \frac{b-a}{a-c} = \delta_T$ (since $\delta < 1$ always), and if $\frac{b-a}{a-c} < 1$. The same analysis applies to the enterprise player, so (substituting Greek letters for Latin, and exchanging β and γ[3]) if the hacker is playing tit-for-tat, cooperation can be sustained for $\delta > \frac{\gamma-\alpha}{\alpha-\beta}$ and $\frac{\gamma-\alpha}{\alpha-\beta} < 1$.

The Grim Trigger case is somewhat simpler. For the hacker:

$$p_H(C) > p_H(D, G) \tag{14}$$

$$\frac{a}{1-\delta} > b + \frac{\delta}{1-\delta}c \tag{15}$$

$$a > (1-\delta)b + \delta c \tag{16}$$

$$\delta > \frac{b-a}{b-c} = \delta_G \tag{17}$$

Since $b > a > c$, $0 < \frac{b-a}{b-c} < 1$, so cooperation is sustainable in the Grim Trigger case for $\delta > \frac{b-a}{b-c}$. As above, the same applies to the enterprise player (with the appropriate substitutions), so if the hacker is playing Grim Trigger, cooperation is sustainable if $\delta > \frac{\gamma-\alpha}{\gamma-\beta}$.

Results: Changing the Game

What do the above inequalities mean? While it was useful for the preceding analysis to consider the hacker community and the enterprise community as single players, this is definitely not realistic. Neither community acts as a collective, and neither is made up of members with the same attitudes and goals. While the goal of any anti-hacking program is to reduce the amount of (harmful) hacking, there are some hackers who will never be persuaded to hack for the collective good — it is unlikely that any modified risk/reward structure might have persuaded the Team Evil hackers who defaced several pro-Israel websites (see the report by Beyond Security here) to instead politely point out the site's vulnerabilities. Similarly, enterprises differ in their approaches to security and their willingness to work with benevolent hackers.

In the previous section, δ was defined somewhat loosely, but it is worth a closer look. In typical game-theory analysis, δ is used to quantify the value of

[3]The payoff for the enterprise being aggressive toward a cooperative hacker is γ, not β, but otherwise the algebra is the same.

future payoffs compared to payoffs in the present. It is normally considered a constant, a given value for an iterated game between two players. However, in view of the diversity of the hacker and enterprise communities, it is useful for this analysis to consider the δ as a measurement of an individual hacker's attitude toward risk, desire to cause harm, and other factors that make the hacker more or less likely to perpetrate a harmful hack. For consistency with the algebra above, δ-values are on a scale from 0 to 1, with 0 being a black-hat hacker who is determined to cause harm no matter the potential consequences, and 1 being a truly benevolent grey- or white-hat hacker who only wants to improve data security (and possibly find a good job in the process). The distribution of hackers on the delta scale will be treated in a later paragraph, but independent of the actual distribution, the results above show several possible ways to lower the threshold values of delta in the long term, and thereby increase the number of hackers willing to work for the common good, and decrease the amount of harmful hacking. For reference, here is the normal-form game once more:

	Lenient	*Aggressive*
Cooperate	a, α	c, γ
Harm	b, β	e, ϵ

Recall from above that for the hacker player $\delta_T = \frac{b-a}{a-c}$ and $\delta_G = \frac{b-a}{b-c}$; what parameters can be adjusted to lower each of these thresholds? For the tit-for-tat case, δ_T can be lowered by increasing a [4], decreasing b, or decreasing c, all in a way consistent with the condition $b > a > c$. For the Grim Trigger case, δ_G can be lowered by increasing a, decreasing b [5], or decreasing c, just as for the tit-for-tat case. These parameters correspond to the hacker's payoffs for the strategy profiles (Cooperate, Lenient), (Harm, Lenient), and (Cooperate, Aggressive)[6], respectively.

