Deterrence and Cheap Talk or: How we Learned to Stop Worrying and Love the Signal

Jonathan Welburn, Justin Grana and Karen Schwindt *

November 5, 2019

Abstract

Motivated by the asymmetric information inherent to cyberwarfare, we examine a game of deterrence between an attacker and a defender in which the defender can signal its retaliatory capability but can only imperfectly attribute an attack. We show that there are equilibria in which the defender sends noisy signals to increase its expected payoff. In some equilibria, the defender can use signaling to deter an attacker and increase its payoff. In a different and somewhat counter-intuitive equilibrium, the defender can increase its expected payoff through signaling by inducing the attacker to attack more.

Defensive cyber security best practices have proved to be insufficient for protecting both public and private assets. Cyber aggression against Sony Pictures, the U.S. Office of Personnel Management, the Central Bank of Bangladesh, the Germain Parliament, and ransom-ware attacks WannaCry and NotPetya represent only a small sample of cyber attacks that led to substantial political and economic disruptions and costs. More generally, the threat of cyber attacks against key institutions and critical infrastructure has outpaced defensive efforts to reduce vulnerabilities [8]. As a result, the focus of international cyber defense has shifted toward deterrence. For example, the 2019 National Defense Authorization Act (NDAA) specifically calls for such a U.S. cyber deterrence policy:

> "It shall be the policy of the United States, with respect to matters pertaining to cyberspace, cybersecurity and cyber warfare, that the United States should employ all instruments of national power, including the use of offensive cyber capabilities, to deter if possible, and respond to when necessary, all cyber attacks or other malicious cyber activities [1].

However, traditional deterrence [25, 20] relies on numerous assumptions that in new domains of attack — especially computer networks—are no longer valid. Consider the following two assumptions that are central to classic deterrence theory:

- "General deterrent threats are likely to be more effective when a potential challenger views them as capable [15]." or in other words deterrence requires "the possibility of a clear demonstration of the defender's capabilities [18]."
- "The deterring state must first know who to counterattack [11]."

When considering cyberwarfare, neither of these assumptions are likely to hold. First, it is unlikely that potential attackers know their target's retaliatory capability. The reason is that "cyber weapons rely largely on previously unknown, so called zero-day, vulnerabilities" and thus demonstrating a capability to a potential attacker renders the capability ineffective [29, 5]. Second, properly attributing a cyber attack is a recognized difficult problem due to both the technical acumen required to conduct forensic analysis and the ease in which an attacker can deliberately obscure its identity [26]. These complexities are not limited to digital interactions; imperfect attribution and signaling are also gaining relevance in domains such as traditional warfare and international relations [24, 21, 17] as key elements of deterrence.

Nevertheless, the new tenants of deterrence have not quelled the threat of aggression and retaliation. In addition to the 2019 NDAA where the United States promises to "respond when necessary" major world

*RAND Corporation, 1776 Main St, Santa Monica, CA 90401.{jwelburn, jgrana, schwindt}@rand.org, Acknowledgements

superpowers have made similar retaliatory threats. For example in the 2019 French cyber strategy from the Ministre Des Armées states "We will also be ready to use the cyber weapon in external operations for offensive purposes, alone or in support of our conventional means [19]." Chinese military strategy documents have made similar threats: "A high-level Chinese military organization has for the first time formally acknowledged that the country's military and its intelligence community have specialized units for waging war on computer networks [12]." All told, despite the unverifiable nature of these cyber threats, world powers are still publicizing their intentions to use cber weapons when necessary.

These new features of modern deterrence scenarios demand a formal and rigorous treatment. Such an exercise would provide a needed foundation for the growing literature focused on the feasibility of deterrence in cyber space [16, 14, 13] and establish a standard for expanding traditional deterrence theory. To address this need, we develop a model of an attacker and defender with three main features designed to bridge the gap between traditional and modern deterrence theory:

1. The defender can only imperfectly attribute attacks.
2. The attacker has uncertainty over the defender's retaliatory and defensive capability.
3. The defender can signal its capability not by revealing its true capability but through costless and unverifiable cheap talk.

Our focus is on the relevance and importance of item 3). Specifically, since verifiable signaling is unlikely in many domains, including cyber, we examine whether signaling via cheap talk can be effective in deterring adversaries.

The results of our formal analysis illuminates at least four key insights regarding signaling. First, there is no separating equilibrium in which the defender always noiselessly signals its true retaliatory capability. The reason is that if a defender could convince an adversary that it is indeed signaling its true capability, then the defender would have an incentive to always signal a strong retaliatory capability. Second, there are several babbling equilibria in which the defender's signal provides no information regarding its true capability. While not intrinsically interesting, these babbling equilibria provide a baseline of comparison for any potential signaling equilibria. Third, there exists semi-separating equilibria in which the defender a) releases noisy signals regarding its true retaliatory capability and b) increases deterrence through a reduction in the attack probability relative to a babbling equilibrium and c) increases its expected utility relative to a babbling equilibrium. Or simply put, signaling can be used to increase deterrence.

The fourth and arguably most surprising result is that in some parameter regimes, there exists a semi-separating equilibrium in which the defender increases its expected utility over a babbling equilibrium by inducing the attacker to *increase* the probability of attack. The reasons for this counter-intuitive result are two-fold. First, an increase in the attack probability reduces the frequency in which the defender is punished for an incorrect retaliation. Secondly, the defender can use its signal to induce the attacker to attack when the defender has a higher defensive and retaliatory capability. This result, which we call "anti-deterrence," adds a new consideration to the conversation around cyber deterrence. In contrast to the current discussion that mainly asks "is cyber deterrence possible?" the results of our model suggest that an equally important question is "should cyber deterrence be the goal?"

In addition to work explicitly focused on deterrence theory, our work is related to and partially synthesizes the vast literature on entry deterrence in industrial organization, criminal deterrence and attacker-defender games in the broader context of signaling games. However, our work is undoubtedly most related to [4]. Their model — also motivated by cyber warfare — has a single defender and n possible attackers. The attackers choose whether to attack and the defender receives a noisy signal and chooses whether to retaliate against one or more attackers. The main finding is endogenous complementarity among attackers where increasing aggression from the most aggressive attacker incentivizes increasing aggression from all others. Furthermore, jointly enhancing attack detection and attacker identification (which they jointly refer to as attribution) strengthens deterrence but enhancing only one of either attack detection or identification may weaken deterrence. Our model builds directly from theirs except that we limit our model to one attacker in order to better focus on the impact of signaling [1].

[1] Interestingly, [4] note that an interesting direction of future work would be to ask "Would the ability to signal cyber-capability lead to coordination on a peaceful equilibrium?" While we developed our model before being made aware of [4], the fact that researchers are independently converging on similar topics points to the timeliness of the topic.

Our work is also related to attacker/defender games, especially those that occur in the cyber domain. These games include element of imperfect attribution and blame [9], multiple attackers [30], and signaling [31]. While the full literature of attacker/defender games is vast and extends back to Blotto games [6], there are several large surveys on attacker and defender games and their applications to computer network security [27, 22, 32]. We emphasize however that our model extends beyond just cyber security and we largely abstract from technical details that are particular to cyber attacks.

Outside of warfare and defense, deterrence and signaling is a prominent topic in the industrial organization literature, specifically in regard to market entry. Classic work in this field is concerned with undertaking costly investments to deter a potential market entrant and maintain monopoly status [7, 23, 10]. These models have evolved to include various forms of signaling [28, 3]. Finally, the concept of deterrence and punishment is prevalent in the economics of crime literature [2].

1 Model Outline

We consider a two-player sequential-move game of imperfect information between an attacker and a defender. At the start of the game, nature assigns the defender either a "high" or "low" type, signifying its retaliatory capabilities. In the model, the defender knows its capability with certainty while the attacker does not. Instead, the attacker only knows the probability with which nature assigns the defender's retaliation capability. If our game is interpreted as one instance of an infinitely repeated game, the defender's capability fluctuating between high and low can come about due to the dynamics of bugs and exploits being discovered and patches subsequently released.

After the defender realizes its capability, it chooses how to signal its capability to the attacker. The defender can either signal that it has a high capability or a low capability. We do not place any restrictions on these signals and there is no cost to signaling. That is, regardless of what the defender's true capability is, it can costlessly signal any capability.

Next, the attacker perfectly observes the defender's signal and then chooses whether or not to attack. The attacker's decision to attack is binary and it can only condition its decision on the signal it received from the defender and not the defender's true capability.

Following the attacker's decision to attack, the defender receives a signal that is correlated but *not* perfectly correlated with the attacker's action. As a realistic example, it is possible that an attacker initiates an attack but the defender's threat detection software never notices the attack and thus the defender receives a signal that it is not under attack. This represents an undetected attack. On the other hand, it is possible for the attacker to choose not to attack but the defender receives a signal that it is under attack. This represents a false alarm or possibly an attack by an exogenous and unmodeled attacker. This signal generating process captures the imperfect attribution aspect of the model. That is to say, even when the attacker chooses to attack, the defender does not know with certainty whether it was actually attacked. We note that this signal generating procedure is the same as in the one attacker case of [4].

After observing the signal, the defender moves next by choosing whether to retaliate against the attacker. There is no restriction on the defender's actions conditional on the signal. So for example, a real-world defender can choose to retaliate even when it didn't receive a signal that it was attacked. This might happen if a defender knows that its detection capabilities are poor and it is likely that an attacker chose to attack and subverted detection methods. Similarly, a defender can forego retaliation even if it receives a signal that its systems are under attack.

The payoffs depend on the defender's capability, the attacker's decision to attack and the defender's decision to retaliate. The attacker incurs a reward for attacking but also incurs a cost if the defender retaliates. If the attacker does not attack but the defender retaliates anyway, the attacker still incurs a cost. The attacker incurs a higher cost of retaliation from a defender that has a high capability. The defender incurs a cost when it is attacked but receives a small benefit (less than the cost of being attacked) for correctly retaliating. The defender incurs a cost if it incorrectly retaliates against the attacker. That is, if the attacker chooses not to attack but the defender retaliates, the defender incurs a cost. If the attacker does not attack and the defender does not retaliate, neither player incurs rewards or costs.

There are several justifications as to why a defender would receive a benefit from retaliating. One reason, as noted in [4] is that a retaliation can be reinterpreted as stopping an ongoing attack. Another source of

benefit is that there may be political pressure to retaliate after an attack. Yet another motivation for the defender receiving a benefit by retaliating is to establish a long-run reputation as a player that is willing to retaliate, which may deter other potential attackers in the future.

A key assumption in our model is that a defender with high capability both delivers a stronger strike to the attacker and also receives a higher benefit for a correct retaliation than a defender of low capability. This assumption is justified in several ways. First, once again, if a retaliation is thought of as stopping an ongoing attack, then a defender that delivers a strong strike to the adversary does more damage to the adversary's systems and thus is more effective at stopping the ongoing attack. However, an additional interpretation is that the damage to the adversary and the benefit to the defender is independent of the defender's type but the probability of a successful retaliation is higher for a defender of high capability. Therefore, the parameters describing the defender's benefit from a correct retaliation and the harm inflicted on the attacker can be interpreted as the *expected* benefit and harm.

2 Model Specification

The game has two players, a and d, and a nature player to capture stochastic elements. In the first stage of the game, nature chooses the type of the defender. The defender is of type H—representing high capability—with probability γ and is type L with probability $(1-\gamma)$.

After the defender is assigned its type, it signals either s_H or s_L. Specifically, the defender's *pure* strategy at this stage is a mapping from $\{H, L\}$ to $\{s_H, s_L\}$. Therefore, the defender's mixed strategy is a mapping $F : \{H, L\} \to [0, 1]$. That is, the defender chooses the probability for which it signals a high capability for each of its possible types. This function can be represented by two real numbers α_H and α_L. Specifically, α_H is the probability that the defender signals s_H— a high capability signal—given it was assigned a high capability and α_L is the probability the defender signals s_H given that nature assigned it a low capability. Analogously, $(1-\alpha_H)$ and $(1-\alpha_L)$ is the probability that the defender signals s_L—a low capability signal—given that nature assigned it a high and low capability, respectively.

After the defender's signal, the attacker observes the signal and chooses whether or not to attack. The attacker's *pure* strategy is then a mapping from $\{s_H, s_L\}$ to $\{A, DA\}$. and thus the attacker's mixed strategy is a mapping $G : \{s_H, s_L\} \to [0, 1]$. Intuitively, the attacker's mixed strategy assigns the probability of attack, A, conditional on the signal that it received. This strategy can be represented by two real numbers β_H and β_L where β_H is the probability the attacker chooses A given that it received the signal s_H and β_L is the probability that attacker chooses A given it received the signal s_L. Of course, $(1-\beta_H)$ and $(1-\beta_L)$ is the probability the attacker doesn't attack (chooses action DA), given it received signal s_H and s_L, respectively.

After the attacker's action is drawn according to the attacker's mixed strategy, the defender's observation is drawn by nature. Specifically, the defender either observes o_1 or o_2 but the probability of each signal depends on the attacker's action. Specifically, if the attacker chooses to attack, the defender observes o_1 with probability π_1 and o_2 with probability $(1-\pi_1)$. If the the attacker does not attack, then the defender observes o_1 with probability π_2 and and o_2 with probability $(1-\pi_2)$. Intuitively, π_1 and π_2 represent the defender's ability to attribute an attack. For example, if $\pi_1 = \pi_2$, then the signal does not depend on the attacker's action and the defender does not learn anything from the signal. If $\pi_1 = 1$ and $\pi_2 = 0$, then the defender can perfectly attribute attacks. Without loss of generality, we assume that $\pi_1 \geq \pi_2$.

Finally, the defender must choose whether to retaliate given it's observation. The defender's pure strategy maps its capability, observation and the signal it sent to an action $\{R, DR\}$ (R for retaliate and DR for don't retaliate). That means that the defender's mixed strategy is a function $F : \{o_1, o_2\} \times \{s_H, s_L\} \times \{H, L\} \to [0, 1]$. This strategy represents the probability that the defender retaliates—chooses action R—given its signal, observation and type. Let $\rho(x, y, z)$ be the probability the defender retaliates after observing observation x, signaling y and having type H. For example $\rho(o_1, s_H, H)$ is the probability that a defender of high capability that signaled s_H and observed o_1 chooses to retaliate. Let ρ (without subscripts) be shorthand for the set of $\rho(x, y, x)$ in the defender's strategy that give the retaliation probabilities.

The payoffs depend on the attacker's action, the defender's capability and the defender's choice to retaliate. If the attacker attacks, it accrues payoff of 1. However, if the defender retaliates, the attacker incurs a cost of c_H if the defender has high capability and c_L if the defender has low capability. If the attacker does not attack but the defender retaliates, we assume the attacker incurs a cost of v, regardless of the defender's

type. Since a defender of high capability is more able to punish, we assume $c_H > c_L$. We also assume that $c_H > 1+v$. This assumption implies that when the defender has high capability, the attacker would prefer to not attack and not be retaliated against over attacking and incurring a retaliation. Secondly, we assume that $c_L > v$. This means that the cost to being correctly retaliated against is always worse than being incorrectly retaliated against. For technical convenience, we assume that $1-\pi_1 c-\pi_2 v \neq 0$ and $1-\pi_1 c-\pi_2 v \neq 0$. This is an innocuous assumption that allows us to ignore sets of parameters that have measure 0.[2]

For the defender, if it is attacked it incurs a cost of -1. If it correctly retaliates, it earns r_H if it is high capability and r_L if it is low capability. We assume $r_H > r_L$. We also assume that $r_H, r_L < 1$ which means that the defender would rather not be attacked than be attacked and correctly retaliate. If the defender retaliates when the attacker didn't actually attack, it incurs a cost of w. If there is no attack and no retaliation, both players earn 0. The extensive form version of the game is given in figure 1 which illustrates the sequence of events, the information sets and the payoffs.

