Deterrence with Imperfect Attribution*

Sandeep Baliga
Kellogg SOM, Northwestern University

Ethan Bueno de Mesquita
Harris School, University of Chicago

Alexander Wolitzky
Department of Economics, MIT

May 2, 2018

Abstract

Motivated by cyberwarfare, we study deterrence when attacks cannot be perfectly attributed. Each of n attackers may attack the defender. The defender observes a signal that probabilistically attributes the attack. The defender may retaliate against one or more attackers but retaliating against innocent attackers is costly. We uncover an endogenous strategic complementarity among the attackers: if one attacker becomes more aggressive, *all* attackers become more aggressive. Improving the defender's ability to detect attacks or identify the attacker can trigger more attacks, but simultaneously improving detection and identification reduces attacks. If the defender can commit, she might retaliate *less* after some signals.

*We have received helpful comments and feedback from Daron Acemoglu, Scott Ashworth, Wiola Dziuda, Drew Fudenberg, Roger Lagunoff, Konstantin Sonin, and seminar audiences at Chicago, Georgetown, Political Economy in the Chicago Area (P.E.C.A.) conference and UBC. Zhaosong Ruan provided excellent research assistance.

"Whereas a missile comes with a return address, a computer virus generally does not."

–William Lynn, U.S. Deputy Secretary of Defense, 2010

The ability to maintain peace through deterrence rests on a simple principle: the credible threat of sufficiently strong retaliation in response to an attack prevents forward-looking adversaries from launching attacks in the first place (Schelling [49], Snyder [55], Myerson [38]). The traditional concern about the effectiveness of deterrence is that retaliation might not be credible. But technological changes, especially the rise of cyberwarfare, have brought a new set of considerations to the fore. Central among these new issues is the *attribution problem*: the potential difficulty in determining who is responsible for an attack, or even if an attack occurred at all.

Obviously, attribution problems weaken deterrence "by reducing an assailant's expectation of unacceptable penalties" (Kello [29] p.130; see also Clark and Landau [14], Edwards et al. [17], Goldsmith [20], Lindsay [35], and Nye [40]): multiplying a penalty by the probability of correct attribution reduces the expected penalty. But the effects of imperfect attribution on deterrence are much richer than this, and the precise effects—as well as how a state can optimally deter attacks under imperfect attribution—have yet to be studied. As General Michael Hayden [23], former director of the National Security Agency, put it in testimony before Congress, "[c]asually applying well-known concepts from physical space like deterrence, where attribution is assumed, to cyberspace, where attribution is frequently the problem, is a recipe for failure." The current paper takes up Hayden's challenge by analyzing deterrence under imperfect attribution.

While attribution problems are endemic to cyberwarfare, they also arise in many other environments where deterrence matters. Even in conventional warfare, it can be difficult to determine who initiated a given attack.[1] The problem is amplified is counterinsurgency, where there is often uncertainty as to which of multiple terrorist or insurgent factions is responsible for an attack (Berman et al. [9], Shaver and Shapiro [52], Trager and Zagorcheva [58]). Turning to non-conflict environments, it is possible to measure pollution, but it may be difficult to assign responsibility to one potential polluter over another (Weissing and Ostrom [61]). Similar issues can arise in other areas of law and economics (Png [42], Lando [33]). Without minimizing these alternative applications, the current paper focusses on cyberwarfare.

We offer a model of deterrence with imperfect attribution with multiple potential attackers and one defender. An attacker gets an opportunity to strike the defender. The defender observes a noisy

[1] For example, the soldiers who entered Ukraine in March 2014 wore no insignia, and Russia initially denied involvement (Shevchenko [53]).

signal, which probabilistically indicates whether an attack occurred and who attacked. Attribution problems include three kinds of potential mistakes. There is a *false alarm* if the defender perceives an attack when no attack occurred. There is a *detection failure* if the defender fails to detect an attack that did occur. And there is *misidentification* if the defender assigns responsibility for an attack to the wrong attacker. In our model, the defender suffers a cost if she is attacked. She receives a private benefit that defrays some of this cost if she retaliates against the right attacker, but she suffers an additional cost if she retaliates against the wrong one. Each attacker gets a private benefit from attacking but suffers a cost if the defender retaliates against him. There are no direct externalities among attackers—one attacker's payoff does not depend on whether another attacks or faces retaliation.

The main mechanism our analysis uncovers is that the attribution problem generates an endogenous strategic complementarity among the potential attackers. As will become clear, this effect makes deterrence under imperfect attribution inherently multilateral. To see the idea, suppose attacker i becomes more aggressive. Then the defender's belief that attacker i was responsible increases for every signal where she detects an attack, and her belief that any other potential attacker was responsible decreases. This makes the defender more likely to retaliate against attacker i and less likely to retaliate against all other attackers. But this in turn leads the other attackers to become more aggressive—in effect, all other attackers can "hide behind" the aggressiveness of attacker i. Thus, a rise in the aggressiveness of a single attacker increases the probability with which every attacker attacks in equilibrium (and in particular increases the overall probability of an attack). Note that these externalities among attackers are purely due to the attribution problem, as there is no direct connection among the attackers' payoffs. This phenomenon is new to the theory of deterrence.

Our main results explore the implications of this mechanism for how changes in the environment—including the defender's information regarding attribution and the defender's ability to commit to a retaliatory strategy—affect deterrence.

Given the challenges attribution problems create for deterrence, a standard intuition is that improving attribution will improve deterrence. We establish several results showing that this is not always the case: sometimes, improving the defender's information *weakens* deterrence. First, reducing detection failures is not always beneficial. This is because reducing detection failures—distinguishing an attack by i from no attack—can exacerbate misidentification—making it more difficult to distinguish an attack by i from an attack by j. This can result in all attackers becoming

more aggressive, and hence in an increase in the total probability of an attack. Second, reducing misidentification is not always beneficial either. This is because increasing the likelihood of getting a clear signal when i attacks can make the defender less likely to retaliate after an ambiguous signal, which can reduce the overall probability of retaliation following an attack. This can again make all attackers more aggressive and thus make attacks more likely.

We also provide positive results, which characterize informational improvements that do always strengthen deterrence. Although improving detection or identification alone need not improve deterrence, we show that deterrence *is* always improved by simultaneously strengthening both detection and identification—in the sense that some attacks that previously went undetected are now unambiguously attributed to the perpetrator. Deterrence is also strengthened by reducing false alarms.

Next, we characterize the optimal deterrence policy when the defender can commit to a retaliatory strategy in advance. We show that deterrence is stronger when the defender can commit, in that every attacker attacks with lower probability. However, the defender should not necessarily commit to retaliate more after *every* signal. This is because the signals' information content changes when the attackers become less aggressive. Specifically, under commitment there may be a greater chance of misidentification or false alarm after some signals, and the defender may want to back off after such signals. In general, the optimal policy balances the commitment to large punishments from traditional deterrence theory à la Schelling with the risk of retaliating in error as discussed in the newer informal literature on cyberwarfare à la Kello [29] or Singer and Friedman [51].

Finally, we briefly consider an extension in which one attacker, a *provocateur*, benefits when the defender retaliates against another attacker. We show that introducing benefits from provoking the defender only strengthens our results when there are two possible attackers, but that matters are more complex with three or more attackers.

By way of further motivation for our model, we note that false alarms, detection failures, and misidentification have all arisen in major cyber incidents.

The Stuxnet worm was used to disrupt the Iranian nuclear facility at Natanz by causing centrifuges to malfunction over the course of more than a year. During the attack, the Iranians believed the problems with their centrifuges were the result of faulty parts, engineering incompetence, or domestic sabotage (Singer and Freedman [51]). Stuxnet was eventually uncovered not by the Ira-

nians, but by European cybersecurity researchers who found a worm that was infecting computers all over the world but was configured to do damage only in very specific circumstances tailored to the facility at Natanz. This was a case of *detection failure*.

In 1998, the United States Department of Defense discovered a series of attacks exploiting operating system vulnerabilities to retrieve large amounts of sensitive data from military computer networks. The United States was preparing for possible military action in support of UN weapons inspections in Iraq, and the cyberattacks emanated from Abu Dhabi. A Department of Defense investigation, referred to as Solar Sunrise, initially attributed the attacks to Iraq. The U.S. went so far as to send a strike team to Abu Dhabi, only to find a room full of computer servers. Ultimately, the attacks turned out to be the work of two sixteen-year olds in San Francisco and an eighteen-year old Israeli (Adams [3], Kaplan [28]). Conversely, the hacking of the Democratic National Committee servers during the 2016 U.S. Presidential election was initially attributed to a lone Romanian hacker who went by the moniker Guccifer 2.0. Later, U.S. authorities determined the hacking was done by Russian security agencies who had tried to cover their tracks by pretending to be Guccifer 2.0 (see the findings of the cybersecurity company ThreatConnect [57]). These are cases of *misidentification.*

Finally, in 2008, a worm on Department of Defense computers was found to have gained access to enormous quantities of US war planning materials. The leading theory to emerge from the resulting clean up and forensic operation, known as Buckshot Yankee, was that the worm was the work of a foreign intelligence agency (probably Russian) that infiltrated the "air gap" surrounding military computer networks through a USB drive sold to an American soldier in Afghanistan. In response, the Department of Defense banned all USB drives for years. But others point to the worm's relative unsophistication and argue it could have accidentally made its way onto the computer networks without malicious intent (Shachtman [50]). This may, then, have been a case of a *false alarm.*

The key mechanism of our model—"less suspect" attackers' desire to hide their attacks behind "more suspect" attackers—is also reflected in several incidents. According to American authorities, the Russian military agency GRU executed a cyberattack during the opening ceremony of the Pyeongchang Winter Olympics. In a so-called "false flag" operation, the GRU used North Korean IP addresses to make the hack appear to be the work of North Korea (Nakashima [39]). The GRU hoped to deflect suspicion onto North Korea, which was already highly suspect because of its hack of Sony Pictures and a variety of other cyber operations. Looking ahead, the International Olympic Committee (I.O.C.) can respond by expelling one or more countries from the next Olympics. As

in our model, it is natural to think the I.O.C. loses on net if it expels an innocent country, but not if it expels the guilty one. We discuss additional examples after presenting the model.

This paper relates to several literatures. Of course, a large literature explores aspects of deterrence other than the attribution problem. Schelling [49] explained the logic of deterrence and the importance of commitment. Jervis [27] pointed out that, when the motives of a player who acquires arms for deterrence are not known to his opponent, the opponent may react by acquiring arms to protect himself from predation. This "security dilemma" may equally apply to cyberweapons (Buchanan [11]). The security dilemma has been formalized using the idea that arms might be strategic complements rather than substitutes (Kydd [31], Baliga and Sjöström [7], Chassang and Padró i Miquel [12]). For example, Chassang and Padró i Miquel [12] show that, in a coordination game, arms acquisition can increase a preemptive incentive to go to war faster than it reduces the incentive to predate. Hence, arms acquisition may cause escalation rather than deterrence. Acemoglu and Wolitzky [2] incorporate an attribution problem into a dynamic coordination game with overlapping generations. A player does not know whether an ongoing conflict was started by the other "side" or by a past member of his own side. This leads to cycles of conflict as players occasionally experiment with peaceful actions to see if the other side plays along.[2] Another literature explores the search for credibility, especially the role played by domestic politics (see, for example, Fearon [18], Powell [44], Smith [54], Di Lonardo and Tyson [60]). We abstract from these themes in order to focus on the implications of attribution problems for deterrence with multiple attackers.

Our model also relates to the literature on inspection games. In such a game, an inspectee may or may not act legally, and an inspector decides whether to call an alarm as a function of a signal of the inspectee's action (see Avenhaus, von Stengel and Zamir [5] for a survey). This literature usually allows only one inspectee, though some of our comparative statics results also apply to that case. In particular, we show by example that an improvement in information in the sense of Blackwell can make the defender worse off (without commitment)—this appears to be a novel result in the inspection game literature. Some inspection game models do allow multiple inspectees, but these models study issues other than attribution, such as the allocation of scarce detection resources across sites (Avenhaus and Kilgour [5], Hohzaki [24]).

Inspection games appear in economics in the guise of "auditing games," where a principal tries to

[2] Rohner, Thoehig and Zillibotti [46] study the impact of trust on trade in a two period game where one player learns whether the other side is aggressive through his first period action. There is no attribution problem in their model.

catch agents who "cheat." These games have many interesting features. For example, the principal might commit to random audits to save on auditing costs (Mookherjee and Png [37]). The principal also faces a commitment problem, as she may not have an incentive to monitor the agent ex post (Graetz, Reinganum and Wilde [21], Khalil [30]). However, the attribution problem we study does not arise in these models.

In law and economics, there is a question of whether deterrence is undercut by the fact that even the innocent might be convicted (see Lando [33] and Section 8 of Polinsky and Shavell [43]). This approach assumes full commitment to fines and subsidies. More importantly, it does not fully formalize the strategic setting as a multi-player game, so key properties like the strategic complementarity of equilibrium actions cannot be uncovered. Indeed, the attackers in our model can be interpreted as criminals and the principal as a judge who seeks to punish the guilty but not the innocent. Hence, our model or some variant thereof might be of interest in law and economics.[3]

Finally, repeated games with imperfect monitoring model multilateral moral hazard without commitment (Radner [45], Green and Porter [22], Abreu, Pearce, and Stacchetti [1]). Our model collapses the infinite horizon into a principal who plays a best response. This approach might also be a useful shortcut in other contexts. For example, Chassang and Zehnder [13] study a principal with social preferences who cannot commit to a contract and instead makes an ex post transfer from an active agent to a passive agent towards whom the active agent may or may not have taken a pro-social action. Their approach is an alternative to relational contracting models of intertemporal incentives (Baker, Gibbons, and Murphy [6]).