Based on this analysis, there are two changes that enterprises can make to decrease the proportion of harmful hackers: increasing the rewards for cooperative hackers (a), or by trying to decrease the payoff to hostile hackers (b). Which of these measures would be most effective depends on the actual distribution of hackers in δ-space; one possible measurement of this distribution comes from

[4] This is because $\frac{\partial \delta_T}{\partial a} = \frac{b-a}{(a-c)^2} - \frac{1}{a-c}$, which is negative for all valid values of a, b, and c.

[5] This is because $\frac{\partial \delta_G}{\partial b} = \frac{b-a}{(b-c)^2} + \frac{1}{b-c}$, which is positive for all valid values of a, b, and c.

[6] The c parameter, which corresponds to the (Cooperate, Aggressive) result, only applies after the hacker player has broken the agreement, and so doesn't make sense to consider from a policy standpoint as the damage is already done.

the Verizon Business 2009 Data Breach Report. Figure 28: "Distribution of breach size by number of records", on page 35 of this document, shows a strong peak (39%) in the 10,000–100,000 bin, and 71% of breaches fall into this bin or below. These are certainly large attacks, but not the largest, and this figure suggests that the distribution of hackers in δ-space is dense in the middle — thus any attempt to either encourage or coerce hackers toward the gray-hat end of the scale would show a strong increase in effectiveness at a certain level of implementation.

The methods to decrease the payoffs for both types of hacker are obvious: aggressive pursuit by law enforcement, legislation, and longer prison terms would be likely to work. However, methods to increase the rewards for cooperative hackers may need to be somewhat more creative. As "Love the Worm" argues, white-hat hacking could be encouraged by cash incentives, a safe-harbor program providing immunity to prosecution for past crimes, or even simply a reputation-boosting recognition of a job well-done could all incentivize hackers to move toward more benevolent activities.

What can be done to increase the cooperativeness of enterprises? Recall $\delta_T = \frac{\gamma-\alpha}{\alpha-\beta}$ and $\delta_G = \frac{\gamma-\alpha}{\gamma-\beta}$ for the enterprise player. As above, it is reasonable to assume that the enterprises are distributed throughout δ-space (though most enterprises are presumably motivated by profit rather than idealism, so the distribution should be rather sparse near the ends); then the δs can be decreased by increasing α or by decreasing γ[7]. So it should be possible to increase the proportion of cooperative, hacker-friendly enterprises by either increasing the rewards for being lenient (α), or by decreasing the rewards for being aggressive (γ).

Being lenient toward benevolent hackers is its own reward: if an enterprise is known for cooperating or being generous with gray-hat hackers (increasing a), more hackers will be likely to try to do that company a good turn in the hopes of a reward, thereby increasing α as well. Similarly, being hostile to the hacker community may simply increase hostile hacks, decreasing γ, but it is more orderly to decrease γ legislatively. It would be sensible (if this is not already the case) to legislate that convicted hackers are not liable for the costs of upgrading compromised systems, if those systems should have been upgraded in the first

[7] As above, the β parameter only applies after the enterprise player has broken the cooperation agreement, so it will not be considered.

place (i.e. in cases other than zero-day exploits). This "feedback" feature, if it applies in the real world, is a powerful means of altering the nature of the hacker-enterprise relationship. Such a feedback would result from the actions of the hacker player as well, but as the enterprise community is of necessity more organized, it would be easier and more efficient to affect the strategic relationship by changing policy on the enterprise side.

Analyzing the relationship between hackers and enterprises using game theory supports both the traditional criminal solution of harsher penalties and stricter enforcement, and the more cooperative strategy suggested by "Love the Worm". While the model is necessarily simplified, and it is difficult to estimate the effectiveness of either strategy without a better idea of the current values of the many parameters involved, the criminal solution is more expensive and time-consuming than encouraging the culture of cooperation that already exists in the IT world. Though lowering the penalties for the most damaging attacks would be foolish, over-deterring the comparatively harmless attacks causes the enterprise community to miss out on the overall benefits of those incidents.