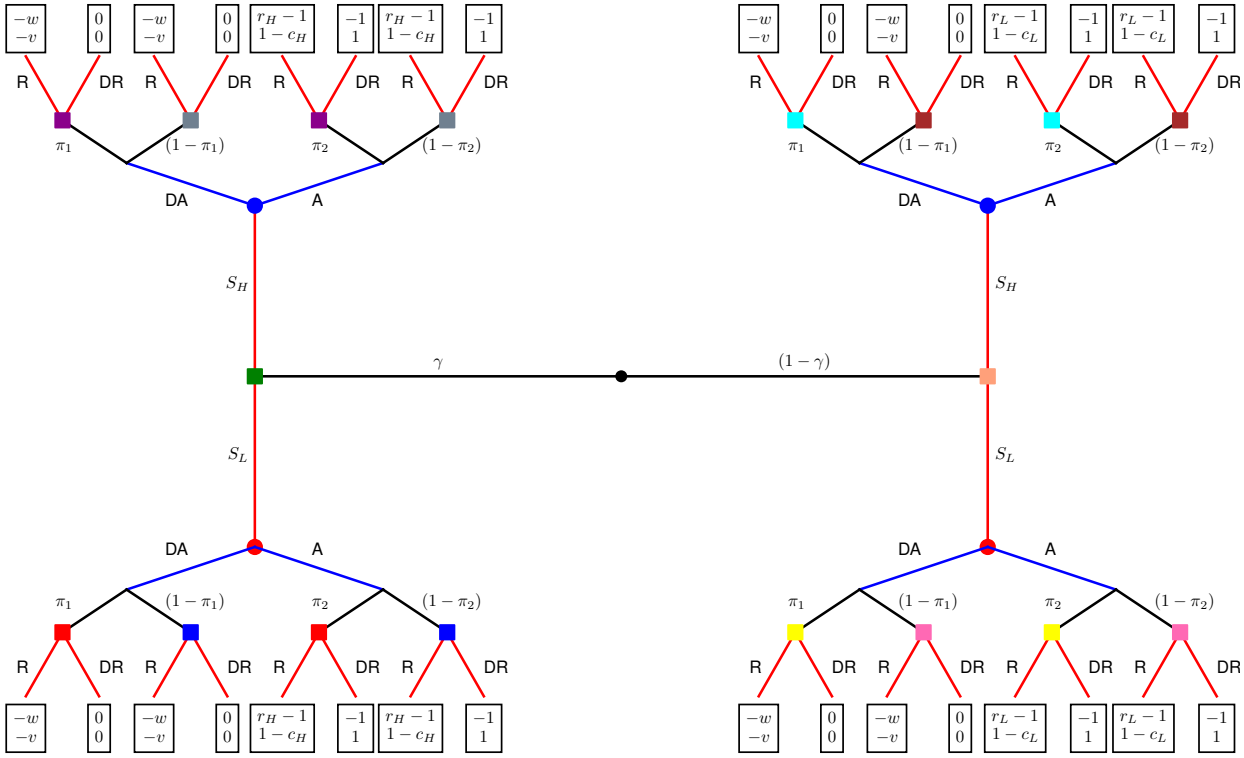


Figure 1: Extensive form representation of the signaling game. "Circle" nodes are attacker nodes and "square" nodes are defender nodes. Nodes of the same color are in the same information set. Probabilities for the nature player are given in Greek letters and actions for the attacker and defender are given in Latin letters.

The solution concept we will use to analyze this game is the Perfect Bayesian Equilibrium (henceforth equilibrium). Under such a solution concept, player's strategies are optimal given their beliefs and beliefs are derived using Bayes rule wherever possible. We do not need any further equilibrium refinement because our main result holds for all possible actions off of the equilibrium path.

[2]A more technical justification for this assumption is to assume that at least one of the parameters are drawn from a continuous probability distribution at the start of the game and all players observe the parameter value. Then the probability that $1-\pi_1 c-\pi_2 v=0$ is exactly zero and such a case can be ignored without changing the fundamental nature of the game.

3 Results and Analysis

To more cleanly present the results, we first analyze an **attribution game** and then analyze the associated **signaling game**. The attribution game is the same as the game described above except the defender's capability is fixed and common knowledge to the attacker and the defender (this would happen if, for example, $\gamma = 1$ or $\gamma = 0$). This renders signaling unnecessary since the attacker knows the defender's capability with certainty. Therefore, in the attribution game the sequence of events are: 1) the attacker chooses whether or not to attack, 2) the defender receives a signal that is correlated with the attack and then 3) the defender chooses whether or not to retaliate. After establishing intuition in the attribution game, we return to the full signaling game in which the defender's capability is not common knowledge and the defender can signal.

3.1 The Attribution Game

Since in the attribution game we assume that the defender's capability is common knowledge in the attribution game, we drop subscripts and let r be the reward the defender receives from correctly retaliating and c be the cost to the attacker from being retaliated against after attacking. In the proofs, we assume that $1 - c < -v$ in the attribution game. Otherwise, there is a trivial equilibrium where the attacker always attacks and the defender always retaliates. However, when considering the full signaling game later, a key equilibrium arises when $1 - c_H < -v$ but $1 - c_L > -v$, which is obscured in the attribution game. All formal proofs are given in the appendix.

Proposition 1 (Equilibrium in the Attribution Game)**.** *Suppose $1 - c < -v$. Let β be the probability the attacker attacks in the attribution game. Then:*

1. *If $1 - \pi_1 c + \pi_2 v < 0$, there exists an equilibrium of the attribution game where the attacker randomizes with probability $\beta_1^* = \frac{\pi_2 w}{\pi_1 r + \pi_2 w}$ and the defender never retaliates after observing o_2 and randomizes between retaliating and not retaliating after o_1 with probability $\rho_1^* = \frac{1}{\pi_1 c - \pi_2 v}$.*
2. *If $1 - \pi_1 c + \pi_2 v > 0$, there exists an equilibrium of the attribution game where the attacker randomizes with probability $\beta_2^* = \frac{(1-\pi_2)w}{(1-\pi_1)r + (1-\pi_2)w}$ and the defender always retaliates after observing o_1 and randomizes between retaliating and not retaliating after o_2 with probability $\rho_2^* = \frac{1 - \pi_1 c + \pi_2 v}{(1-\pi_1)c - (1-\pi_2)v}$.*

Corollary 1 (Uniqueness of Equilibrium in Attribution Game)**.** *The equilibria in proposition 1 are unique*

Proposition 1 and corollary 1 establish that there is a unique equilibrium in the attribution game but the nature of the equilibrium depends on the value of the parameters. To better understand the equilibrium, first consider why it is impossible for there to be an equilibrium in pure strategies. If the attacker always attacks, then the defender has a best response to retaliate, regardless of its signal. However, if the defender always retaliates, the attacker is better off not attacking. Similarly, from the perspective of the defender, if it chooses the pure strategy of never retaliating, the attacker's best response is to always attack, in which case the defender would have a profitable deviation to always retaliate. Therefore, there cannot be a pure strategy equilibrium.

A necessary condition for a mixed strategy equilibrium is that both the defender and the attacker are indifferent among at least two of their strategies. The defender only has to consider three out of its four pure strategies:

1. Always retaliate
2. Never retaliate
3. Retaliate after o_1 and don't retaliate after o_2.

The strategy "Retaliate after o_2 and don't retaliate after o_1" is dominated because for any fixed value of attack probability, β , Bayesian beliefs necessitate that an attack was more likely if the defender observes o_1 then if it observed o_2. Therefore, retaliating after o_2—when the defender is less certain there was an attack—and not retaliating after o_1—when the defender is more certain there was an attack—is a dominated strategy.

Figure 2 illustrates the defender's expected payoff U_d for each of its three strategies as a function of the attack probability, β. The purple line extending from the origin is the defender's expected utility from never retaliating. The blue line with an intercept at $-\pi_2 w$ is the defender's expected utility from retaliating after observing o_1 and not retaliating after o_2. The red line is the defender's expected utility from always retaliating. Finally, the gray shaded line outlines the defender's best response for each value of β. Specifically, for $\beta < \frac{\pi_2 w}{\pi_1 r + \pi_2 w}$, the defender's best response is to never retaliate. For $\frac{\pi_2 w}{\pi_1 r + \pi_2 w} < \beta < \frac{(1-\pi_2) w}{(1-\pi_1) r + (1-\pi_2) w}$, the defender's best response is to retaliate only after observing o_1. Finally, for $\beta > \frac{(1-\pi_2) w}{(1-\pi_1) r + (1-\pi_2) w}$, the defender's best response is the always retaliate. The legend in the figure lists the equations of each of the lines.

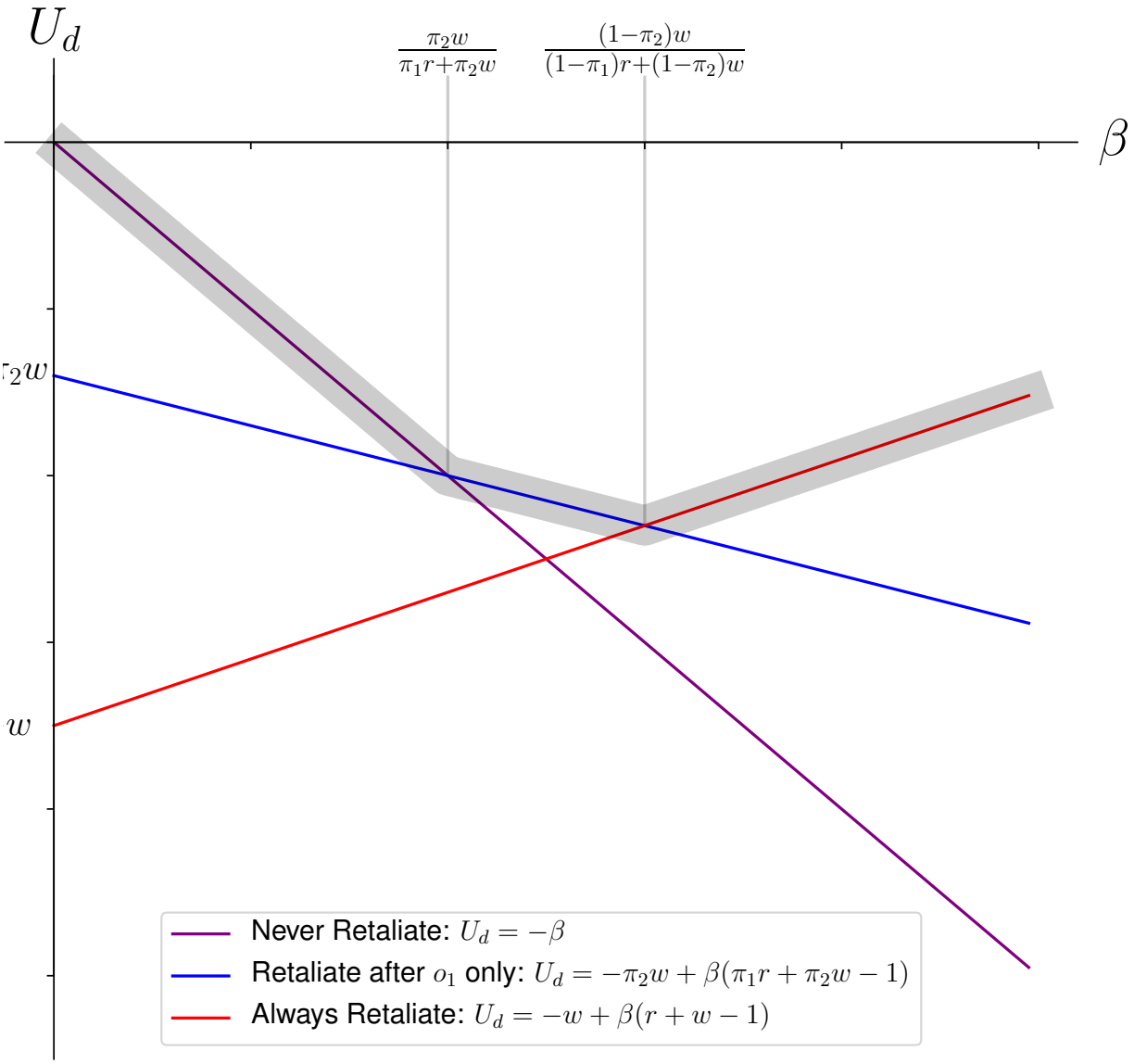
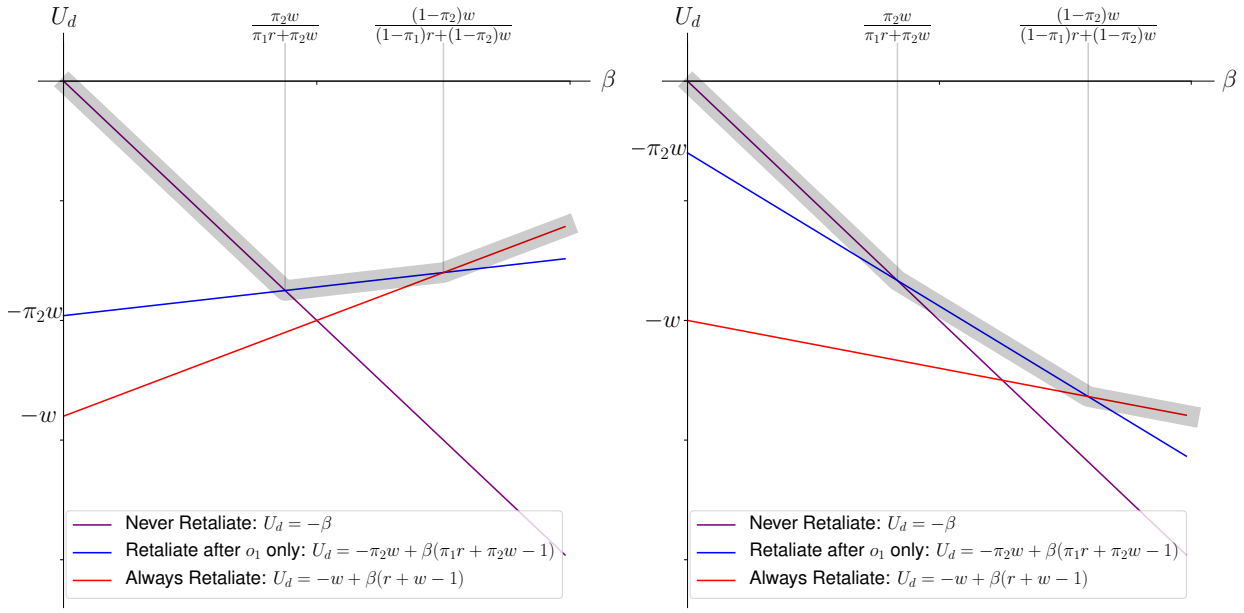


Figure 2: Defender's expected utility for each of its strategies

.



(a) $\pi_1 r + \pi_2 w > 1$ (b) $r + w < 1$

Figure 3: Two different versions of figure 2 with different slopes for defender strategies.

The slope of the blue line and the red line in figure 2 are determined by $\pi_1 r + \pi_2 w - 1$ and $r + w - 1$, respectively. By assumption, these are never 0. However, there is no restriction on their sign (except that $\pi_1 r + \pi_2 w - 1 < r + w - 1$). Therefore, the defender's best response curve may appear qualitatively different, as shown in figure 3.

Independent of the slope of the curves, there are two points where the defender is indifferent between two of its strategies. These occur at $\beta_1^* = \frac{\pi_2 w}{\pi_1 r + \pi_2 w}$ and $\beta_2^* = \frac{(1-\pi_2)w}{(1-\pi_1)r+(1-\pi_2 w)}$. Since the defender cannot have a pure strategy in equilibrium, it must be willing to randomize between at least two strategies. Therefore, the equilibrium attacker randomization probability must be at either one of these two values of β.

From the attacker's perspective, it is willing to randomize if it is indifferent between attacking and not attacking. That means the defender must randomize either after observing o_1 or after observing o_2 to make the attacker indifferent. Suppose the defender randomizes after o_1 and never retaliates after o_2. The randomization probability that would make the attacker indifferent between attacking and not attacking is $\frac{1}{\pi_1 c - \pi_2 v}$. Of course, this is only a proper probability if $\pi_1 c - \pi_2 v > 1$. This means that if the cost of an attacker getting caught is too low (low value of c) or if its penalty of being incorrectly retaliated against is too high (high value of v), the defender cannot retaliate with a high enough probability after o_1 to make the attacker indifferent between attacking and not attacking. In other words, if the cost of being correctly retaliated against is sufficiently close to the cost of being incorrectly retaliated against, the attacker should always just attack since the cost it incurs due to a retaliation does not significantly depend on whether or not it attacked.