1 Model

There are $n+1$ players: n attackers and one defender. They play a two-stage game:

1. With probability $\gamma \in (0,1)$, one of the n attackers is randomly selected. That attacker chooses whether or not to attack. With probability $1-\gamma$, no one has an opportunity to attack.

2. The defender observes a signal s drawn from a finite set S. If attacker i attacked in stage 1, the probability of signal s is π_i^s. If no one attacked in stage 1 (i.e., if some attacker had an opportunity to attack but chose not to, or if no one had an opportunity to attack), the

[3] The one-inspectee inspection game also arises in law and economics. Specifically, Tsebelis [59] studies costly monitoring by the police. The police cannot commit to monitoring effort, so in equilibrium the police mix between working and shirking and criminals mix between criminality and law-abidingness.

probability of signal s is π_0^s. The defender then chooses whether to retaliate against one or more of the attackers.

The attackers differ in their aggressiveness. An attacker with aggressiveness $x_i \in \mathbb{R}$ receives a payoff of x_i if he attacks. Each attacker also receives a payoff of -1 if he is retaliated against. Each attacker i's aggressiveness x_i is his private information and is drawn from a continuous distribution F_i with bounded support.

The defender receives a payoff of $-K$ is she is attacked. In addition, for each attacker i, if she retaliates against i she receives an additional payoff of $y_i \in \mathbb{R}_+$ if i attacked and receives an additional payoff of $y_i - 1$ if i did not attack. The vector $y = (y_i)_{i=1}^n$ is the defender's private information and is drawn from a continuous distribution G with bounded support and marginals $(G_i)_{i=1}^n$. We assume that $G_i(K) = 1$ for all i. This implies that the defender would rather not be attacked than be attacked and successfully retaliate.

In general, a strategy for attacker $i \in I := \{1, .., n\}$ is a mapping from his aggressiveness x_i to his probability of attack, $p_i(x_i) \in [0, 1]$. A strategy for the defender is a mapping from $y = (y_i)_{i \in I}$ and the signal s to the probability with which she retaliates against each attacker, $r^s(y) = (r_i^s(y))_{i \in I} \in [0, 1]^n$.[4] However, it is obvious that every best response for both the attackers and the defender takes a cutoff form, where attacker i attacks if and only x_i exceeds a cutoff $x_i^* \in [0, 1]$, and the defender retaliates against attacker i after signal s if and only if y_i exceeds a cutoff $y_i^{s*} \in [0, 1]$.[5] We can therefore summarize a strategy profile as a vector of cutoffs $(x^*, y^*) \in [0, 1]^n \times [0, 1]^{n|S|}$. Equivalently, we can summarize a strategy profile as a vector of attack probabilities $p \in [0, 1]^n$ for the attackers and a vector of retaliation probabilities $r \in [0, 1]^{n|S|}$ for the defender, as for attacker i choosing attack probability p_i is equivalent to choosing cutoff $x_i^* = F_i^{-1}(1 - p_i)$, and for the defender choosing retaliation probability r_i^s is equivalent to choosing cutoff $y_i^{s*} = G_i^{-1}(1 - r_i^s)$.

The solution concept is sequential equilibrium (*equilibrium* henceforth).

We assume that S contains a "null signal," $s = 0$, which probabilistically indicates that no attack has occurred. The interpretation is that $s = 0$ corresponds to the defender failing to detect an attack. We make the following two assumptions.

1. For each player i, the probability of each non-null signal $s \neq 0$ is greater when i attacks than

[4] We implicitly assume that the defender's $-K$ payoff from being attacked is either measurable with respect to her signals or arrives after she decides whether to retaliate, so that any actionable information the defender receives from her payoff is encapsulated by the signals.

[5] Behavior at the cutoff is irrelevant as F_i and G_i are assumed continuous. Our mains results go through when F_i and G_i admit atoms, but the exposition is slightly more complicated.

when no one attacks: for all $i \in I$ and all $s \neq 0$,

$$\pi_i^s \geq \pi_0^s.$$

Note that this implies $\pi_i^0 \leq \pi_0^0$ for all $i \in I$, as $(\pi_i^s)_{s \in S}$ and $(\pi_0^s)_{s \in S}$ must sum to 1.

2. It is not optimal for the defender to retaliate after receiving the null signal: for all $i \in I$,

$$G_i\left(\frac{n\left(1-\gamma\right)\pi_0^0}{n\left(1-\gamma\right)\pi_0^0+\gamma\pi_i^0}\right)=1. \tag{1}$$

Note that this implies $y_i < 1$ with probability 1, so the defender never benefits from retaliating against an innocent attacker.

Finally, we assume that either (i) $\pi_0^s > 0$ for all $s \in S$, or (ii) $F_i\left(1\right) < 1$ for all $i \in I$ and $S = \bigcup_{i \in I \cup \{0\}, s \in S} \operatorname{supp} \pi_i^s$. Either assumption guarantees that every signal $s \in S$ arises with positive probability in equilibrium (and hence the defender's beliefs are determined by Bayes' rule), which is the only role of this assumption.

We offer a few comments on the interpretation of the model before turning to its analysis.

First, the assumption that $y_i \geq 0$ implies that retaliation would be credible for the defender if she knew who attacked. We thus abstract from the classic "search for credibility" in the traditional deterrence literature (Schelling [49], Snyder [55], Powell [44]) to isolate the new issue of how imperfect attribution affects deterrence. In general, the benefits of successful retaliation may include the disruption of an ongoing attack, reputational benefits vis a vis other potential attackers, or a taste for vengeance.

Second, as the Stuxnet attack highlights, it is possible for a cyberattack to occur without the defender recognizing she is under attack. The model captures the possibility of such detection failures through the null signal.

The presence of the null signal is also important for the strategic complementarity at the heart of our model. By Assumption 1, when attacker i becomes more aggressive, he becomes more "suspect" after every non-null signal and becomes less suspect after the null signal. By Assumption 2, this increases retaliation against attacker i and decreases retaliation against all other attackers, as retaliation occurs only following non-null signals.

Third, we consider a static model where at most one potential attacker has an opportunity to attack. This approach is equivalent to considering the Markov perfect equilibrium in a continuous-

time dynamic model where, for each attacker, an independent and identically distributed Poisson clock determines when that attacker has an attack opportunity. As the probably that independent Poisson clocks tick simultaneously is zero, in such a model it is without loss of generality to assume that two attackers can never attack at exactly the same time.

Finally, the payoff functions admit several different interpretations. We have normalized both the cost to an attacker of facing retaliation and the cost to the defender of retaliating in error to 1. This means that x_i and y measure the benefit of a successful attack/retaliation *relative* to the cost of facing retaliation/retaliating in error. Thus, an increase in x_i (for example) can represent either an increase in the benefit of attacking or a decrease in the cost of facing retaliation.

There are of course a variety of benefits from successful cyberattacks. The Chinese were able to use cyber espionage to acquire plans for the F-35 stealth fighter from a US military contractor, allowing them to build a copy-cat stealth fighter at accelerated speed and low cost. The United States and Israel were able to use cyberattacks to disrupt the Iranian nuclear program. Cyberattacks have also been used to incapacitate an adversary's military capabilities—for instance by disrupting communications—by the United States (against Iraqi insurgents), Russia (in Ukraine, Georgia, and Estonia), Israel (against Syrian air defenses), and others. To the extent that retaliation to cyberattacks remains within the cyber domain, variation in the costs of retaliation could derive from variation in the vulnerability of a country's civil or economic infrastructure to cyberattack. Thus, for example, North Korea may be more aggressive in the cyber domain than the United States because it does not have a vulnerable tech industry that could be disrupted by cyber retaliation. Finally, as technologies for hardening targets, denying access, and improving security improve, the distribution of benefits may worsen (Libicki, Ablon, and Webb [34]).

Similarly, an increase in y can represent either an increase in the benefit of successful retaliation or a decrease in the cost of retaliating in error. We have already mentioned several benefits of successful retaliation. A change in y might result from technological innovations that alter the extent to which the damage from an attack can be mitigated, or from political, economic, or strategic shifts that affect the value of reputation, the risk of escalation, or the potential for spillovers to civilian domains. A decrease in the cost of retaliating in error might result from either a decreased fear of escalation beyond the cyber domain or a technological shift that allowed for more targeted retaliation, among other possibilities.

2 Equilibrium Characterization

In this section, we characterize equilibrium and show that the attackers' strategies are *endogenous strategic complements*: if one attacker attacks with higher probability, they all attack with higher probability. This complementarity is the key mechanism uncovered by our analysis.

We first characterize the attackers' cutoffs x^* as a function of the defender's retaliation probabilities r (all missing proofs are in the Appendix). The following formula for x_i^* results because an attack by i increases the probability of each signal s by $\pi_i^s - \pi_0^s$, and thus increases the total probability that i faces retaliation by $\sum_s (\pi_i^s - \pi_0^s) r_i^s$.

Lemma 1 *In every equilibrium, for every $i \in I$, attacker i's cutoff is given by*

$$x_i^* = \sum_s (\pi_i^s - \pi_0^s) r_i^s. \tag{2}$$

Next, we characterize the defender's cutoffs y^* as a function of the attackers' attack probabilities p. Note that, if attacker i attacks with probability p_i when given the opportunity, his unconditional probability of attacking is $\frac{\gamma}{n} p_i$. Therefore, given a vector of (conditional) attack probabilities $p \in [0,1]^n$, the probability that attacker i attacked conditional on signal s equals

$$\beta_i^s(p) = \frac{\gamma p_i \pi_i^s}{\gamma \sum_j p_j \pi_j^s + \left(n - \gamma \sum_j p_j\right) \pi_0^s}. \tag{3}$$

Lemma 2 *In every equilibrium, for every $i \in I$ and $s \in S$, the defender's cutoff is given by*

$$y_i^{s*} = 1 - \beta_i^s(p). \tag{4}$$

We also note that the defender never retaliates after the null signal, by Assumptions 1 and 2.

Lemma 3 *In every equilibrium, $r_i^0 = 0$ for all $i \in I$.*

Out first result combines Lemmas 1, 2, and 3 to give a necessary and sufficient condition for a vector of attack and retaliation probabilities $(p, r) \in [0,1]^n \times [0,1]^{n|S|}$ to be an equilibrium.

Proposition 1 *A vector of attack and retaliation probabilities (p, r) is an equilibrium if and only*

if

$$F_i^{-1}(1-p_i) = \sum_{s\neq 0}(\pi_i^s - \pi_0^s)\left(1 - G_i\left(1-\beta_i^s(p)\right)\right) \tag{5}$$

$$= \sum_{s\neq 0}(\pi_i^s - \pi_0^s)\left(1 - G_i\left(\frac{n\pi_0^s + \gamma\sum_{j\neq i}p_j\left(\pi_j^s - \pi_0^s\right) - \gamma p_i \pi_0^s}{n\pi_0^s + \gamma\sum_{j\neq i}p_j\left(\pi_j^s - \pi_0^s\right) + \gamma p_i\left(\pi_i^s - \pi_0^s\right)}\right)\right) \tag{6}$$

and

$$r_i^s = 1 - G_i\left(1-\beta_i^s(p)\right)$$

for all $i \in I$ *and* $s \in S$.

Equation (5) is key for understanding our model. The left-hand side is attacker i's cutoff (recall, $x_i^* = F_i^{-1}(1-p_i)$). The right-hand side is the increase in the probability that attacker i faces retaliation when he attacks, noting that the probability that an attacker faces retaliation after any signal equals the probability that the defender's propensity to retaliate (y_i) exceeds the probability that the attacker did not attack conditional on the signal ($y_i^{s*} = 1-\beta_i^s(p)$). Equilibrium equates these two quantities.

Our central result on endogenous strategic complementarity now follows from the fact that $\beta_i^s(p)$ is increasing in p_i and decreasing in p_j for all $j \neq i$. To see the idea, suppose attacker i attacks with higher probability: p_i increases. This makes attacker i more "suspect" after every non-null signal and makes every attacker $j \neq i$ less suspect: for every $s \neq 0$, β_i^s increases and β_j^s decreases. In turn, this makes the defender retaliate more against i and less against j: for every $s \neq 0$, r_i^s increases and r_j^s decreases. Finally, this makes attacker j attack with higher probability: x_j^* decreases. Intuitively, when one attacker becomes more likely to attack, this makes the other attackers attack with higher probability, as they know their attacks are more likely to be attributed to the first attacker, making it less likely that they will face retaliation following an attack. This complementarity is the key multilateral aspect of deterrence with imperfect attribution.

To formalize this endogenous strategic complementarity, it is useful to introduce a new function.

Definition 1 *The* endogenous best response function $h : [0,1]^n \to [0,1]^n$ *is defined by letting* $h_i(p)$ *be the unique solution* $p_i' \in [0,1]$ *to the equation*

$$p_i' = 1 - F_i\left(\sum_{s\neq 0}(\pi_i^s - \pi_0^s)\left(1 - G_i\left(\frac{n\pi_0^s + \gamma\sum_{j\neq i}p_j\left(\pi_j^s - \pi_0^s\right) - \gamma p_i'\pi_0^s}{n\pi_0^s + \gamma\sum_{j\neq i}p_j\left(\pi_j^s - \pi_0^s\right) + \gamma p_i'\left(\pi_i^s - \pi_0^s\right)}\right)\right)\right) \tag{7}$$

for all $i \in I$*, and letting* $h(p) = (h_i(p))_{i \in I}$.

Intuitively, if the attack probabilities of all attackers other than i are fixed at $p_{-i} \in [0,1]^{n-1}$, then $h_i(p)$ is the unique equilibrium attack probability for attacker i in the induced two-player game between attacker i and the defender. Note that $h_i(p)$ is well-defined, as the right-hand side of (7) is always between 0 and 1 and is continuous and non-increasing in p_i'. Note also that $p \in [0,1]^n$ is an equilibrium vector of attack probabilities if and only if it is a fixed point of h.

The following lemma formalizes the endogenous strategic complementarity described above: if attacker j attacks more often, this makes attacker i less suspect, so attacker i also attacks more often.