The attacker's willingness to attack can also be phrased in terms of the defender's attribution probabilities. If the defender's attribution ability are low (π_1 is only slightly greater than π_2), then the attacker knows that its attack is not correlated with the defender's signal and thus attacking has little effect on whether or not the defender will retaliate. If this is the case, it is always in the attacker's best interest to attack. Conversely, if the defender's attribution ability is high (π_1 significantly greater than π_2) the attacker knows that if it attacks, it is likely to generate a signal leading to detection and therefore the defender is capable of randomizing to make the attacker indifferent between attacking and not attacking.

Now consider the case when the defender always retaliates after o_1 and randomizes its retaliation after

o_2. The attacker can only be made indifferent between attacking and not attacking if $\pi_1 c - \pi_2 v < 1$. In this case, if c is relatively high and v is relatively low and the defender will always retaliate after o_1, the attacker will incur a high cost when it does attack and is retaliated against. Since the defender's strategy says to always retaliate after o_1, the defender cannot randomize with a low enough probability after o_2 to ever induce the attacker to attack because the potential cost to attacking is so high.

Again, the analysis can be phrased in terms of the defender's attribution parameters. If the defender has high attribution ability (π_1 much greater than π_2) then the attacker knows that if it attacks, it is likely that the defender retaliates. This is because the defender retaliates after observing o_1 and when π_1 is high, the defender is more likely to observe o_1. Therefore, the defender cannot retaliate with a low enough probability after observing o_2 to compensate for the loss the attacker incurs when it attacks and is retaliated against because the attack generated signal o_1.

	$1 - \pi_1 c + \pi_2 v \leq 0$	$1 - \pi_1 c + \pi_2 v \geq 0$
U_d	$-\beta_1^*$	$\beta_2^*(r-1) - (1-\beta_2^*)w$
U_a	$-\pi_2 \rho_1^* v$	$-(\pi_2 + (1-\pi_2)\rho_2^*)v$

Table 1: Equilibrium expected utilities in the attribution game

Since $\beta_1^* < \beta_2^*$, the attacker attacks with a lower probability when $\pi_1 c - \pi_2 v > 1$, which occurs when either the defender has the ability to attribute with a high degree of certainty or the punishment for correctly retaliating is significantly higher than the punishment for incorrect retaliation. However, this does *not* mean that the defender is better off in the equilibrium where the attacker attacks less. Table 1 shows the defender's expected utility under each equilibrium. If $\pi_1 r_H + \pi_2 w > 1$, then the defender's expected utility is higher when the attacker attacks *more*. We will return to this point later when we discuss the question "should deterrence be right goal?"

3.2 The Signaling Game

We now turn our attention to the signaling game. Specifically, we examine whether there are equilibria in which the defender attempts to signal its true capability to the attacker. An important parametric assumption is whether $1 - c_L > -v$ (by assumption $1 - c_H > -v$ always holds). If $1 - c_L > -v$, then the attacker has a dominant strategy to attack *if it knew with certainty* that the defender is type L. This is because the attacker's payoff for attacking and getting punished $(1 - c_L)$ is greater than its payoff from not attacking and getting punished $(-v)$. Therefore, if the attacker knew the defender were type L and would always retaliate, its best response would be to always attack.

3.2.1 Separating Equilibria

First we establish that there is no separating equilibrium in which the defender's signal truthfully reveals its type, regardless of the sign of $1 - c_L + v$:

Proposition 2 (No Separating Equilibrium)**.** *Assume $1 - c_L < -v$. Then, there is no equilibrium where the defender truthfully signals its type in the signaling game. Formally, there is no PBE where $\alpha_1 = 1$ and $\alpha_2 = 0$.*

Proposition 3 (No Separating Equilibrium II)**.** *Assume $1 - c_L > -v$. Then, there is no equilibrium where the defender truthfully signals its type in the signaling game. Formally, there is no PBE where $\alpha_1 = 1$ and $\alpha_2 = 0$.*

Although the formal proofs of propositions 2 and 3 proceed differently, the logic is similar. If the defender truthfully signals its type to the attacker, then after the signal, the attacker and defender just play the attribution game analyzed in the previous section. However, the defender of type L or type H would be better off if it could deceive the attacker in playing a different attribution game. In other words, in some cases a defender of type H would be better off if it could convince the attacker it is type L and have the attacker choose its strategy *as if* the defender is type L. In other cases, the defender of type L would be

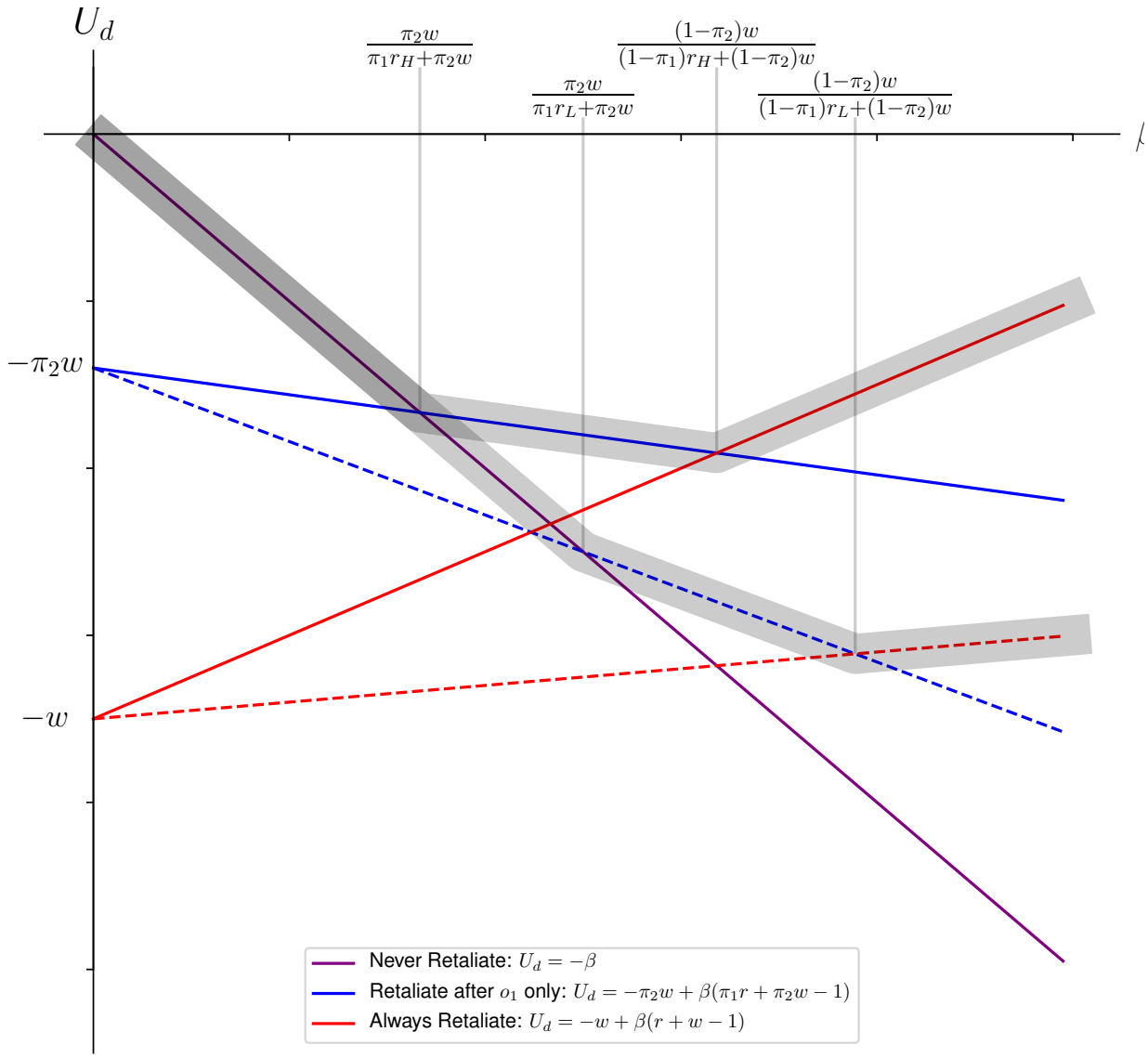


Figure 4: Defender's best responses in the signaling game. The solid purple line extending from the origin is the defender's payoffs from never retaliating. The solid lines represent the defender's payoff when it is type H and the dashed lines represent its payoffs when it is type L. The four labeled values of β denote the points where the defender is indifferent between two of its strategies

better off if it could convince the attacker that it was type H and have the attacker play the attribution game *as if* the defender were type H.

Figure 4 illustrates the payoff for the defender of type L and type H in the signaling game and illustrates why there can't be a separating equilibrium. If the defender did truthfully signal its type, then after the signal, the attacker and defender play the attribution game described above. Since there is a unique equilibrium in the attribution game, the attacker's randomization probabilities after receiving signal s_H are either $\frac{\pi_2 w}{\pi_1 r_H + \pi_2 w}$ or $\frac{(1-\pi_2 w)w}{(1-\pi_1)r_H + (1-\pi_2)w}$ and after receiving signal s_L, randomizes with probability $\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ or $\frac{(1-\pi_2 w)w}{(1-\pi_1)r_L + (1-\pi_2)w}$. All four of these randomization probabilities are annotated in figure 4.

For any two of the equilibrium probabilities annotated in the figure, it is clear that either a defender of type H or a defender of type L would have an incentive to switch its signal. For example, suppose the parameters were such that in the attribution game, when the defender is type H the attacker randomizes with

	β	(L, o_1)	(L, o_2)	(H, o_1)	(H, o_2)
1	$\frac{\pi_2 w}{\pi_1 r_H + \pi_2 w}$	DR	DR	P_1	DR
2	$\frac{(1-\pi_2)w}{(1-\pi_1)r_L + (1-\pi_2)w}$	R	P_2	R	R
3	$\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$	P_3	DR	R	DR
4	$\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$	P_4	DR	R	R
5	$\frac{(1-\pi_2)w}{(1-\pi_1)r_H + (1-\pi_2)w}$	DR	DR	R	P_5
6	$\frac{(1-\pi_2)w}{(1-\pi_1)r_H + (1-\pi_2)w}$	R	DR	R	p_6

Table 2: Possible Pooling Equilibria. Column's 2-5 give the defender's actions give its type and observation. For example column (L, o_1) gives the defender's action when the defender is type L and observes o_1. The defender's action can be either pure—R or DR— or mixed. When the defender's action is mixed, the probability is the probability that the defender retaliates. $P_1 = \frac{1}{\gamma(\pi_1 c_H - \pi_2 v)} P_2 = \frac{1-(1-\gamma)(c_L \pi_1 - \pi_2 v) - \gamma(c_H - v)}{(1-\gamma)(c_L(1-\pi_1) - v(1-\pi_2))}, P_3 = \frac{1-\gamma(\pi_1 c_H - \pi_2 v)}{(1-\gamma)(\pi_1 c_L - \pi_2 v)}$,
$P_4 = \frac{1-\gamma(c_H - v)}{(1-\gamma)(\pi_1 c_L - \pi_2 v)}, P_5 = \frac{1-\gamma(\pi_1 c_H - \pi_2 v)}{\gamma(c_H(1-\pi_1) - v(1-\pi_2))}, P_6 = \frac{1-\gamma(c_H - c_L)\pi_1 - c_L \pi_1 + \pi_2 v}{\gamma(c_H(1-\pi_1) - v(1-\pi_2))}$

probability $\beta_H = \frac{\pi_2 w}{\pi_1 r_H + \pi_2 w}$ and when the defender is type L randomizes with probability $\beta_L = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$. These points are where the solid blue line and the dashed blue line intersect the purple line extending from the origin, respectively. In this case, the defender of type L would have an incentive to signal s_H because its expected utility along its best response curve is higher at β_H. Of course, there are other possible equilibrium randomization probabilities in the attribution game and other versions of the graph (versions with the blue lines sloping upward) but in all cases, either a defender of type H or a defender of type L would have an incentive to not truthfully signal.

3.2.2 Pooling Equilibria

Before analyzing semi-separating equilibria, we present the possible pooling equilibria. In a pooling equilibrium, the defender's signal conveys no information regarding its true type and thus the attacker ignores the signal and only chooses one value of β for which to randomize its attack.

There exists a pure strategy equilibrium if $\gamma(1 - c_H) + (1 - \gamma)(1 - c_L) > -v$. In such an equilibrium, the attacker always attacks and the defender always retaliates. This is because the (net) cost to the attacker of being correctly retaliated against when the defender is type L is relatively small compared to its cost of being incorrectly retaliated against. Therefore, if the defender is significantly likely to be of type L (low value of γ), then an equilibrium attacker strategy is to always attack because the frequency in which the defender is type H and retaliates is not enough to deter the attacker from attacking when the defender is type L and has a relatively weak ability to punish.

If $\gamma(1 - c_H) + (1 - \gamma)(1 - c_L) < -v$, there is no pure strategy equilibrium in which the attacker always attacks or never attacks. This implies that the defender must also play a mixed strategy in order to make the attacker indifferent between attacking and not attacking[3]. For the defender to be willing to randomize at one of its information sets, it must be indifferent between two actions at that information set. This implies that a pooling equilibrium must have the attacker randomize with one of the four probabilities given in figure 4. Table 2 gives the possible pooling equilibria.

Not all of the equilibria in table 2 are possible simultaneously. First, the parameters must be such that the defender's randomization probabilities are between 0 and 1. In addition, the equilibria in lines 3 and 4 cannot exist simultaneously and lines 5 and 6 cannot exist simultaneously. To see why the equilibria in lines 5 and 6 cannot exist simultaneously, consider figure 3. Again, the gray highlighted line traces the defender's best response for each of its types. If the attacker randomizes with probability $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r_H + (1-\pi_2)w}$, then a defender of type H is indifferent between always retaliating and retaliating after o_1 only. However, a defender of type L may have a best response of either retaliating after o_1 and not retaliating after o_2, as shown in figure 5 or to never retaliate, as shown in figure 6. With the exception of a measure zero set of parameters,

[3]Again, there is a measure zero set of parameters where the defender would not have to randomize to make the attacker indifferent but we ignore such a case for the reasons given above.

the defender defender cannot be indifferent between never retaliating and retaliating after o_1 only when the attacker randomizes with probability $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r_H+(1-\pi_2)w}$, and thus only one of the two equilibria can exist for a given value of the parameters. The same type of argument can be used to show that only one of the equilibria in rows 5 and 6 can exist simultaneously[4].

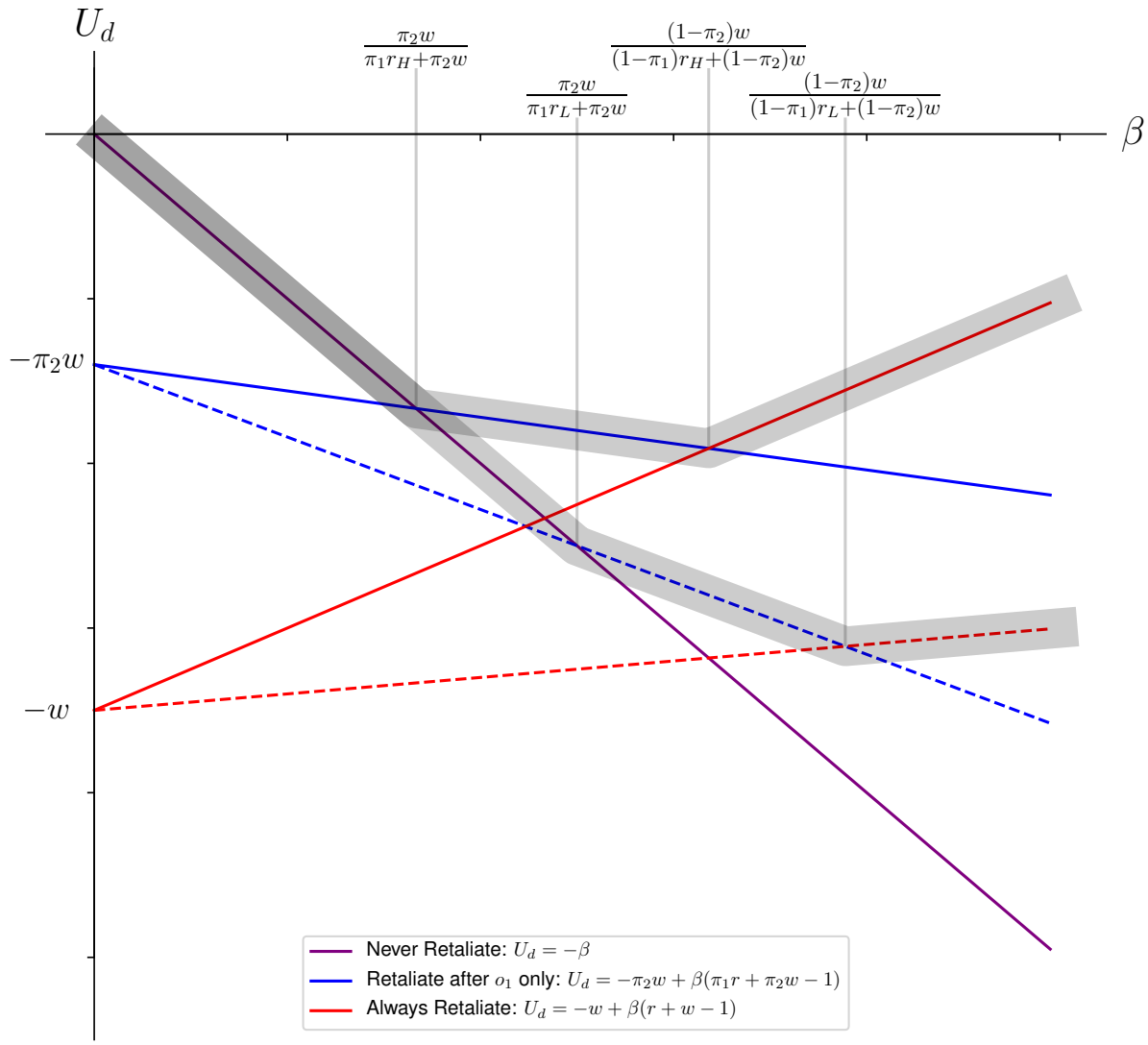


Figure 5: Defender's of type L's best response at $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r_H+(1-\pi_2)w}$ is to retaliate after o_1 and not retaliate after o_2

Since there always exists a babbling equilibrium in cheap talk games, at least one equilibrium in table 2 exists. [5] On the other hand, for some values of the parameters, there are multiple babbling equilibria. For example, when $\pi_1 = .9, \pi_2 = .1, c_H = 4, c_L = 2, v = 2$ and $\gamma = .4$, the randomization probabilities in

[4]Formally, when $r_H > \frac{\pi_1}{\pi_2}\frac{1-\pi_2}{1-\pi_1}$, row 4 and row 5 are possible equilibria. Otherwise, row 3 and 6 are possible equilibria. The measure zero parameter set we ignore occurs when $r_H = \frac{\pi_1}{\pi_2}\frac{1-\pi_2}{1-\pi_1}$

[5]See Farrell, Joseph, and Matthew Rabin. "Cheap talk." Journal of Economic perspectives 10.3 (1996): 103-118.

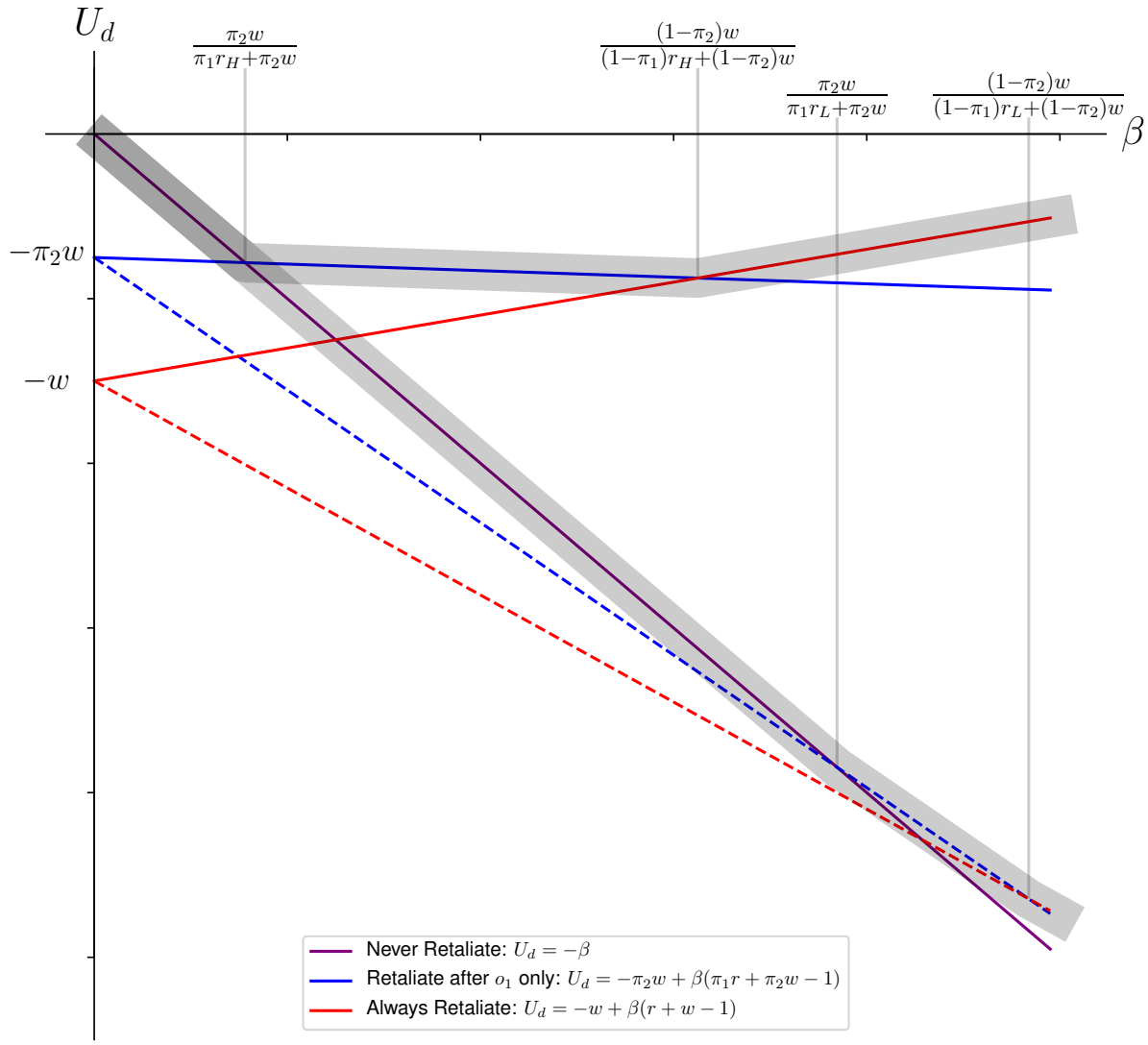


Figure 6: Defender's of type L best response at $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r_H + (1-\pi_2)w}$ is to never retaliate.

row 1,2,4 and 5 are all proper probabilities and thus there are multiple equilibria. Additionally, at those parameter values, the equilibrium in which the attacker always attacks and the defender always retaliates also exists. We will continue the analysis of pooling equilibria when we discuss each equilibrium relative to the semi-separating equilibrium derived in the following section.

3.2.3 Semi-Separating Equilibria

Thus far, we have established that there are never separating equilibria, and depending on the parameter regime, many possible pooling equilibria and one possible pure strategy equilibrium. In this section, we establish the conditions in which there are equilibria where the defender's signal contains some — but not perfect— information regarding its true type.

Proposition 4 (No Semi-Separating Equilibria when $1 - c_L < -v$)**.** *Assume* $1 - c_L < -v$*. Then:*

1. *There is no equilibrium where the defender of type L always signals s_L and a defender of type H randomizes between signaling s_L and s_H and the attacker randomizes with probability β_L after receiving s_L and β_H after receiving s_H and $\beta_L \neq \beta_H$.*

2. *There is no equilibrium where the defender of type H always signals s_H and a defender of type L randomizes between signaling s_L and s_H and the attacker randomized with probability β_L after receiving s_L and β_H after receiving s_H and $\beta_L \neq \beta_H$.*

Proposition 4 says that if an attacker of type L has relatively high ability to punish (relatively high value of c_L), then there is no equilibrium in which the defender truthfully signals when it is one type and randomizes its signal when it is another type. While this proposition, in isolation is a "negative result," it is useful as context for the following result, established in proposition 5.

Proposition 5 (Semi-Separating Equilibrium with Low Punishment Power)**.** *Assume*

1. $1 - c_L > -v$,
2. $\pi_1 c_L - \pi_2 v > 1$,
3. $\pi_2 < \pi_1 r_L + \pi_2 w < 1$.
4. $r_L + w > 1$
5. $\frac{\gamma}{1-\gamma} < \frac{c_L - v - 1}{1 - c_H + v}$
6. $w(\pi_1 r_L + \pi_2 w - \pi_2) < (r_L + w - 1)(\pi_1 r_L + \pi_2 w)$

Then there exists an equilibrium where:

- *A defender of type H always signals s_H and always retaliates.*
- *A defender of type L signals s_H with probability $\alpha_L = \frac{\gamma}{1-\gamma}\frac{1-c_H+v}{c_L-v-1}$. After signaling s_H, the defender always retaliates. After signaling s_L, the defender retaliates with probability $\rho(o_1, s_L, L) = \frac{1}{\pi_1 c_L - \pi_2 v}$ after observing o_1 and never retaliates after observing o_2*
- *An attacker that receives signal s_H attacks with probability $\beta_H = \frac{w(\pi_1 r_L + \pi_2 w - \pi_2)}{(r_L + w - 1)(\pi_1 r_L + \pi_2 w)}$.*
- *An attacker that receives signal L attacks with probability $\beta_L = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$.*

To understand proposition 5 it is beneficial to understand the parameter regime in which the equilibrium exists. The first condition ($1 - c_L > -v$) says that an attacker's best response to a defender of type L that always retaliates is to always attack. Intuitively, this means that a defender of type L has a relatively low capability to deliver an impactful correct retaliation. Condition 2 says that if the attacker knew with certainty that the defender were type L, there is an equilibrium in the induced attribution game where the attacker randomizes with probability $\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$. This condition together with condition 3 establishes that the defender has a relatively high attribution capability (high values of π_1 and low values of π_2 expand the parameter region). Condition 4 says that a defender of type L has enough ability to punish that as the attacker attacks more, the utility of the defender from always punishing increases. Condition 5 says that the defender is not type H too often. Finally, condition 6 says that there is a wide enough range of attacker randomization probabilities where the defender's best response is to always retaliate. In summary, for the semi-separating equilibrium to exist, the defender must have relatively high attribution capability, not have a high retaliatory capability too often and that the defender of type L does not deliver a relatively strong punishment when it successfully retaliates.

To see how such an equilibrium exists, consider figure 7. The two values of β that are labeled are the attacker randomization probabilities in the semi-separating equilibrium. When the attacker attacks with probability β_L, the defender of type L is indifferent between never retaliating and retaliating after o_1. This is where the purple line intersects the dotted blue line. When the attacker attacks with probability β_H, the type L defender's best response is to always retaliate. The horizontal green line illustrates that a defender of type L is indifferent between these two outcomes and thus is willing to randomize its signal when it is type

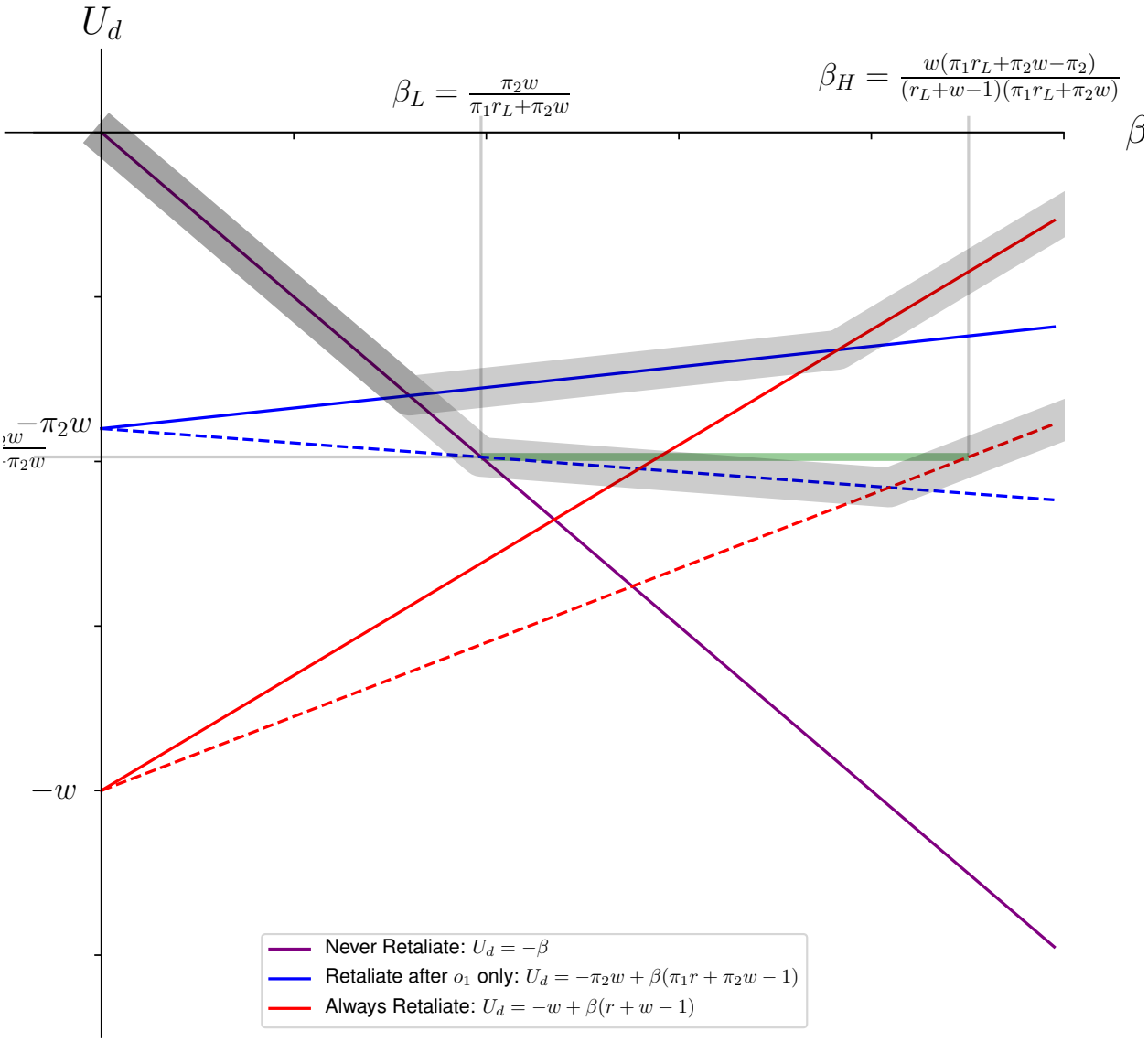


Figure 7: Semi-Separating Equilibrium

L. A defender of type H receives a higher utility when the attacker attacks with probability β_H and always retaliates (solid red line) than when the attacker attacks with probability β_L and the attacker retaliates after o_1 only (dotted blue line). Therefore, the defender of type H would always signal s_H, as indicated in the semi-separating equilibrium.