Lemma 4 *For all distinct* $i, j \in I$*,* $h_i(p)$ *is non-decreasing in* p_j.

Proof. Note that the right-hand side of (7) is non-decreasing in p_j for all $j \neq i$. Hence, an increase in p_j shifts upward the right-hand side of (7) as a function p_i' and thus increases the intersection with p_i'. Formally, the result follows from, for example, Theorem 1 of Milgrom and Roberts [36]. ■

3 Equilibrium Properties and Comparative Statics

This section establishes equilibrium uniqueness and presents comparative statics with respect to F_i and G_i, the distributions of the attackers' and defender's aggressiveness.

3.1 Unique Equilibrium

We first show that there is always a unique equilibrium. To see the intuition, suppose there are two equilibria and attacker i's attack probability increases by the greatest fraction (among all attackers) in the second equilibrium relative to the first. Then attacker i is more suspect after every signal in the second equilibrium, so the defender will retaliate against attacker i more often. But then attacker i should attack less in the second equilibrium, not more.

Proposition 2 *There is a unique equilibrium.*

3.2 Complementary Aggressiveness

Lemma 4 shows that, if one attacker attacks with higher probability, this induces all attackers to attack with higher probability. Of course, attack probabilities are endogenous equilibrium objects.

To understand how such a change in behavior might result from changes in model primitives, we now turn to studying comparative statics with respect to the distributions F_i and G_i.

As we have already discussed, the parameter x_i represents attacker i's benefit from a successful attack relative to the cost of facing retaliation. Similarly, the parameter y represents the benefit of successful retaliation relative to the cost of retaliating against the wrong target. Thus, a change in the distributions F_i or G might result from an change in the distribution of benefits or the distribution of costs. In what follows, we say that attacker i (resp., the defender) *becomes more aggressive* if F_i (resp., G_i for all $i \in I$) increases in the first-order stochastic dominance sense.

3.2.1 Attackers' Aggressiveness

If any attacker becomes more aggressive, then in equilibrium *all* attackers attack with higher probability, and as a consequence the total probability of an attack increases. The intuition is as above: if one attacker attacks more often, the other attackers become less suspect and therefore face retaliation less often, which leads them to attack more often as well. This again expresses the multilateral nature of deterrence with imperfect attribution—increasing the aggressiveness of a single attacker reduces the extent to which all attackers are deterred.

Proposition 3 *Suppose attacker i becomes more aggressive, in that his type distribution changes from F_i to $\tilde{F}_i$, where $\tilde{F}_i(x_i) \leq F_i(x_i)$ for all x_i. Let (p, r) (resp., $(\tilde{p}, \tilde{r})$) denote the equilibrium attack and retaliation probabilities under F_i (resp., $\tilde{F}_i$). Then,*

1. *$p_i \leq \tilde{p}_i$ and $p_j \leq \tilde{p}_j$ for every $j \neq i$.*
2. *For every $j \neq i$, there exists $s \in S$ such that $r_j^s \geq \tilde{r}_j^s$.*

Proof.

1. Let h (resp., $\tilde{h}$) denote the endogenous best response function under F_i (resp., $\tilde{F}_i$). Note that $h_j(p') \leq \tilde{h}_j(p')$ for all $j \in I$ and $p' \in [0,1]^n$. As h and $\tilde{h}$ are monotone, it follows that $h^m((1,\ldots,1)) \leq \tilde{h}^m((1,\ldots,1))$ for all m, where h^m (resp., $\tilde{h}^m$) denotes the m^{th} iterate of the function h (resp., $\tilde{h}$). As h and $\tilde{h}$ are also continuous, and p and $\tilde{p}$ are the greatest fixed points of h and $\tilde{h}$, respectively, $\lim_{m\to\infty} h^m((1,\ldots,1)) = p$ and $\lim_{m\to\infty} \tilde{h}^m((1,\ldots,1)) = \tilde{p}$. Hence, $p \leq \tilde{p}$.

2. Immediate from part 1 of the proposition and (5).

■

The notion that deterrence with imperfect attribution is fundamentally multilateral because the attribution problem creates strategic complementarities among attackers is consistent with descriptions from the qualitative literature.

We have already mentioned Russia's false flag operation in North Korea during the Pyeongchang Winter Olympics. Similar issues arise with China. In 2009, the Information Warfare Monitor uncovered the GhostNet plot, an infiltration of government and commercial computer networks the world over, originating in China. The report indicates that there were "several possibilities for attribution." One likely possibility involved the Chinese government and military. But the report also notes that the evidence was consistent with alternative explanations, including "a random set of infected computers that just happens to include high profile targets of strategic significance to China," criminal networks, or patriotic hackers acting independently of the state. Finally, the report acknowledges, the attack could have been the work of "a state other than China, but operated physically within China... for strategic purposes... perhaps in an effort to deliberately mislead observers as to the true operator(s)." (See [26], pp.48–49.) Similar conclusions were reached half a decade earlier regarding the difficulty in attributing the Titan Rain attacks on American computer systems, which were again traced to internet addresses in China (Rogin [47]). In both cases, the United States government appears to have been highly reluctant to retaliate.

Given China's reputation for aggressiveness in cyberspace, why is the United States so reluctant to retaliate for cyberattacks attributed to China? It seems a key factor is precisely the strategic complementarity created by the attribution problem. In plain language, China's reputation makes it particularly tempting for other actors to hide behind America's suspicion of the Chinese. Singer and Friedman [51] describe the problem as follows:

> It is easy to assume that the [Chinese] government is behind most insidious activities launched by computers located within China. But, of course, this also means that bad actors elsewhere may be incentivized to target Chinese computers for capture and use in their activities, to misdirect suspicions. This very same logic, though, also enables Chinese actors to deny responsibility. (p. 74)

In Singer and Friedman's account, defenders receive signals indicating that China has engaged in a cyberattack. These signals are credible because China is highly suspect. That said, there is an attribution problem, because signals that point to Chinese computers may result from attacks

by foreign actors that have hacked their way into China, or from attacks by non-governmental domestic actors. Indeed, such "third-party" hacking is particularly attractive precisely because China is so suspect. The resulting prevalence of third-party hacking to some extent lets China deny responsibility. This reduces the willingness of defenders to retaliate, which in turn makes it more tempting for China (and everyone else) to attack.

3.2.2 Defender's Aggressiveness

As compared to an increase in an attacker's aggressiveness, an increase in the defender's aggressiveness has the opposite effect on deterrence: all attackers attack with lower equilibrium probability because retaliation is more likely, and consequently the total probability of an attack goes down. Thus, greater aggressiveness on the part of the defender strengthens deterrence. This fact will be a key to understanding our results on the optimal retaliatory policy (with commitment) in Section 5.

Proposition 4 *Suppose the defender becomes more aggressive, in that her type distribution changes from G to $\tilde{G}$, where $\tilde{G}_i(y_i) \leq G_i(y_i)$ for all $i \in I$ and all y_i. Let (p, r) (resp., $(\tilde{p}, \tilde{r})$) denote the equilibrium attack and retaliation probabilities under G (resp., $\tilde{G}$). Then*

1. *$p_i \geq \tilde{p}_i$ for every $i \in I$.*
2. *For every $i \in I$, there exists $s \in S$ such that $r_j^s \leq \tilde{r}_j^s$.*

Proof. Analogous to Proposition 3, noting that increasing G in the FOSD order shifts h down. ■

Proposition 4 says that increased aggressiveness on the part of a defender improves deterrence. For an implication of this result, consider the United States' recently released new nuclear strategy, the Nuclear Posture Review. This policy allows for the possibility of first-use of nuclear weapons in response to devastating but non-nuclear attacks, including cyberattacks (Sanger and Broad [48]). Such a policy may well increase the costs of mistaken retaliation relative to the benefits of "correct" retaliation, which worsens the distribution G. Proposition 4 thus suggests that this threat of harsher punishment could weaken deterrence.

Of course, there is a countervailing effect. The threat of harsher punishments simultaneously affects both the distributions G and F. While the new U.S. nuclear posture likely raises the costs of mistaken retaliation (worsening G), it presumably also raises the costs of being retaliated against. This latter effect corresponds to a worsening of the distribution F, which Proposition 3 indicates

improves deterrence. Thus, taken together, Propositions 3 and 4 indicate that the policy has two effects that pull in opposite directions, with an ambiguous net effect on deterrence.

3.3 Equilibrium Mutes Attacker Heterogeneity

If we put a little more structure on the model, we can make two further observations about attacker aggressiveness. First, not surprisingly, inherently more aggressive attackers attack with higher probability in equilibrium. Second, notwithstanding this fact, equilibrium mutes attacker heterogeneity—that is, inherently more aggressive attackers use a more demanding cutoff (i.e., a higher x_i^*). This follows because inherently more aggressive attackers are more suspect, and therefore face more retaliation.

This result implies another sense in which settings with imperfect attribution are fundamentally multilateral. Suppose attacker 1 is inherently much more aggressive than attacker 2. A naïve analysis would suggest that attacker 2 can be safely ignored. But this neglects attacker 2's great advantage of being able to hide behind attacker 1: if all attacks were assumed to come from attacker 1, attacker 2 could attack with impunity. Hence, equilibrium requires some parity of attack probabilities, even between attackers who are highly asymmetric ex ante.

To isolate the effect of heterogeneous aggressiveness, in this section we restrict attention to symmetric information structures. The information structure is *symmetric* if, for every permutation ρ on I, there exists a permutation ρ' on $S\backslash\{0\}$ such that $\pi_i^s = \pi_{\rho(i)}^{\rho'(s)}$ for all $i \in I$ and $s \in S\backslash\{0\}$.

Proposition 5 *Suppose the information structure is symmetric. Then, for every equilibrium and every $i, j \in I$, the following are equivalent:*

1. *i attacks with higher probability than j: $p_i > p_j$.*
2. *i has a higher threshold than j: $x_i^* > x_j^*$.*
3. *i is "inherently more aggressive" than j: $F_i(x_i^*) < F_j\left(x_j^*\right)$, and hence $F_i(x) < F_j(x)$ for all $x \in \left[x_j^*, x_i^*\right]$.*
4. *i is "more suspect" than j: for every permutation ρ on I mapping i to j and every corresponding permutation ρ' on $S\backslash\{0\}$, $\beta_i^s > \beta_j^{\rho'(s)}$ for all $s \in S\backslash\{0\}$.*

4 The Role of the Information Structure

As we have seen, attribution problems significantly complicate deterrence. As such, a natural intuition is that improving the defender's information—and thus her ability to attribute attacks—will improve deterrence. For instance, in a much discussed 2012 speech, then Secretary of Defense Leon Panetta said the following regarding US cybersecurity (Panetta [41]):

> Over the last two years, DoD has made significant investments in forensics to address this problem of attribution and we're seeing the returns on that investment. Potential aggressors should be aware that the United States has the capacity to locate them and to hold them accountable for their actions that may try to harm America.

In this section, we probe this intuition by studying how changes in the defender's information structure—the matrix $\pi = (\pi_i^s)_{i \in I \cup \{0\}, s \in S}$—affect deterrence. We will see that the conventional wisdom that better information improves deterrence is not always correct, but we also provide formal support for a somewhat weaker and more nuanced version of this claim.

We organize our results by noting three ways in which the defender might receive information that leads her to make a mistake, and then considering how reducing each kind of mistake affects deterrence. First, we consider a reduction in the probability of a false alarm—that is, reducing the probability that the defender gets a non-null signal when no attack has occurred. Second, we consider an improvement in detection—that is, increasing the probability that the defender gets a non-null signal when an attack occurs. Third, we consider an improvement in identification—that is, increasing the probability that the signal points unambiguously to attacker i when i attacks.

We show that a reduction in false alarms is always good for deterrence. Surprisingly, an improvement in either detection or identification alone can be either good or bad for deterrence—we give examples where improvements in either detection or identification lead all attackers to become more aggressive. Finally, we show that, if the defender can simultaneously improve both detection and identification, this unambiguously improves deterrence.

Throughout this section, we consider changes in the defender's information structure from π to $\tilde{\pi}$, and let variables with (resp., without) tildes denote equilibrium values under information structure π (resp., $\tilde{\pi}$).

4.1 False Alarms

If false alarms become less likely (i.e., π_0^s decreases for $s \neq 0$), then all attackers attack with lower probability in equilibrium. Each non-null signal now invites greater suspicion, and hence more retaliation. Also, the marginal impact of an attack on the probability of each non-null signal increases. Both forces increase the marginal impact of an attack on the probability of facing retaliation, and hence reduce the incentive to attack.

Proposition 6 *Suppose false alarms decrease: $\pi_0^s \geq \tilde{\pi}_0^s$ for all $s \neq 0$ and $\pi_0^0 \leq \tilde{\pi}_0^0$, while $\pi_i = \tilde{\pi}_i$ for all $i \in I$. Then $p_i \geq \tilde{p}_i$ for all $i \in I$. Also, $\tilde{r}_i^s \geq r_i^s$ for all $s \neq 0$ and for all $i \in I$.*

4.2 Detection

We say that the defender gets better at detecting attacks by attacker i if π_i^0 decreases (and hence π_i^s increases for some $s \neq 0$). Such an improvement in detection can actually hinder deterrence, making the defender worse-off. This is because an increase in π_i^s has two effects. It makes it easier to distinguish an attack by attacker i from no-attack. But it may also make it harder to distinguish an attack by attacker i from an attack by attacker j. Example 1 shows that the second effect can dominate, so an improvement in detection can make all attackers more aggressive.