3.2.4 Gains from Signaling

Finally we investigate whether it is possible for the defender to gain from signaling. To carry out such an analysis, we say that there is a gain from signaling if for a fixed set of parameters, the semi-separating equilibrium exists and the defender's expected utility in the semi-separating equilibrium is higher than it's expected utility in a pooling equilibrium that exists simultaneously. Of course, this ignores any notion of

Parameter	Value
π_1	.8
π_2	.45
c_H	4
c_L	3
v	2.6
γ	.4
r_H	.9
r_L	.65
w	.8

Table 3: Parameter values where there are gains from signaling through deterrence

equilibrium selection and how the players might arrive at such an equilibrium, which is an interesting future endeavor. We also say that it is possible from the defender to gain with signaling through a *deterrence effect* if the defender's expected utility under the semi-separating equilibrium is higher than in the pooling equilibrium *and* the attacker's attack probability is lower in the semi-separating equilibrium than in the pooling equilibrium.

First, we illustrate that there exist parameter regimes where the defender can gain by deterring the attacker. Consider the parameter values in table 3.[6] In this parameter regime, the conditions in proposition 5 are satisfied and the semi-separating equilibrium exists. At this equilibrium, the attacker attacks with probability .715 and the defender's expected utility before realizing its type is $-.298$. At these parameter values, the equilibrium in row 2 of table 2 also exists. At this equilibrium, the attacker attacks with probability .772 and the defender earns an expected utility of $-.36$.

Figure 8 illustrates the deterrent effect of the semi-separating equilibrium. In the pooling equilibrium, the attacker randomizes with probability $\beta_p = \frac{(1-\pi_2)w}{(1-\pi_1)r_L+(1-\pi_2)w}$. In the semi-separating equilibrium, the attacker will randomizes either at probability β_H or β_L. While β_H is slightly higher than β_p, β_L is sufficiently low such that on average, the attacker attacks less and the defender's expected utility increases by reducing the probability in which the attacker attacks. Since the defender of type L has a higher expected utility at β_H and β_L then at β_p, the defender gains with signaling through a deterrence effect.

After establishing that the defender can benefit through signaling by deterring an attacker, we now present our final result that shows that a defender can benefit through signaling by inducing the attacker to attack more. Consider the parameter set in table 4. In this parameter regime, the semi-separating equilibrium exists and the attacker attacks with probability .729 and the defender earns an expected payoff of $-.249$. At those parameter values, the equilibrium in row 1 of table 2 also exists. This equilibrium is the pooling equilibrium in which the attacker randomizes its attack with the lowest probability. At that pooling equilibrium, the attacker attacks with probability .260 and the defender's expected payoff is $-.260$. This means that the defender can increase its expected utility from signaling by inducing the attacker to attack *more*.

There are two reasons for this counter-intuitive result. The first reason has to do with the trade-off the defender faces between an incorrect retaliation and an undetected attack. If the cost to an incorrect retaliation is relatively high, then the defender has little incentive to retaliate because it risks the possibility of being incorrect. However, if the attacker were to attack with a higher probability, then the defender would incorrectly retaliate less often. If the cost of an incorrect retaliation is high enough, then the defender would benefit from being attacked slightly more but incorrectly retaliating less often.

The second effect is due to the defender inducing the attacker to attack against a defender of type H where the defender gains the most from a correct retaliation. Consider figure 9. The figure illustrates the pooling equilibrium at β_p and the semi-separating equilibrium where the attacker randomizes either at β_L and β_H, depending on the signal it gets from the defender. The defender of type L's expected utility when the attacker attack with probability β_p is higher than if the attacker were to attack with probability β_L or β_H, indicating that a defender of type L is worse off when the attacker attacks more. However, the expected

[6]The intuition that we demonstrate in this section holds for a set of parameters of positive measure. We only select specific parameter values to highlight the main insights.

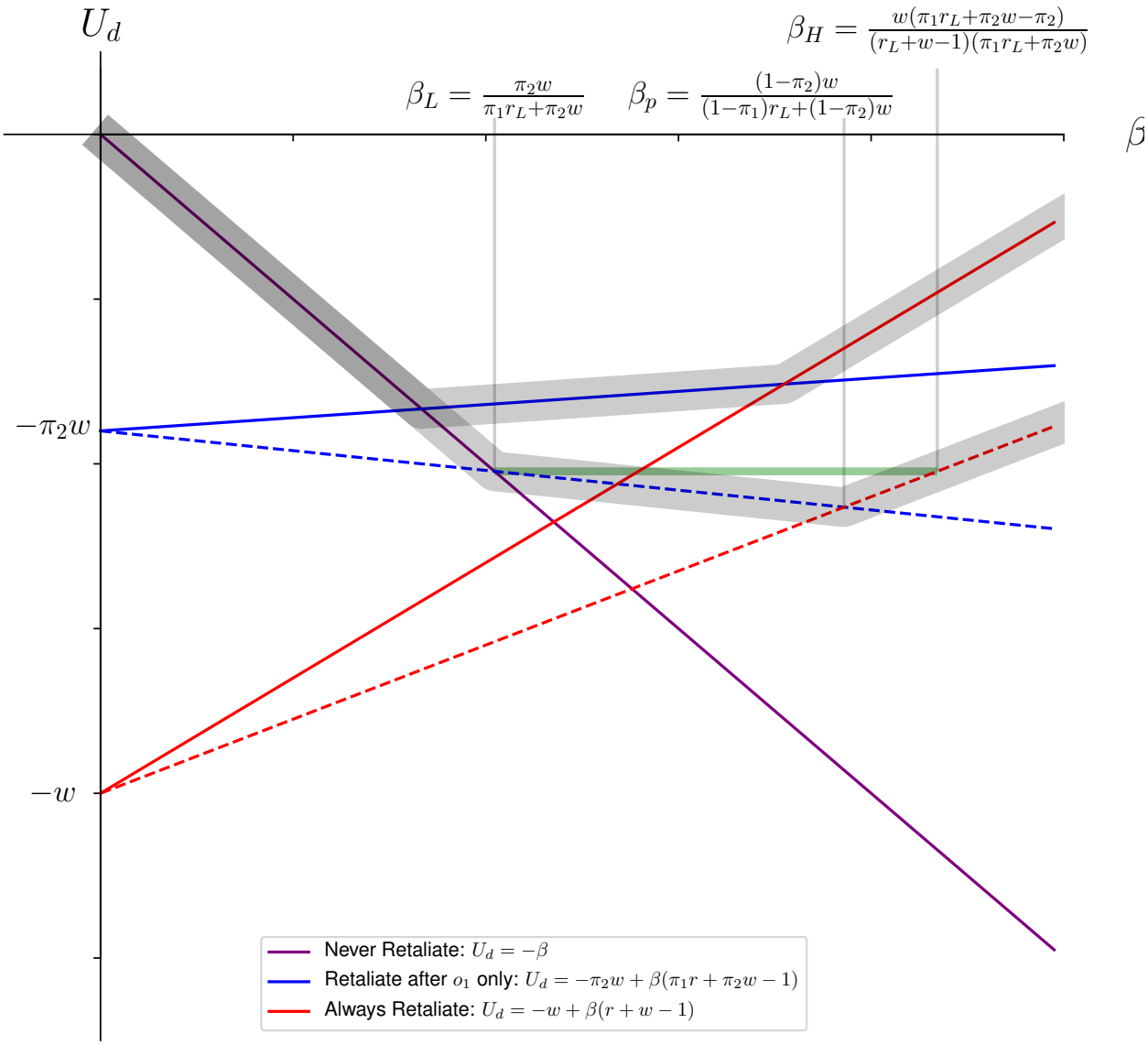


Figure 8: The attack probabilities for the pooling and semi-separating equilibrium

utility of an attacker of type H is higher at β_H than at β_p. Therefore, if the defender can randomize its signal such that the benefit it receives from retaliating when it is type H outweighs the loss it would incur from a higher attack probability when the defender is type L, then the defender stands to gain from a higher attack probability. In other words, the defender can gain when the attacker attacks more as long as the increased attack probability mostly occurs when the defender is type H and can more effectively retaliate.

As a final illustration of the equilibrium and its properties, figure 10 shows the parameter region where there are gains to be made from signaling by inducing the attacker to attack more. This region is defined by the conditions in proposition 5 and the region where $\frac{1}{\gamma(\pi_1 c_H - \pi_2 v)}$ is a proper probability. Generally speaking, the lower right corner of the plot is where attribution capabilities are the highest (high value of π_1 and low value of π_2. The figure shows that as the cost of an incorrect retaliation increases, the defender's attribution

Parameter	Value
π_1	.95
π_2	.5
c_H	5
c_L	3
v	3
γ	.32
r_H	.9
r_L	.7
w	.6

Table 4: Parameter values where there are gains from signaling through deterrence

ability must increase in order to support an equilibrium where there are gains from signaling though an anti-deterrence effect.

4 Conclusion — Towards a Cyber Deterrence Policy

This work responds to an active conversation on the strategy of cyber deterrence where calls for a cyber deterrence policy have been met with debate on the desirability and feasibility of deterring aggression in cyberspace. The challenges emanating from the cyber domain on conflict are a marked departure from the challenges of the nuclear domain. Thus, while the fundamental building blocks of deterrence outlined by Schelling persist, the lessons of nuclear deterrence may not. We offer a model with two players, imperfect information, and signaling to analyze the viability of deterrence in domains with imperfect attribution and signaling. In addition to highlighting the significant obstacles to cyber deterrence, our findings elucidate unique opportunities for deterrence and the curious value of anti-deterrence.

Unfortunately, complete deterrence will not come easy for the foreseeable future; with imperfect attribution, there are no equilibria in which the attacker will never attack and therefore we find complete deterrence unlikely. However, that does not render some deterrence unfeasible. To the contrary, when a defender is able to signal its capability, it may partially deter an attacker from launching an attack. Specifically, we show that there are semi-separating equilibria in which the attacker attacks with a lower probability than in a game without signaling and the defender receives the benefits from the reduced attacks. Thus, while signaling does not bring the attack probability to zero, it can reduce the chances of an attack. Signaling, therefore, is a key feature of a deterrence policy worthy of further exploration.

Our findings also point to a curious concept of anti-deterrence that raises the question: is deterrence the only option? In a world without perfect information and verifiable signals, our results suggest that anti-deterrence may be a means of increasing the defenders well-being while also welcoming a higher probability of attack. We show that in some cases, a defender would be willing to take on a higher probability of being attacked if 1) the higher probability of attack reduces the probability (and therefore costs) of an incorrect retaliation and 2) the increase in the probability of attack is more heavily weighted to when the defender is most able to respond to attack and not when it cannot effectively respond.

The conversation on cyber deterrence is ripe for further debate. Additional work could enhance the present analysis by endogenizing the defenders capability in which the defender undertakes costly investment to improve its, retaliatory, detection and attribution abilities. Furthermore, while we introduce costless signaling to the discussion on cyber deterrence, additional work should consider costly signaling. Specifically, a fruitful extension would be to consider the single use nature of cyber weapons and whether a defender with a stockpile of single use weapons can add credibility to its signal by strategically using some of its weapons as merely a signal.

References

[1] 115th Congress. H.r.5515 - john s. mccain national defense authorization act for fiscal year 2019.

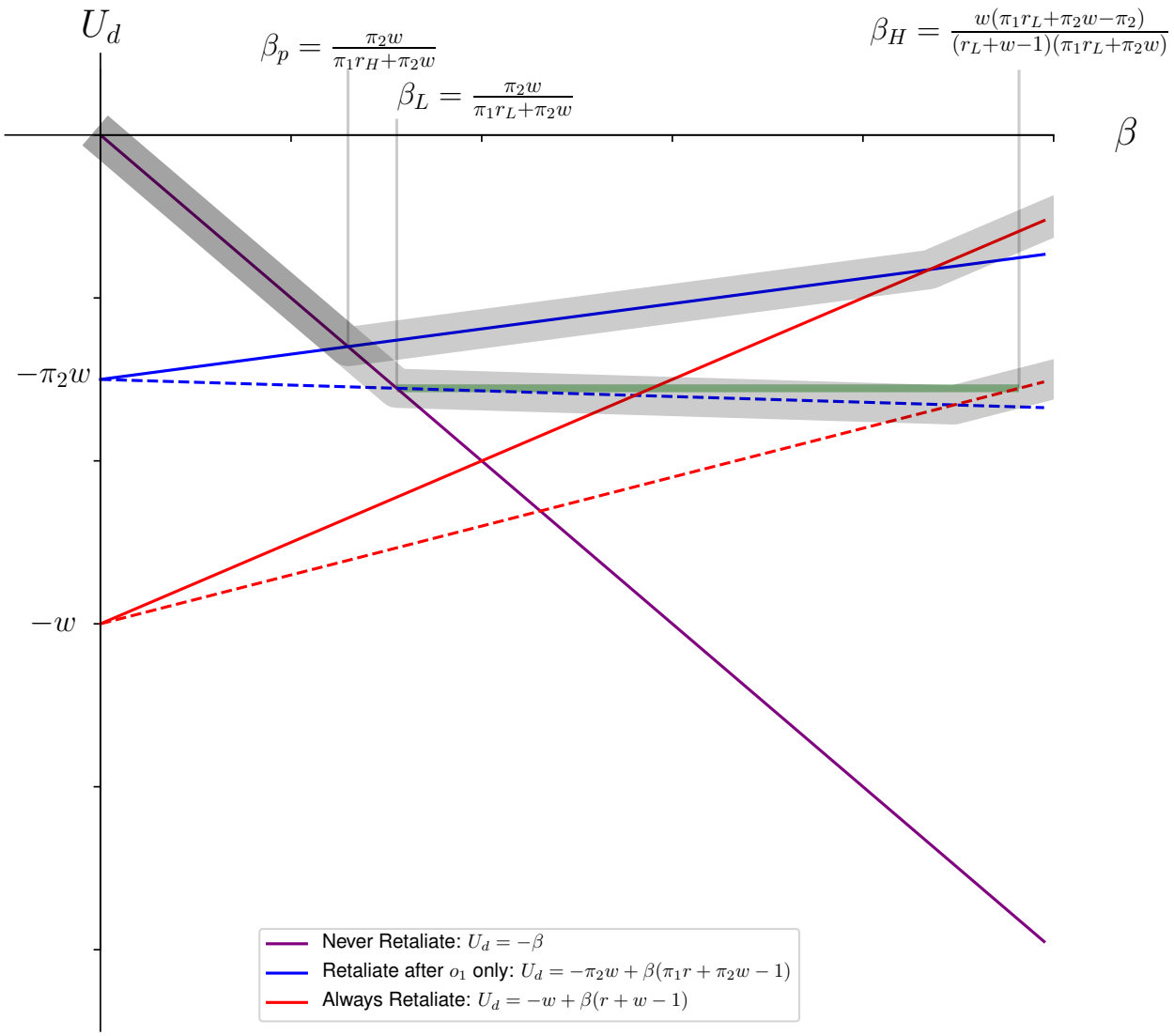


Figure 9: Semi-separating equilibrium where the defender gains by inducing the attacker to attack more

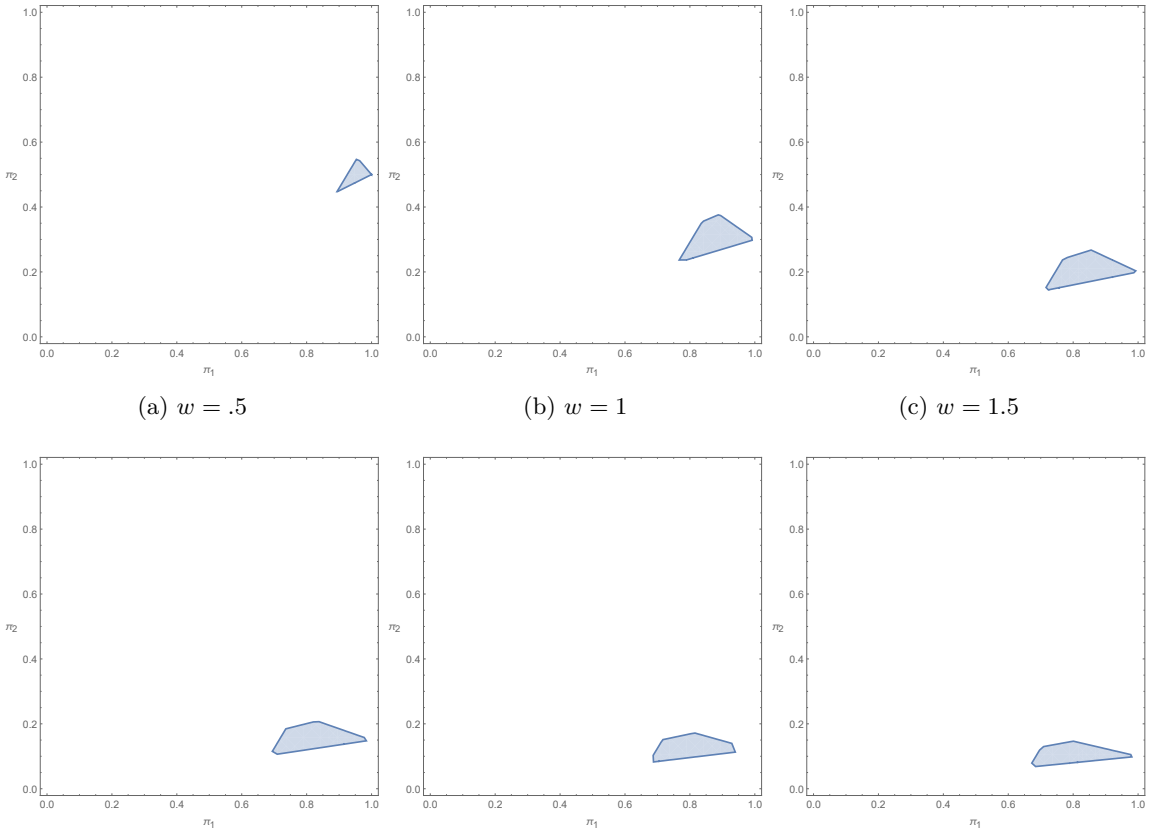
Technical report, Dec 2018.

[2] James Andreoni. Reasonable doubt and the optimal magnitude of fines: should the penalty fit the crime? *The RAND Journal of Economics*, pages 385–395, 1991.

[3] Kyle Bagwell. Signalling and entry deterrence: A multidimensional analysis. *The RAND Journal of Economics*, 38(3):670–697, 2007.

[4] Sandeep Baliga, SOM Kellogg, Ethan Bueno de Mesquita, and Alexander Wolitzky. Deterrence with imperfect attribution. *Working Paper*, 2019.

[5] Annegret Bendiek and Tobias Metzger. Deterrence theory in the cyber-century. *INFORMATIK 2015*, 2015.



(a) $w = .5$ (b) $w = 1$ (c) $w = 1.5$

(d) $w = 2$ (e) $w = 2.5$ (f) $w = 3$

Figure 10: Region in π_1, π_2 space where the semi-separating equilibrium exists and the defender can gain from signaling by inducing the attacker to attack more. The parameters other than π_1, π_2 and w are as in table 4.

[6] Emile Borel. a theorie du jeu et les equations integrales a noyan symetrique. *Comptes Rendus de lAcademicdes Science*, 1921.

[7] Richard E Caves, Michael E Porter, et al. From entry barriers to mobility barriers: Conjectural decisions and contrived deterrence to new competition. *Quarterly journal of Economics*, 91(2):241–261, 1977.

[8] Defense Science Board. Task force on cyber deterrence. Technical report, Feb 2017.

[9] Benjamin Edwards, Alexander Furnas, Stephanie Forrest, and Robert Axelrod. Strategic aspects of cyberattack, attribution, and blame. *Proceedings of the National Academy of Sciences*, 114(11):2825–2830, 2017.

[10] Nancy T Gallini. Deterrence by market sharing: A strategic incentive for licensing. *The American Economic Review*, 74(5):931–941, 1984.

[11] Will Goodman. Cyber deterrence: Tougher in theory than in practice? Technical report, SENATE (UNITED STATES) WASHINGTON DC COMMITTEE ON ARMED SERVICES, 2010.

[12] Shane Harris. China reveals its cyberwar secrets, 2017.

[13] Emilio Iasiello. Is cyber deterrence an illusory course of action? *Journal of Strategic Security*, 7(1):54–67, 2014.

[14] Eric Talbot Jensen. Cyber deterrence. *Emory Int'l L. Rev.*, 26:773, 2012.

[15] Jesse C Johnson, Brett Ashley Leeds, and Ahra Wu. Capability, credibility, and extended general deterrence. *International Interactions*, 41(2):309–336, 2015.

[16] Martin C Libicki. *Cyberdeterrence and cyberwar*. Rand Corporation, 2009.

[17] Marc Lynch. Why engage? china and the logic of communicative engagement. *European Journal of International Relations*, 8(2):187–230, 2002.

[18] Patrick M Morgan. *Deterrence now*, volume 89. Cambridge University Press, 2003.

[19] Madame Florence Parly. Stratgie cyber des armes. 2019.

[20] Robert Powell. *Nuclear deterrence theory: The search for credibility*. Cambridge University Press, 1990.

[21] Bruce Riedel. Al qaeda strikes back. *Foreign Affairs*, pages 24–40, 2007.

[22] Sankardas Roy, Charles Ellis, Sajjan Shiva, Dipankar Dasgupta, Vivek Shandilya, and Qishi Wu. A survey of game theory as applied to network security. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

[23] Steven C Salop. Strategic entry deterrence. *The American Economic Review*, 69(2):335–338, 1979.

[24] Vladislav Saran. Media manipulation and psychological war in ukraine and the republic of moldova. *Centre for European Studies (CES) Working Papers*, 8(4), 2016.

[25] Thomas C Schelling. *The strategy of conflict*. Harvard university press, 1980.

[26] Paulo Shakarian, Gerardo I Simari, Geoffrey Moores, and Simon Parsons. Cyber attribution: An argumentation-based approach. In *Cyber Warfare*, pages 151–171. Springer, 2015.

[27] Sajjan Shiva, Sankardas Roy, and Dipankar Dasgupta. Game theory for cyber security. In *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research*, page 34. ACM, 2010.

[28] Kannan Srinivasan. Multiple market entry, cost signalling and entry deterrence. *Management Science*, 37(12):1539–1555, 1991.

[29] Mariarosaria Taddeo. The limits of deterrence theory in cyberspace. *Philosophy & Technology*, 31(3):339–355, 2018.

[30] Zhiheng Xu and Jun Zhuang. A study on a sequential one-defender-n-attacker game. *Risk Analysis*, 2019.

[31] Xiaoyan Zhou, Jincai Huang, and Guangquan Cheng. Attacker-defender signaling game in multi-period based on technology accumulation and bayesian learning. In *2015 3rd International Conference on Machinery, Materials and Information Technology Applications*. Atlantis Press, 2015.

[32] Quanyan Zhu and T Başar. Decision and game theory for security, 2013.

A Appendix

FOR ONLINE PUBLICATION

- **Proposition 1** To prove proposition 1, we begin with two lemmas that establish there are no pure strategy equilibria and then prove the proposition by looking for mixed strategy equilibria.

  Lemma 1 (No Pure Strategy Equilibrium for the Attacker in the Attribution Game)**.** *If $1 - c < -v$ there is no equilibrium in the attribution game where the attacker plays a pure strategy.*

  Proof. Suppose the attacker's pure strategy is to never attack. The defender's best response against such a strategy is to never retaliate. However, the attacker's best response to the defender never retaliating is to always attack. Therefore, there is no pure strategy equilibrium where the attacker never attacks. Now, suppose the attacker's pure strategy is to always attack. In this case, the defender's best response is to always retaliate. However, the attacker's best response to the defender always retaliating is to never attack (since by assumption, the total payoff to attacking and being correctly retaliated against, $1 - c$, is less than the total payoff of being incorrectly retaliated against $-v$.) Therefore, there is no pure strategy equilibrium where the attacker always attacks. □

  Lemma 2 (No Pure Strategy Equilibrium for the Defender in the Attribution Game)**.** *If $1 - c < -v$ there is no equilibrium in the attribution game where the defender plays a pure strategy.*

  Proof. Suppose the defender's pure strategy is to always retaliate. Then the attacker's best response would be to never attack and thus the defender's best response would be to never retaliate. Therefore, always retaliating cannot be part of an equilibrium. A parallel argument shows that never retaliating cannot be part of of equilibrium. The only other pure strategy for the defender in the attribution game is to always retaliate after receiving signal o_1 and never retaliate after signal o_2 (since $\pi_1 > \pi_2$, the strategy always retaliate after o_2 and never retaliate after o_1 is strictly dominated). If the defender adopts this strategy, the attacker's expected utility from attacking and not attacking is given by:

  $$\begin{aligned} U_a(A,(R,DR)) = \quad & Pr(o_1|A)(1-c) + Pr(o_2|A) = \quad \pi_1(1-c) + (1-\pi_1) \\ U_a(NA,(R,DR)) = \quad & Pr(o_1|NA)(-v) + Pr(o_2|NA) \times 0 = \quad \pi_2 v \end{aligned}$$

  where $U_a(X,(Y,Z))$ is the expected utility of the attacker for choosing action X where the defender chooses (the probability of) action Y after observing o_1 and (the probability of action) Z after observing o_2. These expected utilities implies that if $1 - \pi_1 c - \pi_2 v > 0$, then the attacker would always attack and if $1 - \pi_1 c - \pi_2 v < 0$ then attacker would never attack both of which by lemma 1 cannot be part of an equilibrium (Recall that by assumption $1 - \pi_1 c - \pi_2 v \neq 0$). □

  Lemmas 1 and 2 establish that there cannot be an equilibrium in which either the attacker or the defender play pure strategies. Therefore, we prove proposition 1 by searching for mixed strategy equilibria only.

  Proof of Proposition 1. Suppose the attacker randomizes with probability β. Then equilibrium defender beliefs are determined as follows:

  $$\begin{aligned} Pr(A|o_1) &= \frac{Pr(o_1|A)Pr(A)}{Pr(o_1)} = \frac{\pi_1\beta}{\pi_1\beta + \pi_2(1-\beta)} \\ Pr(DA|o_1) &= \frac{Pr(o_1|A)Pr(DA)}{Pr(o_1)} = \frac{\pi_2(1-\beta)}{\pi_1\beta + \pi_2(1-\beta)} \\ Pr(A|o_2) &= \frac{Pr(o_2|A)Pr(A)}{Pr(o_2)} = \frac{(1-\pi_1)\beta}{(1-\pi_1)\beta + (1-\pi_2)(1-\beta)} \\ Pr(DA|o_2) &= \frac{Pr(o_2|A)Pr(DA)}{Pr(o_2)} = \frac{(1-\pi_2)(1-\beta)}{(1-\pi_1)\beta + (1-\pi_2)(1-\beta)} \end{aligned}$$

With these probabilities, it is possible to write the defender's expected utility from retaliating and not retaliating for each of its observations They are given by:

$$\begin{aligned}
U_d(R,\beta;o_1) &= Pr(A|o_1)(r-1) - Pr(DA|o_1)w \\
&= \frac{\pi_1\beta}{\pi_1\beta+\pi_2(1-\beta)}(r-1) - \frac{\pi_2(1-\beta)}{\pi_1\beta+\pi_2(1-\beta)}w \\
U_d(DR,\beta;o_1) &= -Pr(A|o_1) - 0\times Pr(DA|o_1) \\
&= -\frac{\pi_1\beta}{\pi_1\beta+\pi_2(1-\beta)} \\
U_d(R,\beta;o_2) &= Pr(A|o_2)(r-1) - Pr(DA|o_2)w \\
&= \frac{(1-\pi_1)\beta}{(1-\pi_1)\beta+(1-\pi_2)(1-\beta)}(r-1) - \frac{(1-\pi_2)(1-\beta)}{(1-\pi_1)\beta+(1-\pi_2)(1-\beta)}w \\
U_d(DR,\beta_H;o_2) &= -Pr(A|o_2) - 0\times Pr(DA|o_2) \\
&= -\frac{(1-\pi_1)\beta}{(1-\pi_1)\beta+(1-\pi_2)(1-\beta)}
\end{aligned}$$

where $U_d(X,\beta;o)$ is the expected utility of the defender by choosing action X given the attacker randomizes with probability β and the defender observed observation o. Since there is no equilibrium where the defender plays a pure strategy and must randomize, it must be indifferent at (at least) one of its information sets.

After observing o_1 the defender is indifferent between retaliating and not retaliating when:

$$\begin{aligned}
\frac{\pi_1\beta}{\pi_2(1-\beta)} &= \frac{w}{r} \\
\implies \beta &= \frac{\pi_2 w}{\pi_1 r+\pi_2 w} \qquad (1)
\end{aligned}$$

where it is optimal for the defender to retaliate if β is greater than the right hand side (RHS) of equation 1 and not retaliate if β is less than the RHS.

After observing o_2, the defender is indifferent between retaliating and not retaliating when

$$\begin{aligned}
\frac{(1-\pi_1)\beta}{(1-\pi_2)(1-\beta)} &= \frac{w}{r} \\
\implies \beta &= \frac{(1-\pi_2)w}{(1-\pi_1)r+(1-\pi_2)w} \qquad (2)
\end{aligned}$$

where it is optimal for the defender to retaliate if β is greater than the right hand side (RHS) of equation 2 and not retaliate if β is less than the RHS. Since $\pi_1 > \pi_2$ by assumption, the RHS of 1 is less than the RHS of 2.

Since the defender must be indifferent at at least one of its information sets, equations 1 and 2 give the only two possible values of equilibrium attacker randomization probability. We will now establish the sufficient conditions for the values in equations 1 and 2 to be part of a PBE.

Case 1: $\beta = \frac{\pi_2 w}{\pi_1 r+\pi_2 w}$ Suppose the attacker randomizes with probability $\beta = \frac{\pi_2 w}{\pi_1 r+\pi_2 w}$, then the defender will never retaliate after receiving o_2 and is indifferent after observing o_1, and therefore is willing to randomize with probability ρ after observing o_1. The necessary and sufficient condition for the attacker to be willing to randomize is if its expected utility from attacking is equal to its expected utility from not attacking. This is satisfied when:

$$\begin{aligned}
U_a(A,(\rho,DR)) &= U_a(NA,(\rho,DR)) \\
\implies Pr(o_1|A)Pr(R|o_1)(1-c) + Pr(o_1|A)Pr(NR|o_1) + Pr(o_2|A) &= Pr(o_1|A)Pr(R|o_1)(-v) \\
\implies \pi_1\rho(1-c) + \pi_1(1-\rho) + 1 - \pi_1 &= -\pi_2\rho v \\
\implies \rho &= \frac{1}{\pi_1 c - \pi_2 v}. \qquad (3)
\end{aligned}$$

Since ρ is a probability, it only takes on the values between 0 and 1, which holds only when $1-\pi_1 c+\pi_2 v \leq 0$. Therefore, if $1-\pi_1 c+\pi_2 v \leq 0$, then there is a Nash equilibrium where the attacker randomizes with probability $\beta = \frac{\pi_2 w}{\pi_1 r+\pi_2 w} = \beta_1^*$ and the defender never retaliates after observing o_2 and randomizes with probability $\rho = \frac{1}{\pi_1 c-\pi_2 v} = \rho_1^*$ after observing o_1. This proves item 1 of the proposition.