Example 1

There are two attackers and three signals. Let $\gamma = \frac{2}{3}$, so with equal probability attacker 1 can attack, attacker 2 can attack, or no one can attack. The information structure $\pi = (\pi_i^s)$ is

$$\begin{array}{lll} \pi_0^0 = 1 & \pi_0^1 = 0 & \pi_0^2 = 0 \\ \pi_1^0 = \frac{1}{3} & \pi_1^1 = \frac{2}{3} & \pi_1^2 = 0 \\ \pi_2^0 = \frac{1}{3} & \pi_2^1 = \frac{1}{3} & \pi_2^2 = \frac{1}{3} \end{array}$$

Let $x_1 \in \left\{x_1^L = \frac{1}{2}, x_1^H = 1\right\}$, with $\Pr\left(x_1 = x_1^H\right) = \frac{4}{5}$.
Let $x_2 \in \left\{x_2^L = \frac{1}{4}, x_2^H = 1\right\}$, with $\Pr\left(x_2 = x_2^H\right) = \frac{1}{2}$.
Let $y = \frac{1}{4}$ with probability 1.[6]

Claim 1 *In the unique equilibrium with information structure π, attacker 1 attacks if and only if $x_1 = x_1^H$ and attacker 2 attacks if and only if $x_2 = x_2^H$. Thus, $p_1 = \frac{4}{5}$ and $p_2 = \frac{1}{2}$.*

[6] This type distribution is discrete. However, if we approximate with a continuous distribution, the equilibrium attack probabilities change continuously. The same remark applies to Examples 2 and 3 below.

Now suppose the information structure changes to

$$\begin{array}{lll} \tilde{\pi}_0^0 = 1 & \tilde{\pi}_0^1 = 0 & \tilde{\pi}_0^2 = 0 \\ \tilde{\pi}_1^0 = 0 & \tilde{\pi}_1^1 = \frac{2}{3} & \tilde{\pi}_1^2 = \frac{1}{3} \\ \tilde{\pi}_2^0 = \frac{1}{3} & \tilde{\pi}_2^1 = \frac{1}{3} & \tilde{\pi}_2^2 = \frac{1}{3} \end{array}$$

That is, when attacker 1 attacks, the attack is now always detected.

Claim 2 *In the unique equilibrium with information structure $\tilde{\pi}$, both attackers attack whenever they have the opportunity. Thus, $p_1 = p_2 = 1$.*

4.3 Identification

An improvement in identification can also weaken deterrence. To make this point, it suffices to consider a very restrictive notion of identification: an increase in $\pi_i^{s_i}$, where s_i is a signal that can arise *only* if attacker i attacks. Example 2 shows that an improvement in identification in this strong sense can increase all attack probabilities and make the defender worse-off. Furthermore, the change in information in the example is an improvement in the sense of Blackwell [10], and the example involves only a single attacker and is thus an example of a classical inspection game (Avenhaus, von Stengel, and Zamir [5]). Thus, the example shows that, if the defender's information improves in the sense of Blackwell [10], this can reduce the defender's payoff in the inspection game.[7]

To see the intuition, note that, if the benefit of attacking is high enough, even certain retaliation following the perfect signal may not be enough to deter an attack. The defender must then also be willing to retaliate following an imperfect signal. Moreover, the imperfect signal is less indicative of an attack when the perfect signal is more likely, as the probability that the imperfect signal is a false alarm increases when the perfect signal is more likely. Finally, for the defender to remain willing to retaliate following the imperfect signal when it is less indicative of an attack, the attacker must be attacking with higher probability. Thus, the attacker must attack with higher probability when the perfect signal is more likely.

Example 2

There is one attacker and three signals. Let $\gamma = 1$. The information structure is

$$\begin{array}{lll} \pi_0^0 = \frac{3}{4} & \pi_0^1 = \frac{1}{4} & \pi_0^2 = 0 \\ \pi_1^0 = \frac{1}{4} & \pi_1^1 = \frac{3}{4} & \pi_1^2 = 0 \end{array}.$$

[7] As far as we know, this is a novel observation.

Let $x = \frac{1}{3}$ and $y = \frac{1}{2}$.

Claim 3 *In the unique equilibrium with information structure π, the attacker attacks with probability $\frac{1}{4}$, and the defender retaliates with probability $\frac{2}{3}$ when $s = 1$.*

Suppose the information structure changes to

$$\begin{array}{lll} \tilde{\pi}_0^0 = \frac{3}{4} & \tilde{\pi}_0^1 = \frac{1}{4} & \tilde{\pi}_0^2 = 0 \\ \tilde{\pi}_1^0 = \frac{1}{4} & \tilde{\pi}_1^1 = \frac{1}{2} & \tilde{\pi}_1^2 = \frac{1}{4} \end{array}.$$

Now an attack is perfectly detected with probability $\frac{1}{4}$. Note that $\tilde{\pi}$ is Blackwell more informative than π: by simply conflating signals 1 and 2, the defender can recover π from $\tilde{\pi}$.

Claim 4 *In the unique equilibrium with information structure $\tilde{\pi}$, the attacker attacks with probability $\frac{1}{3}$, and the defender retaliates with probability $\frac{1}{3}$ when $s = 1$ and retaliates with probability 1 when $s = 2$.*

Thus, when the cost of being attacked K is sufficiently large, the defender is better off with less information.

4.4 Detection and Identification

While improving detection or identification alone can weaken deterrence, improving them jointly always improves deterrence. Specifically, we show that, conditional on an attack by attacker i, shifting weight from a signal where attacker i does not face retaliation to a signal where only attacker i faces retaliation makes all attackers less aggressive. The intuition is that such a change in the information structure reduces all three types of mistakes the defender can make—detection failures, false alarms, and misidentification—and thus rules out the countervailing effects highlighted in the above examples.

Proposition 7 *Suppose that, with information structure π, $r_i^s = 0 < r_i^{s'}$ and $r_j^{s'} = 0$ for all $j \neq i$. Suppose also that information structure $\tilde{\pi}$ shifts weight from π_i^s to $\pi_i^{s'}$: that is, $\pi_i^s > \tilde{\pi}_i^s$, $\pi_i^{s'} < \tilde{\pi}_i^{s'}$, and $\pi_j^{\hat{s}} = \tilde{\pi}_j^{\hat{s}}$ for all $(j, \hat{s}) \notin \{(i, s), (i, s')\}$. Then $p_j \geq \tilde{p}_j$ for all $j \in I$.*

A corollary is that shifting weight from non-detection to perfect identification for any attacker makes all attackers less aggressive. In this sense, simultaneously improving both detection and identification necessarily improves deterrence.

Corollary 1 *Suppose that* $G_i(0) < 1$ *and there is a signal* s_i *such that* $\pi_i^{s_i} > 0$, $\pi_0^{s_i} = 0$, *and* $\pi_j^{s_i} = 0$ *for all* $j \neq i$. *If* $\tilde{\pi}$ *shifts weight from* π_i^0 *to* $\pi_i^{s_i}$, *then* $p_j \geq \tilde{p}_j$ *for all* $j \in I$.

Proof. By Lemma 3, $r_i^0 = 0$. In addition, as signal s_i perfectly reveals that i attacked, $r_i^{s_i} = 1 - G_i(0) > 0$, and $r_j^{s_i} = 1 - G_j(1) = 0$ for all $j \neq i$. Hence, the result follows from Proposition 7. ■

It is widely believed that improvements in attribution will improve cyberdeterrence. Indeed, improved deterrence was clearly the goal in the above quote from Leon Panetta, which claimed a significant increase in the United States' capacity to attribute cyberattacks. Moreover, this view is ensconced in the Department of Defense's 2015 official Cyber Strategy [16], which argues:

> Attribution is a fundamental part of an effective cyber deterrence strategy as anonymity enables malicious cyber activity by state and non-state groups. On matters of intelligence, attribution, and warning, DoD and the intelligence community have invested significantly in all source collection, analysis, and dissemination capabilities, all of which reduce the anonymity of state and non-state actor activity in cyberspace. . . . [A]ttribution can play a significant role in dissuading cyber actors from conducting attacks in the first place. The Defense Department will continue to collaborate closely with the private sector and other agencies of the U.S. government to strengthen attribution. This work will be especially important for deterrence as activist groups, criminal organizations, and other actors acquire advanced cyber capabilities over time.

Our model shows that matters are not quite so simple. Reducing misidentification, on its own, creates trade-offs. On the one hand, increasing the frequency with which the defender receives a signal that decisively points to an attacker increases retaliation after that signal. This tends to improve deterrence. On the other hand, the very fact that an attack is more likely to lead to such a decisive signal makes ambiguous signals less indicative of an attack, reducing retaliation following such signals. This tends to make deterrence less effective. Whether improved identification strengthens or weakens deterrence, on net, depends on which effect dominates.

That said, Corollary 1 provides formal support for a more nuanced, and somewhat weaker, version of the conventional wisdom. In particular, deterrence is unambiguously strengthened if identification is improved via improved detection. That is, a technological innovation that both detects some attacks that were previously missed and provides a signal following such attacks that decisively points to the true attacker necessarily improves deterrence.

5 The Role of Commitment

Our last set of results concerns the role of commitment on the part of the defender: how does the defender optimally use her information to deter attacks when she can commit to ex post suboptimal retaliation after some signals?

This question is important because in reality the defender is likely to have some commitment power. For example, a branch of the military can announce a "strategic doctrine," with the understanding that commanders who violate the doctrine are penalized.[8] Indeed, there is serious discussion in the cyber domain (as there was in the nuclear domain) of pre-delegation, whereby military commanders are granted authority to engage in various types of defensive or retaliatory actions without seeking approval from civilian authorities (Feaver and Geers [19]).

We show that, as one would expect, with commitment the defender retaliates more often after some signals, and all attackers attack less often. Thus, generally speaking, the defender should try to commit herself to retaliate aggressively relative to her ex post inclination. But there are some subtleties: as we will see, there may also be some signals after which the defender retaliates *less* often with commitment than without. The intuition is that, since the attackers are less aggressive under commitment, some signals are now more likely to be false alarms, so retaliating after these signals becomes less efficient.

To analyze the commitment model, recall that the attackers' strategies depend only on the defender's retaliation probabilities $(r_i^s)_{i\in I, s\in S}$. Given a vector of retaliation probabilities, the optimal way for the defender to implement this vector is to retaliate against i after s if and only if $y > G^{-1}(1 - r_i^s)$. Hence, a commitment strategy can be summarized by a vector of cutoffs $(y_i^{s*})_{i\in I, s\in S}$ such that the defender retaliates against i after signal s if and only if $y_i > y_i^{s*}$.

What is the optimal vector of cutoffs, and how does it differ from the no-commitment equilib-

[8] For this reason, commitment by the defender is frequently studied as an alternative to no-commitment in the inspection game and related games. The commitment model is sometimes referred to as "inspector leadership" (Avenhaus, von Stengel, and Zamir [5])

rium? The defender's problem is

$$\max_{(y_i^s)_{i\in I, s\in S}}$$

$$\frac{\gamma}{n}\sum_i\left(1-F_i\left(\sum_s\left(\pi_i^s-\pi_0^s\right)\left(1-G_i\left(y_i^s\right)\right)\right)\right)\left[\begin{array}{c}-K\\ +\sum_s\pi_i^s\left[\begin{array}{c}\int_{y_i^s}^{\infty}ydG_i\left(y\right)\\ +\sum_{j\neq i}\int_{y_j^s}^{\infty}\left(y-1\right)dG_j\left(y\right)\end{array}\right]\\ -\sum_s\pi_0^s\sum_j\int_{y_j^s}^{\infty}\left(y-1\right)dG_j\left(y\right)\end{array}\right]$$

$$+\sum_s\pi_0^s\sum_j\int_{y_j^s}^{\infty}\left(y-1\right)dG_j\left(y\right)$$

This uses the fact that $x_i^*=\sum_s\left(\pi_i^s-\pi_0^s\right)\left(1-G\left(y_i^s\right)\right)$, so attacker i attacks with probability $1-F_i\left(\sum_s\left(\pi_i^s-\pi_0^s\right)\left(1-G\left(y_i^s\right)\right)\right)$. In the event attacker i attacks, the defender suffers a loss consisting of the sum of several terms. First, she suffers a direct loss of K. In addition, after signal s, she receives y_i if she retaliates against attacker i (i.e., if $y_i>y_i^s$) and receives y_j-1 if she erroneously retaliates against attacker j (i.e., if $y_j>y_j^s$). If instead no one attacks, then the defender receives y_j-1 if she erroneously retaliates against attacker j.

The first-order condition with respect to y_i^s is

$$f_i\left(x_i^*\right)\left(\pi_i^s-\pi_0^s\right)\left[\begin{array}{c}-K\\ +\sum_s\pi_i^s\left[\int_{y_i^s}^{\infty}ydG\left(y\right)+\sum_{j\neq i}\int_{y_j^s}^{\infty}\left(y-1\right)dG\left(y\right)\right]\\ -\sum_s\pi_0^s\sum_{j=1}^{n}\int_{y_j^s}^{\infty}\left(y-1\right)dG\left(y\right)\end{array}\right]$$

$$-\left(1-F_i\left(x_i^*\right)\right)\pi_i^s y_i^s$$

$$+\sum_{j\neq i}\left(1-F_j\left(x_j^*\right)\right)\pi_j^s\left(1-y_i^s\right)$$

$$+\left(\frac{n}{\gamma}-\sum_{j=1}^{n}\left(1-F_j\left(x_j^*\right)\right)\right)\pi_0^s\left(1-y_i^s\right)=0.$$

The first term is the (bad) effect that increasing y_i^s makes attacker i attack more. The second term is the (also bad) effect that increasing y_i^s makes attacks by i more costly, because the defender successfully retaliates less often. The third term is the (good) effect that increasing y_i^s makes attacks by each $j\neq i$ less costly, because the defender erroneously retaliates less often. The fourth term is the (good) effect that increasing y_i^s increases the defender's payoff when no one attacks,

again because the defender erroneously retaliates less often.