Case 2: $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r+(1-\pi_2)w}$ Suppose the attacker randomizes with probability $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r+(1-\pi_2)w}$, then the defender will always retaliate after receiving o_1 and is willing tor randomize with probability ρ after observing o_2. The necessary and sufficient condition for the attacker to be willing to randomize is if its expected utility from attacking is equal to its expected utility from not attacking. This is satisfied when:

$$\begin{aligned} U_a(A,(R,\rho)) &= U_a(NA,(R,\rho)) \\ \implies Pr(o_1|A)(1-c)+Pr(o_2|A)\rho(1-c)+Pr(o_2|A)(1-\rho) &= Pr(o_1|NA)(-v)+Pr(o_2|NA)\rho(-v) \\ \implies \pi_1(1-c)+(1-\pi_1)\rho(1-c)+(1-\pi_1)(1-\rho) &= -v(\pi_2+(1-\pi_2)\rho) \\ \implies \rho &= \frac{1-\pi_1 c+\pi_2 v}{(1-\pi_1)c-(1-\pi_2)v}. \end{aligned} \tag{4}$$

Since by assumption $c-v>1$ the numerator is less than the denominator and thus the only way that ρ represents a proper probability is if $1-\pi_1 c+\pi_2 v>0$. Therefore, if $1-\pi_1 c+\pi_2 v>0$, then there is a Nash equilibrium where the attacker randomizes with probability $\beta = \frac{(1-\pi_2)w}{(1-\pi_1)r+(1-\pi_2)w} = \beta_2^*$ and the defender always retaliates after observing o_1 and randomizes with probability $\rho = \frac{1-\pi_1 c+\pi_2 v}{(1-\pi_1)c-(1-\pi_2)v} = \rho_2^*$ after observing o_2. This proves item 2 of the proposition. □

- *Proof of Corollary 1.* As shown in the proof of proposition 1, there are only two possible values of β that can be part of a Nash equilibrium. For each value of β, there is only one mixed strategy for the defender that would make the attacker indifferent and thus willing to randomize. This suggests that there may be two equilibria. However, under one value of β the existence of a defender's equilibrium mixed strategy relies on $1-\pi_1 c+\pi_2 v>0$ where for the other value of β, the existence of the defender's mixed strategy relies on $1-\pi_1 c+\pi_2 v<0$. Since both conditions can not hold simultaneously, there is a unique Nash equilibrium determined by the sign of $1-\pi_1 c+\pi_2 v$. □

- *Proof of Proposition 2.* If the defender truthfully signals its capability, then Bayes rule dictates that the attacker assigns probability 1 to the defender's true capability and 0 otherwise. Therefore, after the truthful signal, a separating equilibrium would have the players play the equilibrium profile of the attribution game where the parameters are determined by the defender's true type and the payoffs would be as in table 1 where the values of r and c are given according to the defender's type. This proof shows that for for all possible values of $1-\pi_1 c_k+\pi_2 v$ where $k \in \{H,L\}$, the defender can improve its expected utility by lying to the attacker about its type.

  Formally, let β_L^* be the attacker's equilibrium probability of attacking when the players play the attribution game and the defender is type L and let β_H^* be the attacker's equilibrium probability of attacking when they play the attribution game and the defender is type H.

  - **Case 1:** $1-\pi_1 c_H+\pi_2 v<0$ **and** $1-\pi_1 c_L+\pi_2 v<0$. Consider when the defender truthfully signals s_L, indicating that it has a low capability. In this case, the attacker's equilibrium probability in the attribution game is $\beta_L^* = \frac{\pi_2 w}{\pi_1 r_L+\pi_2 w}$ and the defender's expected utility is $-\beta_L^*$. If instead of signaling s_L the defender signaled s_H, the attacker would randomize with probability $\beta_H^* = \frac{\pi_2 w}{\pi_1 r_H+\pi_2 w}$ and the defender's payoff would be $-\beta_H^* > -\beta_L^*$. Therefore the defender has an incentive to signal it is type H when it is truly type L and thus there is not a separating equilibrium when $1-\pi_1 c_H+\pi_2 v<0$ and $1-\pi_1 c_L+\pi_2 v<0$
  - **Case 2:** $1-\pi_1 c_H+\pi_2 v>0$ **and** $1-\pi_1 c_L+\pi_2 v>0$ In this regime, if the defender were to signal its true capability, the attacker would attack with probability $\beta_k^* = \frac{(1-\pi_2)w}{(1-\pi_1)r_k+(1-\pi_2)w}$ where k is either H or L depending on the defenders signal. The defender of type k has an expected utility of $\beta_k^*(r-1)-(1-\beta_k^*)w = \beta_k^*(r_k-1+w)-w$. If $(r_H-1+w)>0$, then the defender's

utility is increasing in the attack probability. Therefore when the defender is type H it would prefer that the attacker attack with a higher probability thus would signal that it is type L and induce the attacker to attack with probability $\beta_L^* > \beta_H^*$. Therefore, there cannot be a separating equilibrium if $(r_H - 1 + w) > 0$. Now suppose $(r_H - 1 + w) \leq 0$. Then it must be the case that $(r_L - 1 + w) < 0$, which implies that when the defender is type L, its expected utility is decreasing in the attack probability. This means that when the defender is truly type L it would prefer to signal that it was type H so that the attacker randomizes with probability $\beta_H^* < \beta_L^*$. As a result, there cannot be a separating equilibrium when $(r_H - 1 + w) \leq 0$.

– **Case 3:** $1 - \pi_1 c_H + \pi_2 v < 0$ **and** $1 - \pi_1 c_L + \pi_2 v > 0$ In this regime, if the defender is type H and signals such, the attacker randomizes with probability $\beta_H^* = \frac{\pi_2 w}{\pi_1 r_H + \pi_2 w}$ and the defender's expected utility when it is type H is $-\beta_H^*$. If the defender is type L and it signals such, the attacker randomizes with probability $\beta_L^* = \frac{(1-\pi_2)w}{(1-\pi_1)r_L + (1-\pi_2)w}$ and the defender earns an expected utility of $\beta_L * (\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ (which by indifference is the same as $\beta_L^*(r_L - 1 + w) - w$). Since $\pi_1 > \pi_2$ and $r_H > r_L$, it can be shown that $\beta_L^* > \beta_H^*$. Consider the defender's deviation of signaling that it is type L when it is actually type H and changing its strategy to retaliating after o_1 and not retaliating after o_2. In this case, the defender's utility is given by:

$$U_d((R, DR), \beta_L^*) = \pi_1 \beta_L^*(r_H - 1) - (1 - \pi_1)\beta_L^* - (1 - \beta_L^*)\pi_2 v \tag{5}$$

where $U_d((A, B), C)$ is the defender's expected utility from playing A after o_1 and B after o_2 when the attacker randomizes with probability C. For this deviation to not be profitable for the defender it must be that:

$$\begin{aligned}
U_d((DR, DR), \beta_H^*) &\geq & U_d((R, DR), \beta_L^*) \\
-\beta_H^* &\geq & \pi_1 \beta_L^*(r_H - 1) - (1 - \pi_1)\beta_L^* - (1 - \beta_L^*)\pi_2 v \\
-\beta_H^* &\geq & \beta_L^* \pi_1 r_H - \beta_L^* - \pi_2 w + \beta_L \pi_2 w \\
-\beta_H^* &\geq & \beta_L^* \pi_1 r_H - \beta_L^* - \beta_H^*(\pi_1 r_H + \pi_2 w) + \beta_L \pi_2 w \\
\beta_H^*(\pi_2 w + \pi_1 r_H - 1) &\geq & \beta_L^*(\pi_2 w + \pi_1 r_H - 1).
\end{aligned} \tag{6}$$

Since $\beta_H^* < \beta_L^*$ the inequality in equation 6 only holds if $\pi_2 w + \pi_1 r_H - 1 \leq 0$. So if $\pi_2 w + \pi_1 r_H - 1 > 0$, then the deviation is profitable and there is no incentive for the defender to truthfully signal its type. What remains to be shown is that the defender does not have an incentive to truthfully signal its type when $\pi_2 w + \pi_1 r_H - 1 > 0$. To do this, consider the defender's deviation of signaling it is type H when it is actually type L and retaliating after o_1 and not retaliating after o_2. Under this deviation, the defender's expected utility is

$$U_d((R, DR), \beta_H^*) = \beta_H^*(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w \tag{7}$$

For this deviation to not be profitable for the defender it must be that:

$$\begin{aligned}
U_d((R, R), \beta_L^*) &\geq U_d((R, R), \beta_H^*) \\
\beta_L^*(\pi_1 r_L - 1 + \pi_2 w) - \pi_2 w &\geq \beta_H^*(\pi_1 r_L - 1 + \pi_2 w) - \pi_2 w.
\end{aligned} \tag{8}$$

Since $\beta_L^* > \beta_H^*$, the only way for the inequality in equation 8 to hold is if $\pi_1 r_L - 1 + \pi_2 w \geq 0$. However, if $\pi_1 r_L - 1 + \pi_2 w \geq 0$ then $\pi_1 r_H - 1 + \pi_2 w \geq 0$ since $r_H > r_L$. But from the first part of case 3, if $\pi_1 r_H - 1 + \pi_2 w \geq 0$, then the defender would have an incentive to deviate when it is type L. Therefore, when $\pi_1 r_H - 1 + \pi_2 w \geq 0$, the defender has an incentive to deviate from truthful signaling and when $\pi_1 r_H - 1 + \pi_2 w \leq 0$, the defender has an incentive to deviate from truthful signaling. Ignoring the measure 0 case where $\pi_1 r_H - 1 + \pi_2 w = 0$, the defender always has an incentive to deviate from truthful signaling.

All three cases cover all possible parameter values and illustrate the for all values of the parameters, there is always a profitable deviation from truthful signaling for the defender and thus there is no separating equilibrium. □

- *Proof of Proposition 3.* In this case, if the attacker knows the defender is type L, then it is always a best response for the attacker to attack. As a result, it is always the defender's best response to retaliate when it truthfully signals it is type L. To show this cannot be an equilibrium, consider the following two cases.:
  - **Case 1**: Suppose $\pi_1 r_H + \pi_2 w - 1 > 0$. Under a separating equilibrium, when the defender signals it is type L, the attacker attacks with probability 1. Also, when the defender is type H and truthfully signals its type, its expected utility is $-\beta_H$ when $1 - \pi_1 c_H + \pi_2 v \leq 0$ and $\beta_H(r_H + w - 1) - w$ when $1 - \pi_1 c_H + \pi_2 v > 0$. Consider each of the two cases separately:
    - Suppose $1 - \pi_1 c_H + \pi_2 v > 0$. Since by assumption $\pi_1 r_H + \pi_2 w - 1 > 0$, then $r_H + w - 1 > 0$ and thus the attacker's expected utility is increasing in β_H when $1 - \pi_1 c_H + \pi_2 v > 0$. Therefore, if $1 - \pi_1 c_H + \pi_2 v > 0$ and the defender is type H, it would be better off signaling it is type L and thus there cannot be a separating equilibrium in which it truthfully signals its type.
    - Suppose $1 - \pi_1 c_H + \pi_2 v < 0$. Then when the defender is type H and truthfully signals its type, it's expected utility is $-\beta_H^* = \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w}$. If it were to instead switch its strategy by signaling that it is type L and always retaliating, it's expected utility is $r_H - 1$. For the defender to not have an incentive to make this switch it must be that:

$$\begin{aligned} \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w} &\geq r_H - 1 \\ \rightarrow 0 &\geq r_H(\pi_1 r_H + \pi_2 w - \pi_1) \end{aligned} \tag{9}$$

      However by assumption, $\pi_1 r_H + \pi_2 w - 1 > 0$ so it is impossible for the inequality in equation 9 to hold and this there cannot be a separating equilibrium.
  - **Case 2:** Suppose $\pi_1 r_H + \pi_2 w - 1 < 0$. Again, there are two sub-cases:
    - Suppose $r_L + w - 1 < 0$. The defender's expected utility by truthfully signaling when it is type L is $r_L - 1$. If instead it signaled it was type H and always retaliated, its expected utility would be $\beta_H^*(r_L + w - 1) - w$. For it to not gain anything from such a deviation it must be:

$$\begin{aligned} r_L - 1 &\geq \beta_H^*(r_L + w - 1) - w \\ \rightarrow 1 &\leq \beta_H^* \end{aligned} \tag{10}$$

      Since β_H^* is a probability less than 1, the inequality in equation 10 cannot hold and therefore there cannot be a separating equilibrium.
    - Suppose $r_L + w - 1 > 0$. If the defender signals that it is type L when it is type H and always retaliates, it will earn $r_H - 1$. If it signals it is type L when it is truly type L, it earns $r_L - 1$. If $1 - \pi_1 c_H + \pi_2 v < 0$, then when the defender signals it is type H, the attacker attacks with probability $-\beta_H^* = \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w}$. Two possible deviations the defender can make is 1) when L signal that it is type H and never retaliate and 2) when H signal that it is type L and always retaliate. For neither of these deviations to be profitable it must be:

$$\begin{aligned} r_H - 1 &< \beta_H^* \\ r_L - 1 &> \beta_H^* \end{aligned}$$

      Since $r_H > r_L$, it is impossible for both conditions to hold simultaneously. Finally, If $1 - \pi_1 c_H + \pi_2 v > 0$, then when the defender signals it is type H, the attacker attacks with probability $-\beta_H^* = \frac{-\pi_2 w}{\pi_1 r_H - \pi_2 w}$ and the defender earns an expected utility of $\beta_H^*(\pi_1 r_H + \pi_2 v - 1) - w$. If instead, the defender signaled it was type L when it is type H, and always retaliate it would earn $r_H - 1$. For there to be no incentive for the defender to deviate it must be that:

$$\begin{aligned} r_H - 1 &< \beta_H^*(\pi_1 r_H + \pi_2 v - 1) - w \\ \rightarrow \frac{r_H - 1 + w}{\pi_1 r_H + \pi_2 v - 1} &> \beta_H^* \end{aligned} \tag{11}$$

      Since by assumption $r_H - 1 + w > 0 > \pi_1 r_H + \pi_2 v - 1$, the left hand side of equation 11 is negative. Since β_H^* is a proper probability, such an equality can never be satisfied and thus the defender would have an incentive to deviate.

□

- *Proof of proposition 4.* First, recall the equilibrium probabilities from the attribution game and label them as:

$$\begin{aligned}\beta_{H1} &= \frac{\pi_2 w}{\pi_1 r_H + \pi_2 w} \\ \beta_{H2} &= \frac{(1-\pi_2)w}{(1-\pi_1)r_H + (1-\pi_2)w} \\ \beta_{L1} &= \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w} \\ \beta_{L2} &= \frac{(1-\pi_2)w}{(1-\pi_1)r_L + (1-\pi_2)w}\end{aligned} \tag{12}$$

Since $\pi_1 > \pi_2$ and $r_H > r_L$, it can be shown that $\beta_{H1} < \beta_{L1} < \beta_{H2} < \beta_{L2}$. By the same argument as in the attribution game, any equilibrium must have $\beta_{H1} < \beta_H < \beta_{L2}$ and $\beta_{H1} < \beta_L < \beta_{L2}$ The reason is that if any of the attacker's randomization probabilities are outside these bounds, the defender's best response, regardless of its type,is to either always retaliate or never retaliate, which cannot be part of an equilibrium (because then the attacker would no longer be willing to randomize). Given this fact, we will now prove each claim in the proposition.

  - **Proof of Part 1:** Under the strategy profile described in part 1, Bayes rule dictates that when the attacker receives signal s_H, it knows with probability 1 that the defender is type H. Therefore, when the attacker receives signal s_H, any equilibrium must have the players play the attribution game and the attacker attacks with probability $\beta_H = \beta_{H1}$ or $\beta_H = \beta_{H2}$, depending on the sign of $1 - \pi_1 c_H + \pi_2 v$. For the defender to be willing to randomize between signaling s_L and s_H, it must be indifferent between sending the two signals. We examine the cases separately.
    - Suppose $\beta_H = \beta_{H2}$. When the defender signals s_H, it is indifferent between retaliating after o_1 only and always retaliating. Thus, it's expected utility is $\beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w = \beta_H(r_H + w - 1) - w$. Let β_L be the probability the attacker attacks when it receives signal s_L.
      - Suppose $\beta_L < \beta_H$. If $\pi_1 r_H + \pi_2 w - 1 < 0$, then when the defender of type H signals s_L and only attacks after o_1, it earns $\beta_L(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w$ which is strictly greater than its expected utility from signaling s_H because $\beta_L < \beta_H$ and $\pi_1 r_H + \pi_2 w - 1 < 0$ by assumption. Therefore, the attacker would not be willing to randomize between signals when it is type H. If $\pi_1 r_H + \pi_2 w - 1 > 0$, then the attacker would not be willing to signal s_L and only attack after o_1 because $\beta_L < \beta_H$, and the payoff for only attacking after o_1 is increasing in β. Therefore, the only way the defender can be indifferent between signaling s_H and s_L is if β_L is such that the defender's best response is to never retaliates after signaling s_L. This condition is given by

$$\begin{aligned} -\beta_L &= \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w \\ -\beta_L &= \beta_H(\pi_1 r_H + \pi_2 w - 1) - \beta_{H1}(\pi_1 r_H + \pi_2 w) \\ \frac{\beta_H - \beta_L}{\beta_H - \beta_{H1}} &= \pi_1 r_H + \pi_2 w \end{aligned} \tag{13}$$

        For the condition in equation 13 to hold, it must be that $\beta_L < \beta_{H1}$ since by assumption $\pi_1 r_H + \pi_2 w > 1$. However, by the argument above, there is no equilibrium in which the attacker randomizes with a probability $\beta_L < \beta_{H1}$ so there cannot be an equilibrium with $\beta_L < \beta_H$.
      - Suppose $\beta_L > \beta_H$. This means that when the defender of type H signals s_L and the attacker randomizes with probability β_L, the defender's best response is to always retaliate. This implies that for all values of β such that $\beta_H < \beta < \beta_L$, the defender's expected

utility is either *strictly* increasing or *strictly* decreasing in β, depending on the sign of $r_H + w - 1$. Due to strict monotonicity of the defender's utility with respect to β, it cannot be indifferent between signaling β_H and β_L.

Since there cannot be an equilibrium with $\beta_H = \beta_{H2}$ and $\beta_L < \beta_H$ or $\beta_L > \beta_H$, there cannot be an equilibrium with $\beta_H = \beta_{H2}$.

* Suppose $\beta_H = \beta_{H1}$. In this case, the defender of type H is indifferent between never retaliating and retaliating afer o_1 only and earns an expected utility of $-\beta_H = \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w$. By the argument above β_L cannot be less than $\beta H1$. Therefore, suppose $\beta_L > \beta_H$. If $\pi_1 r_H + \pi_2 w - 1 > 0$, then the defender's expected utility of signaling s_L and only retaliating after o_1 is $\beta_L(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w$ which is greater than it's expected utility of after signaling s_H. Therefore, the defender of type H would not be willing to randomize between signaling s_L and s_H. If $\pi_1 r_H + \pi_2 w - 1 < 0$, the only way the defender can be indifferent between signaling s_H and s_L is if it always retaliates after signaling s_L. This implies

$$-\beta_H = \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w = \beta_L(r_H + w - 1) \tag{14}$$

where the first equality comes from the fact that at β_{H1} the attacker must be indifferent between never retaliating and retaliating after o_1 only. Now consider a defender of type L that always signals it is type L. In this case, it's utility is either $-\beta_L$, $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ or $\beta_L(r_L + v - 1)$, depending on which of its strategies are optimal at β_L. If a defender of type L instead signaled s_H and never retaliated, its expected utility would be $-\beta_H = \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w = \beta_L(r_H + v - 1)$. The following inequalities show that this regardless of which one of the defender's strategies is optimal at β_L, there exists a profitable deviation where the defender of type L signals s_H and never retaliates:

$$\begin{aligned} -\beta_L &< -\beta_H \text{ (By assumption)} \\ \beta_L(r_L + w - 1) - w &< \beta_L(r_H + w - 1) \text{ (Because } r_H > r_L \text{)} \\ \beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w &< \beta_H(\pi_1 r_H + \pi_2 w - 1) - \pi_2 w \end{aligned}$$

The last line follows because $r_H > r_L$ and $\pi_1 r_H + \pi_2 w - 1 < 0$. This shows that there cannot be a PBE where $\beta_H = \beta_{H1}$

Since there cannot be an equilibrium where the attacker randomizes with either β_{H1} or β_{H2} after observing s_H, there cannot be a PBE where a defender of type L always signals s_L and a defender of type H randomizes between signaling s_L and s_H.

– **Proof of Part 2:** Under the strategy profile described in part 2, Bayes rule dictates that when the attacker receives signal s_L, it knows with probability 1 that the defender is type L. Therefore, when the attacker receives signal s_L, any equilibrium must have the players play the attribution game and the attacker attacks with probability $\beta_L = \beta_{L1}$ or $\beta_L = \beta_{L2}$, depending on the sign of $1 - \pi_1 c_L + \pi_2 w$. For the defender to be willing to randomize between signaling s_L and s_H, it must be indifferent between sending the two signals. We examine the cases separately.

* Suppose $\beta_L = \beta_{L2}$. In this case, the defender of type L is indifferent between retaliating after o_1 only and always retaliating and earns $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w = \beta_L(r_L + w - 1) - w$. There cannot be an equilibrium where $\beta_H > \beta_L$ because then the defender's best response would be to always retaliate regardless of its type. Therefore, it is sufficient to only consider the case where $\beta_H < \beta_L$. If $\pi_1 r_L + \pi_2 w - 1 < 0$, then the defender of type L has a profitable deviation to signal it is type H and only retaliate after o_1 and earn $\beta_H(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ which is greater than $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$. Now suppose $\pi_1 r_L + \pi_2 w - 1 > 0$. Since there is no equilibrium where the attacker ever randomizes with a probability $\beta < \beta_{H1}$ it must be that $\beta_{H1} < \beta_H < \beta_L$. This means that the attacker's best response at β_H is either to retaliate after o_1 only or always retaliates. However, since $\pi_1 r_L + \pi_2 w - 1 > 0$ then $\pi_1 r_H + \pi_2 w - 1 > 0$ and $r_H + \pi_2 w - 1 > 0$ which means that the expected utility of the defender of type H is increasing in β and thus a defender of type H would prefer to signal s_L. Consequently, there cannot be an equilibrium where $\beta_L = \beta_{L2}$.

* Suppose $\beta_L = \beta_{L1}$. In this case, a defender of type L is indifferent between never retaliating and retaliating after o_1 only and earns $-\beta_L = \beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$. Suppose $\beta_H < \beta_L$, then there defender of type L would not be willing to randomize between s_L and s_H because it can signal s_H, never retaliate and earn $beta_H$, which is greater than its expected utility of $-\beta_L$ by signaling s_L. Now suppose $\beta_H > \beta_L$. If $\pi_1 r_L + \pi_2 w - 1 > 0$, the defender of type L can earn $\beta_H(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ when it signals s_H and only retaliates after o_1. Since this is higher than its maximum expected utility of $\beta_L(\pi_1 r_L + \pi_2 w - 1) - \pi_2 w$ when it signals s_L, a defender of type L would not be willing to randomize its signals. Lastly, if $\pi_1 r_L + \pi_2 w - 1 < 0$, the defender can only be indifferent between signaling s_H and s_L if its best response is to always retaliate when it is type L and signals s_H (because its expected utility of only retaliating after o_1 is strictly monotonic in β). However, if the defender's of type L's best response to an attacker randomizing with probability β_H is to always retaliate, it is also a defender of type H's best response to always retaliate. Since the defender always retaliating after a signal cannot be part of an equilibrium, there cannot be an equilibrium where $\beta_H > \beta_L$.

Since there cannot be an equilibrium where the attacker randomizes with either β_{L1} or β_{L2} after observing s_L, there cannot be an equilibrium where a defender of type H always signals s_H and a defender of type L randomizes between signaling s_L and s_H.

□

- *Proof of Proposition 5.* Begin by considering the attacker. If the attacker receives signal s_L, then Bayes rules necessitate it knows the defender is type L and thus the attacker and defender play the attribution game. As proposition 1 shows, there is an equilibrium in the attribution game where the attacker randomizes with probability $\frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ and the defender retaliates with probability $\frac{1}{\pi_1 c_L - \pi_2 v}$, which by assumption 2 is a proper probability. Now consider the attacker that receives signal s_H. For it to be willing to randomize, it must be indifferent between attacking and not attacking. Assuming the defender retaliates regardless of its type, this condition is given by:

$$\begin{aligned} P(H|s_H)(1-c_H) + P(L|s_H)(1-c_L) &= -v \\ \frac{P(s_H|H)P(H)(1-c_H)}{P(s_H|H)P(H) + P(s_H|L)P(L)} + \frac{P(s_H|L)P(L)(1-c_L)}{P(s_H|H)P(H) + P(s_H|L)P(L)} &= -v \\ \frac{\gamma(1-c_H)}{\gamma(1-c_H) + (1-\gamma)P(s_H|L)} + \frac{\gamma(1-c_L)P(s_H|L)}{\gamma(1-c_H) + (1-\gamma)P(s_H|L)} &= -v \\ P(s_H|L) = \frac{\gamma}{(1-\gamma)} \frac{1-c_H+v}{c_L-v-1} & \end{aligned} \tag{15}$$

By assumption, $1 - c_H + v$ and $c_L - v - 1$ are both less than 0, so the second fraction in equation 15 is positive. By assumption, 5, the entire right hand side of equation 15 is less than 1 and thus a proper probability.

Now consider the defender. To begin, consider a defender of type L. A necessary condition for the defender of type L to be willing to randomize between signaling 1) s_L and earning a payoff of $\frac{-\pi_2 w}{\pi_1 r_L + \pi_2 w}$ and randomizing between never retaliating and retaliating after o_1 only and 2) s_H, inducing the attacker to attack with probability β_H, and always retaliating, it must be indifferent between the two signals. This implies

$$\begin{aligned} \frac{-\pi_2 w}{\pi_1 r_L + \pi_2 w} &= \beta_H(r_L + w - 1) - w \\ \beta_H &= \frac{w(\pi_1 r_L + \pi_2 w - \pi_2)}{(r_L + w - 1)(\pi_1 r_L + \pi_2 w)} \end{aligned} \tag{16}$$

Since, $\pi_1 r_L + \pi_2 w < 1$ by assumption, the defender's expected utility from retaliating after o_1 only is decreasing in β and thus, the defender's best response at β_H is to always retaliate. Therefore, the defender of type L is indifferent between signaling s_H and s_L. Finally, consider the defender

of type H. When the attacker randomizes with probability β_H, because it is a defender's of type L best response to always retaliate, it must also be a defender of type H's best response to always retaliate. The defender's payoff from always retaliating is $\beta_H(r_H + w - 1) - w = -\beta_L + \beta_H(r_H - r_L)$. What remains to be shown is that a defender of type H does not have an incentive to signal s_L. If the defender signals s_L and always retaliates, its payoff must be less than if it signals s_H because its payoff to always retaliaing is increasing in β. It's payoff by signaling s_L and never retaliating is $\frac{-\pi_2 w}{pi_1 r_L + \pi_2 w} = \beta_H(r_L + w - 1) - w < \beta_H(r_H + w - 1) - w$. Finally, it's payoff of signaling s_L and retaliating after o_1 only is $\beta_L(\pi_1(r_H - r_L) - 1)$ which is strictly less than $-\beta_L + \beta_H(r_H - r_L)$. Therefore, the defender does not have an incentive to change its strategy. □

No other semi-separating equilibrium. First, we will show that there is no equilibrium in which the defender randomizes its signal for each of its types. Then we will show there is no equilibrium in which the defender randomizes only when it is type H.

For contractiction, suppose there is a a signaling equilibrium where the defender randomizes its signal for each of its types and induces the attacker to randomize with probability β_H and β_L and without loss of generality, assume $\beta_H > \beta_L$. There is no equilibrium where β_L is so low that the attacker of type H would never retaliate. Therefore, the defender of type H must be indifferent between retaliating after o_1 only and always retaliating. This implies

$$\beta_L(\pi_1 r_H + \pi_2 w - 1) = \pi_2 w = \beta_H(r_H + w - 1) - w \tag{17}$$

of course, this can only happen when $(\pi_1 r_H + \pi_2 w - 1) < 0$ and $(r_H + w - 1)$. For a defender of type L to be indifferent, there are two cases.

- Consider the case where the defender of type L is indifferent between retaliating only after o_1 when the attacker attacks with probability β_L and always retaliating when the attacker attacks with probability β_H. This implies
$$\beta_L(\pi_1 r_L + \pi_2 w - 1) = \pi_2 w = \beta_H(r_L + w - 1) - w \tag{18}$$
However, solving for equations 17 and 18 yields $\beta_H = \pi_1 \beta_L$ which violates the fact that $beta_H > \beta_L$.

- Consider the case where the defender of type L is indifferent between never retaliating when the attacker attacks with probability β_L and always retaliating when the attacker attacks with probability $beta_H$. This implies
$$-\beta_L = \beta_H(r_L + w - 1) - w \tag{19}$$
Equation 17 and 19 together imply that
$$\beta_L = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w} + (r_H - r_L)(\beta_H - \pi_1 \beta_L) \tag{20}$$
However, the soluton to equation 20 yields a value of $\beta_L > \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$, which cannot be part of an equilibrium because at such a value of β_L, the defender would prefer to retaliate after o_1.

Now consider the semi-separating strategy where the defender randomizes its signal when it is type H and always signals s_L when it is type L. If the attacker randomizes its signal when it is type H, then Bayes rule will dictate that when the attacker receives signal s_H, it knows the attacker is type H with certainty; Let β_H be the attacker's randomization probability when it receives signal s_H. Since the attacker knows the defender's type when the defender signals s_H, the players play the attribution game and therefore the attacker either randomizes with $\beta_H = \frac{\pi_2 w}{\pi_1 r_h + \pi_2 w}$ or $\beta_H = \frac{(1-\pi_2)w}{(1-\pi_1)r_H + (1-\pi_2)w}$. We consider these cases separately.

- Suppose $\beta_H = \frac{\pi_2 w}{\pi_1 r_L + \pi_2 w}$ where the defender of type H is indifferent between never retaliating and retaliating after o_1 only. For the defender of type H to be willing to randomize its signal, it must be that the defender is indifferent between signaling s_H and signaling s_L, inducing the attacker to attack with probability β_L and always retaliating. However, at such a β_L, the defender of type L's expected utility for any of it strategies is strictly less than its expected utility from signaling s_H and never retaliating, therefore, the defender would never be willing to signal s_L.

- Suppose $\beta_H = \frac{(1-\pi_2)w}{(1-\pi_1)r_H+(1-\pi_2)w}$ where the defender of type H is indifferent between retaliating after o_1 only and always retaliating. For the defender of type H to be willing to randomize its signal, it must be that the defender is indifferent between signaling s_H and signaling s_L, inducing the attacker to attack with probability β_L and never retaliating. However, at such a value of β_L, the defender of type H or type L would ever retaliate and thus the attacker wouldn't be willing to randomize but instead would attack with probability 1. Therefore, there can't be an equilibrium in which $\beta_H = \frac{(1-\pi_2)w}{(1-\pi_1)r_H+(1-\pi_2)w}$.

Finally consider the semi-separating strategy where the defender randomizes its signal when it is type L and always signals s_H when it is type H. If the attacker randomizes its signal when it is type L, then Bayes rule will dictate that when the attacker receives signal s_L, it knows the attacker is type L with certainty. Let β_L be the attacker's randomization probability when it receives signal s_L. Since the attacker knows the defender's type when the defender signals s_L, the players play the attribution game and therefore the attacker either randomizes with $\beta_L = \frac{\pi_2 w}{\pi_1 r_L+\pi_2 w}$ or $\beta_H = \frac{(1-\pi_2)w}{(1-\pi_1)r_L+(1-\pi_2)w}$. Our main proposition showed that there can be an equilibrium when $\beta_L = \frac{\pi_2 w}{\pi_1 r_L+\pi_2 w}$ so here, we consider the case where $\beta_L = \frac{(\pi_2-)w}{(1-\pi_1)r_L+(1-\pi_2)w}$. The only way the defender can be indifferent between the attacker attacking with probability β_L and β_H is if $\pi_1 r_L+\pi_2 w-1 > 0$ and the defender of type L never retaliates at $\beta_H < \beta_L$. However, since $\pi_1 r_L+\pi_2 w-1 > 0$, the attacker of type H's expected utility is increasing in β and therefore would prefer to signal s_L and not s_H.

All of the cases show that there are no other semi-separating equilibria. □