Denote the negative of the term in brackets (the cost of an attack by i) by $l_i(y^*)$. Then we can rearrange the FOC to

$$y_i^{s*} = \frac{n\pi_0^s + \gamma \sum_{j \neq i} \left(1 - F_j\left(x_j^*\right)\right)\left(\pi_j^s - \pi_0^s\right) - \gamma\left(1 - F_i\left(x_i^*\right)\right)\pi_0^s - \gamma f_i\left(x_i^*\right)\left(\pi_i^s - \pi_0^s\right) l_i\left(y^*\right)}{n\pi_0^s + \gamma \sum_{j \neq i} \left(1 - F_j\left(x_j^*\right)\right)\left(\pi_j^s - \pi_0^s\right) + \gamma\left(1 - F_i\left(x_i^*\right)\right)\left(\pi_i^s - \pi_0^s\right)}.$$

In contrast, in the no-commitment model, y_i^{s*} is given by the equation

$$y_i^{s*} = \frac{n\pi_0^s + \gamma \sum_{j \neq i} \left(1 - F_j\left(x_j^*\right)\right)\left(\pi_j^s - \pi_0^s\right) - \gamma\left(1 - F_i\left(x_i^*\right)\right)\pi_0^s}{n\pi_0^s + \gamma \sum_{j \neq i} \left(1 - F_j\left(x_j^*\right)\right)\left(\pi_j^s - \pi_0^s\right) + \gamma\left(1 - F_i\left(x_i^*\right)\right)\left(\pi_i^s - \pi_0^s\right)}.$$

Thus, the only difference in the equations for y^* as a function of x^* is that the commitment case has the additional term $-f_i\left(x_i^*\right)\left(\pi_i^s - \pi_0^s\right) l_i\left(y^*\right)$, reflecting the fact that increasing y_i^{s*} has the new cost of making attacks by i more likely. (In contrast, in the no-commitment case the attack decision has already been made at the time the defender chooses her retaliation strategy, so the defender trades off only the other three terms in the commitment FOC.) This difference reflects the additional deterrence benefit of committing to retaliate, and suggests that y_i^{s*} is always lower with commitment—that is, that commitment makes the defender more aggressive.

However, this intuition resulting from comparing the FOCs under commitment and no-commitment is incomplete: the x^*'s in the two equations are different, and we will see that it is possible for y_i^{s*} to be *higher* with commitment for some signals. Nonetheless, we can show that with commitment all attackers attack with lower probability and the defender retaliates with higher probability after at least some signals.

Proposition 8 *Let (p, r) be the no-commitment equilibrium and let $(\tilde{p}, \tilde{r})$ be the commitment equilibrium. Then $p_i \geq \tilde{p}_i$ for all $i \in I$, and for every $i \in I$ there exists $s \in S$ such that $r_i^s \leq \tilde{r}_i^s$.*

The second part of the proposition is immediate from the first: if every attacker is less aggressive under commitment, every attacker must face retaliation with a higher probability after at least one signal. The first part of the proposition follows from noting that the endogenous best response function is shifted up under commitment, due to the defender's additional deterrence benefit from committing to retaliate aggressively.

Proposition 8 shows that the defender benefits from committing to retaliate more aggressively after some signals. This is distinct from the search for credibility discussed in the nuclear deterrence

literature (Schelling [49], Snyder [55], Powell [44]). There, one assumes perfect attribution, and the key issue is how to make retaliation credible (i.e., make y_i positive). Here, we take y_i positive for granted, and show that the defender still has a problem of not being aggressive enough in equilibrium.

Although Proposition 8 indicates that the optimal deterrence policy is in some sense more aggressive than equilibrium retaliation, the next example shows that the optimal commitment strategy does not necessarily involve retaliating more aggressively after every signal. In the example, there are three signals: the null signal, an intermediate signal, and a highly informative signal. With commitment, the defender retaliates with very high probability after the highly informative signal. This deters attacks so successfully that the intermediate signal becomes very likely to be a false alarm. In contrast, without commitment, the equilibrium attack probability is higher, and the intermediate signal is more indicative of an attack. The defender therefore retaliates with higher probability following the intermediate signal without commitment.

Example 3

There is one attacker and three signals. Let $\gamma = \frac{1}{2}$. The information structure is

$$\begin{array}{lll} \pi_0^0 = \frac{1}{2} & \pi_0^1 = \frac{1}{3} & \pi_0^2 = \frac{1}{6} \\ \pi_1^0 = \frac{1}{6} & \pi_1^1 = \frac{1}{3} & \pi_1^2 = \frac{1}{2} \end{array}$$

Let $x \in \left\{x^L = \frac{1}{4}, x^H = 1\right\}$, with $\Pr\left(x = x^H\right) = \frac{1}{2}$.
Let $y \in \left\{y^L = \frac{1}{5}, y^H = \frac{3}{5}\right\}$, with $\Pr\left(y = y^H\right) = \frac{1}{2}$. Let $K = 1$.

Claim 5 *In the unique equilibrium without commitment, $p_1 = 1$, and the equilibrium retaliation probabilities $(r^s)_{s \in S}$ are given by*

$$r^0 = 0, r^1 = \frac{1}{2}, r^2 = \frac{1}{2}.$$

Claim 6 *In the unique equilibrium with commitment, $p_1 = \frac{1}{4}$, and the equilibrium retaliation probabilities $(r^s)_{s \in S}$ are given by*

$$r^0 = 0, r^1 = 0, r^2 = \frac{3}{4}.$$

Under some circumstances, we can say more about how equilibrium retaliation differs with and without commitment. Say that signals s and s' are *comparable* if there exists $i^* \in I$ such that $\pi_i^s = \pi_0^s$ and $\pi_i^{s'} = \pi_0^{s'}$ for all $i \neq i^*$. If s and s' are comparable, say that s is *more informative*

than s' if

$$\frac{\pi^s_{i^*}}{\pi^s_0} \geq \frac{\pi^{s'}_{i^*}}{\pi^{s'}_0}.$$

That is, s is more informative than s' if, compared to s', s is relatively more likely to result from an attack by i^* than from no attack (or from an attack by any $i \neq i^*$).

The next Proposition shows that, if s is more informative than s' and the defender is more aggressive after s' with commitment than without, then the defender is also more aggressive after s with commitment than without. (Conversely, if the defender is less aggressive after s with commitment, then the defender is also less aggressive after s' with commitment.) That is, commitment favors more aggressive retaliation following more informative signals. The intuition is that the ability to commit tilts the defender towards relying on the most informative signals to deter attacks, and any offsetting effects resulting from the increased probability of false alarms are confined to less informative signals.

Note that the following result concerns the defender's aggressiveness toward any attacker, not only the attacker i^* used to compare s and s'.

Proposition 9 *Let (x, y) be the no-commitment equilibrium and let $(\tilde{x}, \tilde{y})$ be the commitment equilibrium. Fix an attacker $i \in I$ and signals $s, s' \in S$ such that s and s' are comparable, s is more informative than s', and $\min\left\{y_i^s, y_i^{s'}, y_i^{s'}, \tilde{y}_i^{s'}\right\} > 0$. If $\tilde{y}_i^{s'} \leq y_i^{s'}$, then $\tilde{y}_i^s \leq y_i^s$; and if $\tilde{y}_i^s \geq y_i^s$, then $\tilde{y}_i^{s'} \geq y_i^{s'}$.*

To summarize, Proposition 8 is in broad agreement with recent arguments calling for more aggressive cyberdeterrence (e.g., Hennessy [25]). However, Example 3 shows that improving cyberdeterrence is more subtle than simply increasing aggressiveness across the board. While the optimal policy has the defender retaliating more aggressively some of the time, it does not necessarily involve increased retaliation after all signal realizations that point to an attack. This is because some signal realizations may do a relatively poor job of distinguishing among potential attackers. Increased retaliation following such signal realizations may do little to influence the marginal incentives of attackers while leading to significant costs of triggering erroneous retaliation. Moreover, as retaliatory aggressiveness ramps up and deters ever more attacks, this risk becomes greater, as a larger share of perceived attacks will turn out to be false alarms.

Slightly more broadly, our analysis can usefully be contrasted with the suggestion by Clarke and Knake [15] that cybersecurity would be enhanced by a policy that holds governments responsible for any cyberattack originating from their territory, whether state sanctioned or otherwise. Such a

policy is one way of increasing retaliatory aggressiveness across the board, since it holds governments accountable for an extremely wide range of attacks. The problem with such a policy, from our perspective, is that it could lead to increased retaliation following relatively uninformative signals (e.g., the simple fact that an attack emanates from servers in Abu Dhabi or China). Increased aggressiveness following such uninformative signals heightens the risk of retaliation against an innocent actor. It also increases incentives for deliberate sabotage or provocation, a topic we turn to in the next section.

Finally, we remark that the strategic complementarity among attackers that drove our results in the no-commitment model partially breaks down under commitment. In particular, it is no longer true that an exogenous increase in attacker i's aggressiveness always makes all attackers more aggressive in equilibrium. The reason is that the complementarity effect from the no-commitment model may be offset by a new effect coming from the deterrence term $f_i\left(x_i^*\right)\left(\pi_i^s-\pi_0^s\right)l_i\left(y^*\right)$ in the defender's FOC. Intuitively, if attacker i starts attacking more often, this typically leads the defender to start retaliating more against attacker i (y_i^* decreases) and less against other defenders (y_j^* increases for $j\neq i$). This strategic response by the defender has the effect of increasing $l_j\left(y^*\right)$ for all $j\neq i$: since the defender retaliates more against i and less against j, an attack by j becomes more costly for the defender, as it is more likely to be followed by erroneous retaliation against i and less likely to be followed by correct retaliation against j. This increase in $l_j\left(y^*\right)$ then makes it more valuable for the defender to deter attacks by j (as reflected in the $f_j\left(x_j^*\right)\left(\pi_j^s-\pi_0^s\right)l_j\left(y^*\right)$ term), which leads to an offsetting decrease in y_j^*.

6 An Extension: Provocateurs

Underlying the strategic complementarity in our model is the fact that, when any one attacker strikes the defender, this generates "suspicious" signals and hence increases the probability that the defender retaliates against every attacker. This feature of the model naturally raises questions about the ability of provocateurs to foment conflict. For instance, terrorist organizations often use violence to try to spoil peace talks (Stedman [56], Kydd and Walter [32]). The idea is that some faction opposed to a negotiated peace engages in attacks that might be misattributed to a competing faction that is engaged in the negotiations, leading to the breakdown of the peace process.[9] A related phenomenon is false-flag operations, where one attacker poses as another to

[9] Examples might include the wave of violence by Palestinian factions during the negotiations over the Oslo Accords, or the Real IRA's bombing of Omagh shortly after the signing of the Good Friday Agreement.

evade responsibility or to provoke a conflict between another attacker and the defender (for a discussion in the cyber context, see Bartholomew and Guerrero-Saade [8]). In our setting, one might worry that one attacker (the *provocateur*) might engage in aggressive behavior in order to increase the likelihood of retaliation by the defender against another attacker whom the provocateur considers an adversary.

To address this possibility, we generalize the model by supposing that there is one attacker i, the provocateur, who gets an additional benefit $b_i^j \geq 0$ when the defender retaliates against attacker $j \neq i$. In this case, equation (2) becomes

$$x_i^* = \sum_s \left(\pi_i^s - \pi_0^s\right) \left[r_i^s - \sum_{j \neq i} r_j^s b_i^j\right].$$

As in the baseline model, the defender retaliates against attacker i after signal s if $y > 1 - \beta_i^s$, and the defender never retaliates after signal 0. Hence,

$$x_i^* = \sum_{s \neq 0} \left(\pi_i^s - \pi_0^s\right) \left[1 - G\left(1 - \beta_i^s\right) - \sum_{j \neq i} \left(1 - G\left(1 - \beta_j^s\right)\right) b_i^j\right],$$

or equivalently

$$x_i^* = \sum_{s \neq 0} \left(\pi_i^s - \pi_0^s\right) \left[\begin{array}{c} 1 - G\left(1 - \frac{\gamma\left(1 - F_i\left(x_i^*\right)\right)\pi_i^s}{n\pi_0^s + \gamma \sum_{j'} \left(1 - F_{j'}\left(x_{j'}^*\right)\right)\left(\pi_{j'}^s - \pi_0^s\right)}\right) \\ - \sum_{j \neq i} \left(1 - G\left(1 - \frac{\gamma\left(1 - F_j\left(x_j^*\right)\right)\pi_i^s}{n\pi_0^s + \gamma \sum_{j'} \left(1 - F_{j'}\left(x_{j'}^*\right)\right)\left(\pi_{j'}^s - \pi_0^s\right)}\right)\right) b_i^j \end{array}\right]$$

In the special case of only two attackers, this simplifies to

$$x_i^* = \sum_{s \neq 0} \left(\pi_i^s - \pi_0^s\right) \left[\begin{array}{c} 1 - G\left(1 - \frac{\gamma\left(1 - F_i\left(x_i^*\right)\right)\pi_i^s}{n\pi_0^s + \gamma \sum_{j \in \{i,-i\}} \left(1 - F_j\left(x_j^*\right)\right)\left(\pi_{j'}^s - \pi_0^s\right)}\right) \\ - \left(1 - G\left(1 - \frac{\gamma\left(1 - F_j\left(x_j^*\right)\right)\pi_i^s}{n\pi_0^s + \gamma \sum_{j \in \{i,-i\}} \left(1 - F_j\left(x_j^*\right)\right)\left(\pi_j^s - \pi_0^s\right)}\right)\right) b_i^{-i} \end{array}\right]$$

Note that the right-hand side of this equation is decreasing in x_i^* and increasing in x_{-i}^*. The intuition is that an increase in x_i^* decreases β_i^s and increases β_{-i}^s for all s, while an increase in x_{-i}^* increases β_i^s and decreases β_{-i}^s for all s. As the provocateur dislikes retaliation against herself and likes retaliation against attacker $-i$, this means that the provocateur's benefit of attacking is decreasing in her own equilibrium attack probability and increasing in the other attacker's. Hence,

strategic complementarity is preserved, and all the results from the baseline model go through unchanged. We also get some new results: for example, an increase in b_i^{-i} will make both players more aggressive in equilibrium. That is, the more motivated is the provocateur, the more conflict there will be. It is also clear that the same conclusions apply if both attackers $i \in \{1, 2\}$ benefit from provoking a retaliatory attack against the other. We summarize with a proposition

Proposition 10 *With two attackers, all results from the baseline model remain valid if each attacker $i \in \{1, 2\}$ receives an additional "provocation benefit" $b_i^{-i} \geq 0$ when the defender retaliates against the other attacker. In addition, if attacker i's provocation benefit increases from b_i^{-i} to $\tilde{b}_i^{-i} \geq b_i^{-i}$, then both attackers attack with greater probability in equilibrium: $p_i \leq \tilde{p}_i$ and $p_{-i} \leq \tilde{p}_{-i}$.*

Interestingly, the analysis is more complicated if there are more than two attackers. In particular, the strategic complementarity from our baseline model need no longer hold. To see the problem, suppose there are three attackers: the provocateur (1), and two regular attackers (2 and 3). An increase in attacker 2's attack probability decreases β_3^s and hence decreases the retaliation probability against attacker 3. If the provocateur is strongly motivated by the prospect of encouraging retaliation against attacker 3, an increase in attacker 2's attack probability can thus *discourage* attacks by the provocateur. Thus, while the intuition that highly motivated provocateurs are destabilizing finds support in the two-attacker model, this may not be the case with more than two attackers.[10]

7 Conclusion

Motivated by recent developments in cyberwarfare, we develop a model of deterrence with imperfect attribution. There are several main findings.

First, a form of endogenous strategic complementarity arises among the different potential attackers. Increased aggressiveness on the part of one attacker makes all other attackers more aggressive, due to the possibility of "hiding their attacks" behind the first attacker.

Second, improving the defender's information has subtle—and sometimes counterintuitive—effects on the efficacy of deterrence. Improving either the defender's ability to detect attacks or her

[10] At first glance, this finding appears reminiscent of the fact that Cournot oligopoly has strategic complements with two firms but not three. But the logic is quite different. In Cournot, the issue is that with two firms we can turn substitutes into complements by reversing the order on one firm's strategies. Here, the issue is that increased aggressiveness by attacker 2 encourages attacks by the provocateur through its effect on the defender's beliefs about attacker 2, but can discourage attacks by the provocateur through its effect on the defender's beliefs about attacker 3.

ability to identify attackers can make deterrence less effective. However, simultaneously improving both detection and identification—in that some attacks that previously went undetected are now both detected and correctly attributed—always improves deterrence.

Third, deterrence is always more effective if the defender can commit in advance to a retaliatory strategy. However, the defender should not necessarily commit to retaliate more after every possible signal, and should instead base retaliation on only the most informative signals.

We have considered a very simple and stylized model in order to clarify some basic strategic issues that arise under imperfect attribution of attacks. There are many possible extensions and elaborations. For example, we have studied an asymmetric model where the roles of attacker and defender are distinct. More realistically, players might both carry out attacks and suffer from them. In such a model, player A may attack player B, but player B might be reluctant to retaliate, fearing a future counterattack or an escalation into conventional war. Or player A may be attacked by player B but attribute the attack to player C, and hence retaliate against player C. But this in turn triggers retaliation by player C, and attacks and retaliation may spread through the international system. How can peace be maintained in such a dynamic model with risks of multilateral misattribution and escalation?

A second possible extension would introduce different types of attacks and retaliation, perhaps along with uncertainty as to each actor's capability. In such a model, would deterrence be reserved for the largest attacks, even at the cost of allowing constant low-level intrusions? Would the ability to signal cyber-capability lead to coordination on a peaceful equilibrium, or to perverse incentives leading to conflict? We hope the current paper may inspire further research on these important and timely questions.

Appendix: Omitted Proofs

Proof of Lemma 1. When attacker i's type is x_i, his expected payoff when he attacks is

$$x_i - \sum_s \pi_i^s r_i^s,$$

and his expected payoff when he has the opportunity to attack but does not attack is

$$-\sum_s \pi_0^s r_i^s.$$

Therefore, i attacks when he has the opportunity if $x_i > \sum_s (\pi_i^s - \pi_0^s) r_i^s$, and he does not attack if $x_i < \sum_s (\pi_i^s - \pi_0^s) r_i^s$. ■

Proof of Lemma 2. When the defender's type is y, her (additional) payoff from retaliating against attacker i after signal s is $y_i - 1 + \beta_i^s(p)$. Therefore, she retaliates if $y_i > 1 - \beta_i^s(p)$, and does not retaliate if $y_i < 1 - \beta_i^s(p)$. ■

Proof of Lemma 3. Note that

$$\begin{aligned} y_i^{0*} &= 1 - \beta_i^0(p) \\ &= 1 - \frac{\gamma p_i \pi_i^0}{n\pi_0^0 - \gamma \sum_j p_j \left(\pi_0^0 - \pi_j^0\right)} \\ &\geq 1 - \frac{\gamma \pi_i^0}{n\pi_0^0 - \gamma (n-1) \pi_0^0 - \gamma \left(\pi_0^0 - \pi_i^0\right)} \\ &= \frac{n(1-\gamma)\pi_0^0}{n(1-\gamma)\pi_0^0 + \gamma \pi_i^0}, \end{aligned}$$

where the inequality follows because $\pi_0^0 \geq \pi_j^0$ for all j. The lemma now follows by (1). ■

Proof of Proposition 1. Equation (5) follows from combining (2), (4), $x_i^* = F_i^{-1}(1 - p_i)$, and $y_i^{s*} = G_i^{-1}(1 - r_i^s)$, and recalling that $r_i^0 = 0$. Equation (6) then follows from (3). The equation for r_i^s follows from combining (4) and $y_i^{s*} = G_i^{-1}(1 - r_i^s)$. ■

Proof of Proposition 2. We show that h has a unique fixed point.

By Lemma 4 (and the fact that $h_i(p)$ does not depend on p_i), h is a monotone function on $[0,1]^n$. Hence, by Tarski's fixed point theorem, h has a greatest fixed point: that is, there is a fixed point p^* such that, for every fixed point p^{**}, $p_i^* \geq p_i^{**}$ for all $i \in I$.

Now let p^* be the greatest equilibrium, and let p^{**} be an arbitrary equilibrium. We show that

$p^* = p^{**}$.

Let $i = \text{argmax}_{j \in I} \frac{p_j^*}{p_j^{**}}$. As p^* is the greatest equilibrium, we have $\frac{p_i^*}{p_i^{**}} \geq 1$. Therefore, for every $s \in S$,

$$\begin{aligned}
\beta_i^s(p^*) &= \frac{\gamma p_i^* \pi_i^s}{n\pi_0^s + \gamma \sum_j p_j^* \left(\pi_j^s - \pi_0^s\right)} \\
&= \frac{\frac{p_i^{**}}{p_i^*} \gamma p_i^* \pi_i^s}{\frac{p_i^{**}}{p_i^*} n\pi_0^s + \frac{p_i^{**}}{p_i^*} \gamma \sum_j p_j^* \left(\pi_j^s - \pi_0^s\right)} \\
&\geq \frac{\gamma p_i^{**} \pi_i^s}{\frac{p_i^{**}}{p_i^*} n\pi_0^s + \gamma \sum_j p_j^{**} \left(\pi_j^s - \pi_0^s\right)} \\
&\geq \frac{\gamma p_i^{**} \pi_i^s}{n\pi_0^s + \gamma \sum_j p_j^{**} \left(\pi_j^s - \pi_0^s\right)} = \beta_i^s(p^{**}),
\end{aligned}$$

where the first inequality holds because $\frac{p_i^{**}}{p_i^*} \leq \frac{p_j^{**}}{p_j^*}$ for all $j \in I$ and the second inequality holds because $\frac{p_i^{**}}{p_i^*} \leq 1$. Notice this implies

$$\begin{aligned}
p_i^* &= 1 - F_i\left(\sum_{s \neq 0} (\pi_i^s - \pi_0^s)\left(1 - G_i\left(1 - \beta_i^s(p^*)\right)\right)\right) \\
&\leq 1 - F_i\left(\sum_{s \neq 0} (\pi_i^s - \pi_0^s)\left(1 - G_i\left(1 - \beta_i^s(p^{**})\right)\right)\right) = p_i^{**}.
\end{aligned}$$

As p^* is the greatest equilibrium, this implies $p_i^* = p_i^{**}$. Since $i = \text{argmax}_{j \in I} \frac{p_j^*}{p_j^{**}}$, this implies $p_j^* \leq p_j^{**}$ for all $j \in I$. Hence, as p^* is the greatest equilibrium, $p^* = p^{**}$. ■

Proof of Proposition 5. Fix a permutation ρ on I mapping i to j and a corresponding permutation ρ' on $S \setminus \{0\}$. Then

$$\begin{aligned}
x_i^* &= \sum_{s \neq 0} (\pi_i^s - \pi_0^s)\left(1 - G(1 - \beta_i^s)\right) \\
&= \sum_{s \neq 0} (\pi_i^s - \pi_0^s)\left(1 - G\left(1 - \frac{\gamma\left(1 - F_i\left(x_i^*\right)\right)\pi_i^s}{n\pi_0^s + \gamma \sum_k \left(1 - F_k\left(x_k^*\right)\right)\left(\pi_k^s - \pi_0^s\right)}\right)\right)
\end{aligned}$$

and

$$
\begin{aligned}
x_j^* &= \sum_{s\neq 0}\left(\pi_j^{\rho'(s)}-\pi_0^{\rho'(s)}\right)\left(1-G(1-\beta_j^{\rho'(s)})\right)\\
&= \sum_{s\neq 0}\left(\pi_j^{\rho'(s)}-\pi_0^{\rho'(s)}\right)\left(1-G\left(1-\frac{\gamma\left(1-F_j\left(x_j^*\right)\right)\pi_j^{\rho'(s)}}{n\pi_0^{\rho'(s)}+\gamma\sum_k\left(1-F_k\left(x_k^*\right)\right)\left(\pi_k^{\rho'(s)}-\pi_0^{\rho'(s)}\right)}\right)\right)\\
&= \sum_{s\neq 0}\left(\pi_i^s-\pi_0^s\right)\left(1-G\left(1-\frac{\gamma\left(1-F_j\left(x_j^*\right)\right)\pi_i^s}{n\pi_0^s+\gamma\sum_k\left(1-F_k\left(x_k^*\right)\right)\left(\pi_k^s-\pi_0^s\right)}\right)\right).
\end{aligned}
$$

Hence,

$$
x_i^*>x_j^* \iff F_i\left(x_i^*\right)<F_j\left(x_j^*\right) \iff p_i>p_j \iff \beta_i^s>\beta_j^{\rho'(s)} \text{ for all } s\in S\backslash\{0\}.
$$

■

Proof of Proposition 6. Note that a decrease in π_0^s for all $s\neq 0$ shifts h down. There are two effects that go the same way: β_i^s increases for all $s\neq 0$ and $i\in I$, so the probability of retaliation given signal s increases (i.e., $\tilde{r}_i^s\geq r_i^s$); and $\pi_i^s-\pi_0^s$, the increase in the probability of each retaliation-inducing signal due to an attack, also increases. Hence, as the proof of Proposition 3, $p_i\geq\tilde{p}_i$ for all i. ■

Proof of Claim 1. It suffices to check that these strategies form an equilibrium. Given the conditional attack probabilities and the information structure, the defender's posterior beliefs (β_i^s) are given by

$$
\begin{array}{lll}
\beta_0^0=\frac{51}{64} & \beta_1^0=\frac{8}{64} & \beta_2^0=\frac{5}{64}\\
\beta_0^1=0 & \beta_1^1=\frac{16}{21} & \beta_2^1=\frac{5}{21}\\
\beta_0^2=0 & \beta_1^2=0 & \beta_2^2=1
\end{array}.
$$

Since $y=\frac{1}{4}$, the defender retaliates against attacker i after signal s iff $\beta_i^s>\frac{3}{4}$. Thus, the defender retaliates against attacker 1 iff $s=1$, and the defender retaliates against attacker 2 iff $s=2$. Therefore, $x_1^*=\frac{2}{3}$ and $x_2^*=\frac{1}{3}$. It follows that attacker 1 attacks iff $x_1=x_1^H$ and attacker 2 attacks iff $x_2=x_2^H$. So this is an equilibrium. ■

Proof of Claim 2. Again, we check that these strategies form an equilibrium. Again, combining the conditional attack probabilities and the information structure, the defender's posterior beliefs

are given by

$$\begin{array}{lll} \beta_0^0 = \frac{3}{4} & \beta_1^0 = 0 & \beta_2^0 = \frac{1}{4} \\ \beta_0^1 = 0 & \beta_1^1 = \frac{2}{3} & \beta_2^1 = \frac{1}{3} \\ \beta_0^2 = 0 & \beta_1^2 = \frac{1}{2} & \beta_2^2 = \frac{1}{2} \end{array} .$$

Note that $\beta_i^s < \frac{3}{4}$ for all $i \in \{1,2\}$ and all s. Hence, the defender never retaliates. This implies that $x_1^* = x_2^* = 0$, so both attackers always attack. ■

Proof of Claim 3. It is clear that the equilibrium must be in mixed strategies. Let p be the probability the attacker attacks. The defender's posterior belief when $s = 1$ is $\beta_1^1 = \frac{3p}{1+2p}$. For the defender to be indifferent, this must equal $\frac{1}{2}$. This gives $p = \frac{1}{4}$.

For the attacker to be indifferent, the retaliation probability when $s = 1$ must solve $\left(\frac{3}{4} - \frac{1}{4}\right) r_1 = \frac{1}{3}$, or $r = \frac{2}{3}$. ■

Proof of Claim 4. Clearly, the defender retaliates with probability 1 when $s = 2$. As $x > \tilde{\pi}_1^2$, this is not enough to deter an attack, so the defender must also retaliate with positive probability when $s = 1$. The defender's posterior belief when $s = 1$ is now $\tilde{\beta}_1^1 = \frac{2p}{1+p}$. For the defender to be indifferent, this must equal $\frac{1}{2}$. This gives $p = \frac{1}{3}$.

For the attacker to be indifferent, the retaliation probability when $s = 1$ must solve $\left(\frac{1}{2} - \frac{1}{4}\right) r_1 + \left(\frac{1}{4}\right)(1) = \frac{1}{3}$, or $r = \frac{1}{3}$. ■

Proof of Claim 5. We check that these strategies form an equilibrium. Note that the defender's posterior beliefs (β_i^s) are given by

$$\begin{array}{ll} \beta_0^0 = \frac{3}{4} & \beta_1^0 = \frac{1}{4} \\ \beta_0^1 = \frac{1}{2} & \beta_1^1 = \frac{1}{2} \\ \beta_0^2 = \frac{1}{4} & \beta_1^2 = \frac{3}{4} \end{array}$$

Recall that the defender retaliates iff $\beta_1^s > 1 - y$. Hence, when $y = y^L$ the defender never retaliates, and when $y = y^H$ the defender retaliates when $s \in \{1,2\}$. Therefore,

$$x^* = \left(\pi_1^1 - \pi_0^1\right) r_1 + \left(\pi_1^2 - \pi_0^2\right) r_2 = (0)\,\frac{1}{2} + \left(\frac{1}{2} - \frac{1}{6}\right)\frac{1}{2} = \frac{1}{6}.$$

Hence, the attacker attacks whenever he has an opportunity. ■

Proof of Claim 6. First, note that these retaliation probabilities deter attacks when $x = x^L$, and yield a higher defender payoff than any strategy that does not deter attacks when $x = x^L$. So the commitment solution will deter attacks when $x = x^L$. Note also that it is impossible to deter attacks when $x = x^H$. So the commitment solution must have $p_1 = \frac{1}{4}$.

When $p_1 = \frac{1}{4}$, the defender's posterior beliefs (β_i^s) are given by

$$\begin{aligned} \beta_0^0 &= \tfrac{9}{10} \quad \beta_1^0 = \tfrac{1}{10} \\ \beta_0^1 &= \tfrac{3}{4} \quad \beta_1^1 = \tfrac{1}{4} \\ \beta_0^2 &= \tfrac{1}{2} \quad \beta_1^2 = \tfrac{1}{2} \end{aligned}$$

With these beliefs, ignoring the effect on deterrence, it is not optimal for the defender to retaliate when $s \in \{0,1\}$. Furthermore, retaliating after $s \in \{0,1\}$ weakly increases the attacker's incentive to attack. So the commitment solution involves retaliation only when $s = 2$.

Finally, when $s = 2$, it is profitable for the defender to retaliate when $y = y^H$ and unprofitable to retaliate when $y = y^L$. So the solution involves retaliation with probability 1 when $y = y^H$, and retaliation with the smallest probability required to deter attacks by the $x = x^L$ type attacker when $y = y^L$. This solution is given by retaliating with probability $\frac{1}{2}$ when $y = y^L$. ■

Proof of Proposition 7. Suppose toward a contradiction that there exists j (possibly equal to i) such that $p_j < \tilde{p}_j$. Then $\max_{j \in I} \frac{\tilde{p}_j}{p_j} > 1$. We first show that $i \in \operatorname{argmax}_{j \in I} \frac{\tilde{p}_j}{p_j}$.

To see this, suppose otherwise, and fix $j^* \in \operatorname{argmax}_{j \in I} \frac{\tilde{p}_j}{p_j}$. As $p_{j^*} < \tilde{p}_{j^*}$, we have $r_{j^*}^{\hat{s}} > \tilde{r}_{j^*}^{\hat{s}}$ for some $\hat{s} \in S$. (Otherwise, j^* would not be more aggressive under $\tilde{\pi}$ than π.) As $r_{j^*}^{s'} = 0$, we have $r_{j^*}^{\hat{s}} > \tilde{r}_{j^*}^{\hat{s}}$ for some $\hat{s} \neq s'$. But then

$$\begin{aligned} \tilde{\beta}_{j^*}^{\hat{s}}(\tilde{p}) &= \frac{\gamma \tilde{p}_{j^*} \tilde{\pi}_{j^*}^{\hat{s}}}{n\tilde{\pi}_0^{\hat{s}} + \gamma \sum_j \tilde{p}_j \left(\tilde{\pi}_j^{\hat{s}} - \tilde{\pi}_0^{\hat{s}}\right)} \\ &= \frac{\frac{p_{j^*}}{\tilde{p}_{j^*}} \gamma \tilde{p}_{j^*} \tilde{\pi}_{j^*}^{\hat{s}}}{\frac{p_{j^*}}{\tilde{p}_{j^*}} n\tilde{\pi}_0^{\hat{s}} + \frac{p_{j^*}}{\tilde{p}_{j^*}} \gamma \sum_j \tilde{p}_j \left(\tilde{\pi}_j^{\hat{s}} - \tilde{\pi}_0^{\hat{s}}\right)} \\ &\geq \frac{\gamma p_{j^*} \tilde{\pi}_{j^*}^{\hat{s}}}{n\tilde{\pi}_0^{\hat{s}} + \gamma \sum_j p_j \left(\tilde{\pi}_j^{\hat{s}} - \tilde{\pi}_0^{\hat{s}}\right)} \\ &\geq \frac{\gamma p_{j^*} \pi_{j^*}^{\hat{s}}}{n\pi_0^{\hat{s}} + \gamma \sum_j p_j \left(\pi_j^{\hat{s}} - \pi_0^{\hat{s}}\right)} = \beta_{j^*}^{\hat{s}}(p), \end{aligned}$$

where the first inequality follows because $\frac{p_{j^*}}{\tilde{p}_{j^*}} \leq \frac{p_j}{\tilde{p}_j}$ for all $j \in I$ and $\frac{p_{j^*}}{\tilde{p}_{j^*}} < 1$, and the second inequality follows because $\tilde{\pi}_{j^*}^{\hat{s}} = \pi_{j^*}^{\hat{s}}$, $\tilde{\pi}_0^{\hat{s}} = \pi_0^{\hat{s}}$, and $\tilde{\pi}_j^{\hat{s}} \leq \pi_j^{\hat{s}}$ for all $j \neq j^*$ (as $\hat{s} \neq s'$). This implies $\tilde{r}_{j^*}^{\hat{s}} \geq r_{j^*}^{\hat{s}}$, a contradiction. Hence, $i \in \operatorname{argmax}_{j \in I} \frac{\tilde{p}_j}{p_j}$.

To complete the proof, note that $r_i^{\hat{s}} > \tilde{r}_i^{\hat{s}}$ for some $\hat{s} \in S$. (Otherwise, i would not be more

aggressive under $\tilde{\pi}$ than π.) As $r_i^s = 0$, we have $r_i^{\hat{s}} > \tilde{r}_i^{\hat{s}}$ for some $\hat{s} \neq s$. But then

$$\begin{aligned}
\tilde{\beta}_i^{\hat{s}}(\tilde{p}) &= \frac{\gamma \tilde{p}_i \tilde{\pi}_i^{\hat{s}}}{n\tilde{\pi}_0^{\hat{s}} + \gamma \sum_j \tilde{p}_j \left(\tilde{\pi}_j^{\hat{s}} - \tilde{\pi}_0^{\hat{s}}\right)} \\
&= \frac{\frac{p_i}{\tilde{p}_i} \tilde{p}_i \tilde{\pi}_i^{\hat{s}}}{\frac{p_i}{\tilde{p}_i} n\tilde{\pi}_0^{\hat{s}} + \frac{p_i}{\tilde{p}_i} \gamma \sum_j \tilde{p}_j \left(\tilde{\pi}_j^{\hat{s}} - \tilde{\pi}_0^{\hat{s}}\right)} \\
&\geq \frac{\gamma p_i \tilde{\pi}_i^{\hat{s}}}{n\tilde{\pi}_0^{\hat{s}} + \gamma \sum_j p_j \left(\tilde{\pi}_j^{\hat{s}} - \tilde{\pi}_0^{\hat{s}}\right)} \\
&\geq \frac{\gamma p_i \pi_i^{\hat{s}}}{n\pi_0^{\hat{s}} + \gamma \sum_j p_j \left(\pi_j^{\hat{s}} - \pi_0^{\hat{s}}\right)} = \beta_i^{\hat{s}}(p),
\end{aligned}$$

where the first inequality follows because $\frac{p_i}{\tilde{p}_i} \leq \frac{p_j}{\tilde{p}_j}$ for all $j \in I$ and $\frac{p_i}{\tilde{p}_i} < 1$, and the second inequality follows because $\tilde{\pi}_i^{\hat{s}} \geq \pi_i^{\hat{s}}$ (as $\hat{s} \neq s$), $\tilde{\pi}_0^{\hat{s}} = \pi_0^{\hat{s}}$, and $\tilde{\pi}_j^{\hat{s}} = \pi_j^{\hat{s}}$ for all $j \neq i$. This implies $\tilde{r}_i^{\hat{s}} \geq r_i^{\hat{s}}$, a contradiction. ■

Proof of Proposition 8. By the defender's FOC with commitment, for all $i \in I$,

$$\tilde{p}_i = 1 - F_i\left(\sum_{s \neq 0} (\pi_i^s - \pi_0^s)\left(1 - G_i\left(\frac{n\pi_0^s + \gamma \sum_{j \neq i} \tilde{p}_j \left(\pi_j^s - \pi_0^s\right) - \gamma \tilde{p}_i \pi_i^0 - \bar{l}_i}{n\pi_0^s + \gamma \sum_{j \neq i} \tilde{p}_j \left(\pi_j^s - \pi_0^s\right) + \gamma \tilde{p}_i \left(\pi_i^s - \pi_i^0\right)}\right)\right)\right) \tag{8}$$

for some constant $\bar{l}_i \geq 0$. Fix a vector $\bar{l} = \left(\bar{l}_i\right)_{i=1}^n \geq 0$, and let $\tilde{p}\left(\bar{l}\right) = \left(\tilde{p}_i\left(\bar{l}\right)\right)_{i \in I}$ denote a solution to (8). We claim that $\tilde{p}_i\left(\bar{l}\right) \geq p_i$ for all i.

To see this, recall that p is the unique fixed point of the function $h : [0,1]^n \to [0,1]^n$, where $h_i(p)$ is the unique solution p_i' to (7). Similarly, $\tilde{p}_i\left(\bar{l}\right)$ is the unique fixed point of the function $\tilde{h} : [0,1]^n \to [0,1]^n$, where $\tilde{h}_i(p)$ is the unique solution p_i' to

$$p_i' = 1 - F_i\left(\sum_{s \neq 0} (\pi_i^s - \pi_0^s)\left(1 - G_i\left(\frac{n\pi_0^s + \gamma \sum_{j \neq i} p_j \left(\pi_j^s - \pi_0^s\right) - \gamma p_i' \pi_i^0 - \bar{l}_i}{n\pi_0^s + \gamma \sum_{j \neq i} p_j \left(\pi_j^s - \pi_0^s\right) + \gamma p_i' \left(\pi_i^s - \pi_i^0\right)}\right)\right)\right).$$

Note that $\tilde{h}_i(p)$ is non-decreasing in p_j for all $j \in I$. In addition $h_i(p) \geq \tilde{h}_i(p)$ for all $i \in I$ and $p \in [0,1]^n$. As h and $\tilde{h}$ are monotone and continuous, and p and $\tilde{p}$ are the greatest fixed points of h and $\tilde{h}$, respectively, $\lim_{m \to \infty} h^m((1,\ldots,1)) = p$ and $\lim_{m \to \infty} \tilde{h}^m((1,\ldots,1)) = \tilde{p}$. Hence, $p \geq \tilde{p}$. ■

Proof of Proposition 9. Under the assumption $\min\left\{y_i^s, y_i^{s'}, y_i^{s'}, \tilde{y}_i^{s'}\right\} > 0$, the defender's FOC is

necessary and sufficient for optimality. Under the FOC,

$$y_i^{s'} = 1 - \frac{\gamma p_i \pi_i^{s'}}{n\pi_0^{s'} + \gamma \sum_j p_j \left(\pi_j^{s'} - \pi_0^{s'}\right)},$$
$$\tilde{y}_i^{s'} = 1 - \frac{\gamma \tilde{p}_i \pi_i^{s'} + \gamma f_i(\tilde{x}_i)\left(\pi_i^{s'} - \pi_0^{s'}\right) l_i(\tilde{y})}{n\pi_0^{s'} + \gamma \sum_j \tilde{p}_j \left(\pi_j^{s'} - \pi_0^{s'}\right)}.$$

Hence, $\tilde{y}_i^{s'} \leq y_i^{s'}$ if and only if

$$\frac{\gamma \tilde{p}_i \pi_i^{s'} + \gamma f_i(\tilde{x}_i)\left(\pi_i^{s'} - \pi_0^{s'}\right) l_i(\tilde{y})}{n\pi_0^{s'} + \gamma \sum_j \tilde{p}_j \left(\pi_j^{s'} - \pi_0^{s'}\right)} \geq \frac{\gamma p_i \pi_i^{s'}}{n\pi_0^{s'} + \gamma \sum_j p_j \left(\pi_j^{s'} - \pi_0^{s'}\right)}$$
$$\iff$$
$$\frac{1}{p_i}\left[\tilde{p}_i + f_i(\tilde{x}_i)\left(1 - \frac{\pi_0^{s'}}{\pi_i^{s'}}\right) l_i(\tilde{y})\right] \geq \frac{n\pi_0^{s'} + \gamma \sum_j \tilde{p}_j \left(\pi_j^{s'} - \pi_0^{s'}\right)}{n\pi_0^{s'} + \gamma \sum_j p_j \left(\pi_j^{s'} - \pi_0^{s'}\right)}. \tag{9}$$

If s and s' are comparable and s is more informative than s', then the left-hand side of (9) is greater for s than for s'. Hence, it suffices to show that

$$\frac{n\pi_0^{s} + \gamma \sum_j \tilde{p}_j \left(\pi_j^{s} - \pi_0^{s}\right)}{n\pi_0^{s} + \gamma \sum_j p_j \left(\pi_j^{s} - \pi_0^{s}\right)} \leq \frac{n\pi_0^{s'} + \gamma \sum_j \tilde{p}_j \left(\pi_j^{s'} - \pi_0^{s'}\right)}{n\pi_0^{s'} + \gamma \sum_j p_j \left(\pi_j^{s'} - \pi_0^{s'}\right)}.$$

Fixing i^* such that $\pi_i^s = \pi_0^s$ and $\pi_i^{s'} = \pi_0^{s'}$ for all $i \neq i^*$, this is equivalent to

$$\frac{n\pi_0^s + \gamma\tilde{p}_{i^*}\left(\pi_{i^*}^s - \pi_0^s\right)}{n\pi_0^s + \gamma p_{i^*}\left(\pi_{i^*}^s - \pi_0^s\right)} \leq \frac{n\pi_0^{s'} + \gamma\tilde{p}_{i^*}\left(\pi_{i^*}^{s'} - \pi_0^{s'}\right)}{n\pi_0^{s'} + \gamma p_{i^*}\left(\pi_{i^*}^{s'} - \pi_0^{s'}\right)}$$

$$\Longleftrightarrow$$

$$\left[n + \gamma\tilde{p}_{i^*}\left(\frac{\pi_{i^*}^s}{\pi_0^s} - 1\right)\right]\left[n + \gamma p_{i^*}\left(\frac{\pi_{i^*}^{s'}}{\pi_0^{s'}} - 1\right)\right] \leq \left[n + \gamma\tilde{p}_{i^*}\left(\frac{\pi_{i^*}^{s'}}{\pi_0^{s'}} - 1\right)\right]\left[n + \gamma p_{i^*}\left(\frac{\pi_{i^*}^s}{\pi_0^s} - 1\right)\right]$$

$$\Longleftrightarrow$$

$$\tilde{p}_{i^*}\left(\frac{\pi_{i^*}^s}{\pi_0^s} - 1\right) + p_{i^*}\left(\frac{\pi_{i^*}^{s'}}{\pi_0^{s'}} - 1\right) \leq \tilde{p}_{i^*}\left(\frac{\pi_{i^*}^{s'}}{\pi_0^{s'}} - 1\right) + p_{i^*}\left(\frac{\pi_{i^*}^s}{\pi_0^s} - 1\right)$$

$$\Longleftrightarrow$$

$$\tilde{p}_{i^*}\left(\frac{\pi_{i^*}^s}{\pi_0^s} - \frac{\pi_{i^*}^{s'}}{\pi_0^{s'}}\right) \leq p_{i^*}\left(\frac{\pi_{i^*}^s}{\pi_0^s} - \frac{\pi_{i^*}^{s'}}{\pi_0^{s'}}\right).$$

Since $\tilde{p}_{i^*} \leq p_{i^*}$ (by Proposition 8) and $\frac{\pi_{i^*}^s}{\pi_0^s} \geq \frac{\pi_{i^*}^{s'}}{\pi_0^{s'}}$ (as s is more informative than s'), this inequality is satisfied. ■

References

[1] Dilip Abreu, David Pearce, Ennio Stacchetti (1990): "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica*, Vol. 58, No. 5, pp. 1041-1063

[2] Daron Acemoglu and Alexander Wolitzky (2014): "Cycles of Conflict: An Economic Model," *American Economic Review*, Vol. 104, No. 4, pp. 1350-1367.

[3] James Adams (2001): "Virtual Defense," *Foreign Affairs,* Vol. 80, No. 3, pp. 98-112.

[4] Avenhaus, Rudolf and D. Marc Kilgour (2004): "Efficient Distributions of Arms-Control Inspection Effort," *Naval Research Logistics*, Vol. 51, No. 1, pp. 1-27

[5] Rudolf Avenhaus, Bernhard von Stengel, and Shmuel Zamir (2002): "Inspection Games," *Handbook of Game Theory with Economic Applications 3*, Elsevier, pp. 1947-1987.

[6] George Baker, Robert Gibbons and Kevin Murphy (1994): "Subjective Performance Measures in Optimal Incentive Contracts," *Quarterly Journal of Economics*, Vol. 109, No. 4, pp. 1125–1156.

[7] Sandeep Baliga and Tomas Sjöström (2004): "Arms Races and Negotiations," *Review of Economic Studies*, Vol. 71, No. 2, pp. 351-369.

[8] Bartholomew, Brian and Juan Andres Guerrero-Saade (2016): "Wave Your False flags! Deception Tactics Muddying Attribution in Targeted Attacks," *Virus Bulletin Conference.*

[9] Eli Berman, Jacob N. Shapiro, and Joseph H. Felter (2011): "Can Hearts and Minds be Bought? The Economics of Counterinsurgency in Iraq," *Journal of Political Economy*, Vol. 119, No. 4, pp. 766-819.

[10] David Blackwell (1951): "The Comparison of Experiments," in *Proceedings, Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press.

[11] Ben Buchanan (2017): *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*, Oxford University Press

[12] Sylvain Chassang, Gerard Padró i Miquel (2010): "Conflict and Deterrence under Strategic Risk," *Quarterly Journal of Economics*, Vol. 125, No. 4, pp. 1821–58

[13] Sylvain Chassang and Christian Zehnder (2016): "Rewards and Punishments: Informal Contracting through Social Preferences," *Theoretical Economics*, Vol. 11, No. 3, pp. 1145-1179.

[14] David D. Clark and Susan Landau (2010): "Untangling Attribution," *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for U.S. Policy.* National Academies Press.

[15] Richard A. Clarke and Robert K. Knake (2010): *Cyberwar: The Next Threat to National Security and What To Do About It,* Ecco.

[16] Department of Defense (2015): *The DoD Cyber Strategy.* Available at: `https://www.defense.gov/Portals/1/features/2015/0415_cyber-strategy/Final_2015_DoD_CYBER_STRATEGY_for_web.pdf`

[17] Edwards, Benjamin, Alexander Furnas, Stephanie Forrest, and Robert Axelrod (2017), "Strategic Aspects of Cyberattack, Attribution, and Blame," *Proceedings of the National Academy of Sciences*.

[18] James D. Fearon (1997): " Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs, " *Journal of Conflict Resolution* Vol. 41, No. 1, pp. 68-90.

[19] Peter Feaver and Kenneth Geers (2017): "'When the Urgency of Time and Circumstances Clearly Does Not Permit...': Pre-Delegation in Nuclear and Cyber Scenarios," *Understanding Cyber Conflict: 14 Analogies*, George Perkovich and Ariel E. Levite (eds.), Georgetown University Press.

[20] Goldsmith, Jack (2013): "How Cyber Changes the Laws of War," *European Journal of International Law*, Vol. 24, No. 1, pp. 129-138.

[21] M. J. Graetz, J. E. Reinganum, and L. L. Wilde (1986): "The Tax Compliance Game: Toward an Interactive Theory of Law Enforcement," *Journal of Law, Economics and Organization*, Vol. 2, No. 1, pp. 1-32.

[22] Edward J. Green and Robert H. Porter (1984): "Noncooperative Collusion under Imperfect Price Information," *Econometrica*, Vol. 52, No. 1, pp. 87-100.

[23] House Permanent Select Committee on Intelligence (2011): *The Cyber Threat.* `https://intelligence.house.gov/sites/intelligence.house.gov/files/documents/100411cyberhearinghayden.pdf`

[24] Ryusuke Hohzaki, (2007): "An Inspection Game with Multiple Inspectees", *European Journal of Operational Research*, Vol. 178, No. 3, pp. 894 – 906.

[25] Susan Hennessy (2017): "Deterring Cyberattacks: How to Reduce Vulnerability," *Foreign Affairs*.

[26] Information Warfare Monitor (2009): *Tracking GhostNet: Investing a Cyber Espionage Network*.

[27] Robert Jervis (1978): "Cooperation Under the Security Dilemma," *World Politics*, Vol. 30, No. 2, pp. 167-214.

[28] Fred Kaplan (2016): *Dark Territory: The Secret History of Cyber War,* Simon & Schuster.

[29] Lucas Kello (2017): *The Virtual Weapon*, Yale University Press: New Haven.

[30] Fahad Khalil (1997): "Auditing without Commitment," *RAND Journal of Economics*, Vol. 28, No. 4, pp. 629-640.

[31] Andrew Kydd (1997): "Game Theory and the Spiral Model," *World Politics*, Vol. 49, No. 3, pp. 371-400.

[32] Andrew Kydd and Barbara F. Walter (2002): "Sabotaging Peace: The Politics of Extremist Violence," *International Organization* 56(2):263–96.

[33] Henrik Lando (2006): "Does Wrongful Conviction Lower Deterrence?," *Journal of Legal Studies*, Vol. 35, No. 2, pp. 327-337.

[34] Martin C. Libicki, Lillian Ablon, and Tim Webb (2015): *The Defender's Dilemma: Charting a Course Toward Cybersecurity*, Rand Corporation.

[35] Lindsay, Jon R. (2015), "Tipping the Scales: The Attribution Problem and the Feasibility of Deterrence Against Cyberattack," *Journal of Cybersecurity*, Vol. 1, No. 1, pp. 53-67.

[36] Milgrom, Paul and John Roberts (1994), "Comparing Equilibria," *American Economic Review*, Vol. 84, No. 3, pp. 441-459.

[37] Dilip Mookherjee and Ivan Png (1989): "Optimal Auditing, Insurance, and Redistribution," *Quarterly Journal of Economics*, Vol. 104, No. 2, pp. 399-415

[38] Roger B. Myerson (2009): "Learning from Schelling's Strategy of Conflict," *Journal of Economic Literature* 47(4):1109–1125.

[39] Ellen Nakashima (2018): "Russian Spies Hacked the Olympics and Tried to Make it Look Like North Korea Did it, U.S. Officials Say," *Washington Post*, February 24.

[40] Joseph S. Nye Jr. (2011): "Nuclear Lessons for Cyber Security?" *Strategic Studies Quarterly*, Vol. 5, No. 4, pp. 18-38.

[41] Leon Panetta (2012): "Remarks by Secretary Panetta on Cybersecurity to the Business Executives for National Security," New York City. Available at: `http://archive.defense.gov/transcripts/transcript.aspx?transcriptid=5136`

[42] Ivan Png (1986): "Optimal Subsidies and Damages in the Presence of Judicial Error," *International Review of Law and Economics*, Vol. 6, No. 1, pp. 101-105

[43] A. Mitchell Polinsky and Steven Shavell (2000): "The Economic Theory of Public Enforcement of Law," *Journal of Economic Literature*, Vol. 38, No. 1, pp. 45-76.

[44] Robert Powell (1990): *Nuclear Deterrence theory: The Search for Credibility,* Cambridge University Press/

[45] Roy Radner (1986): "Repeated Principal-Agent Games with Discounting," *Econometrica*, Vol. 53, No. 5, pp. 1173-1198.

[46] Dominic Rohner, Mathias Thoenig and Fabrizio Zilibotti (2013): "War Signals: A Theory of Trade, Trust and Conflict," *Review of Economic Studies*, Vol. 80, No. 3, pp. 1114-1147

[47] Josh Rogin (2010): "The Top 10 Chinese Cyber Attacks (That We Know Of)," *Foreign Policy.*

[48] David Sanger and William Broad (2018): "Pentagon Suggests Countering Devastating Cyberattacks With Nuclear Arms," *New York Times*, January 16.

[49] Thomas C. Schelling (1960): *The Strategy of Conflict*, Harvard University Press: Cambridge.

[50] Noah Shachtman (2010): "Insiders Doubt 2008 Pentagon Hack was Foreign Spy Attack," *Wired*, August 25.

[51] P.W. Singer and Allan Friedman (2014): *Cybersecurity and Cyberwar: What Everyone Needs to Know,* Oxford University Press.

[52] Shaver, Andrew and Jacob N. Shapiro (2017): "The Effect of Civilian Casualties on Wartime Informing: Evidence from the Iraq War," *Journal of Conflict Resolution*, forthcoming.

[53] Vitaly Shevchenko (2014): "'Little Green Men' or 'Russian Invaders'?," *BBC*, March 11.

[54] Alastair Smith (1998): "International Crises and Domestic Politics " *American Political Science Review*, Vol. 92, No. 3, pp. 623-638.

[55] Glenn H. Snyder (1961): *Deterrence and Defense: Toward a Theory of National Security*, Princeton University Press.

[56] Stephen John Stedman (1997): "Spoiler Problems in Peace Processes," *International Security*, Vol. 22, No. 2, pp. 5-53.

[57] ThreatConnect (2017): "Guccifer 2.0: All Roads Lead to Russia," July 26.

[58] Robert F. Trager and Dessislava P. Zagorcheva (2006): "Deterring Terrorism: It Can Be Done," *International Security*, Vol. 30, No. 3, pp. 87-123.

[59] George Tsebelis (1989): "The Abuse of Probability In Political Analysis: The Robinson Crusoe Fallacy", *American Political Science Review*, Vol. 83, No. 1, pp. 77-91

[60] Livio Di Lonardo and Scott A. Tyson (2018): "Political Instability and the Failure of Deterrence," University of Michigan working paper.

[61] Franz J. Weissing and Elinor Ostrom (1991): "Irrigation Institutions and the Games Irrigators Play: Rule Enforcement without Guards," *Game Equilibrium Models II: Methods, Morals, and Markets*, Springer-Verlag, pp. 188-262.