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

April 26, 2019

Abstract

Motivated by recent developments in cyberwarfare, we study deterrence in a world where attacks cannot be perfectly attributed to attackers. In the model, each of n attackers may attack the defender. The defender observes an imperfect signal that probabilistically attributes the attack. The defender may retaliate against one or more attackers, and wants to retaliate against the guilty attacker only. We note an endogenous strategic complementarity among the attackers: if one attacker becomes more aggressive, that attacker becomes more "suspect" and the other attackers become less suspect, which leads the other attackers to become more aggressive as well. Despite this complementarity, there is a unique equilibrium. We identify conditions under which improving attribution strengthens deterrence—namely, improving attack detection independently of any effect on the identifiability of the attacker, reducing false alarms, or replacing misidentification with non-detection. However, we show that other improvements in attribution can backfire, weakening deterence—these include detecting more attacks where the attacker is difficult to identify or pursuing too much certainty in attribution. Deterrence is improved if the defender can commit to a retaliatory strategy in advance, but the defender should not always commit to retaliate more after every signal.

*We have received helpful comments and feedback from Daron Acemoglu, Scott Ashworth, Wiola Dziuda, Drew Fudenberg, Louis Kaplow, Roger Lagunoff, Robert Powell, Konstantin Sonin, Kathy Spier, and seminar audiences at A.S.S.A. 2019, Becker-Friedman Economic Theory Conference 2018, Chicago, ECARES, Georgetown, Harvard, LSE, Political Economy in the Chicago Area (P.E.C.A.) conference, UBC, and the Wallis Conference 2018. Zhaosong Ruan provided excellent research assistance. Wolitzky thanks the Sloan Foundation and the NSF for financial support.

"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 initiating hostilities in the first place (Schelling 1960; Snyder 1961; Myerson 2009). 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 (2017, p. 130); see also Clark and Landau (2010), Edwards et al. (2017), Goldsmith (2013), Lindsay (2015), and Nye (2011)): 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 (2011), 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 sometimes be difficult to determine who initiated a given attack.[1] The problem is amplified in counterinsurgency, where there is often uncertainty as to which of multiple terrorist or insurgent factions is responsible for an attack (Berman, Shapiro and Felter 2011; Shaver and Shapiro Forthcoming; Trager and Zagorcheva 2006). 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 (Segerson 1988; Weissing and Ostrom 1991). Similar issues can arise in other areas of law and economics (Shavell 1985; Png 1986; Lando 2006). 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

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

one defender. An attacker gets an opportunity to strike the defender. The defender observes a noisy 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 none 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 attacker attacks or faces retaliation.

A first observation is that the attribution problem generates an endogenous strategic complementarity among the potential attackers. This effect makes deterrence under imperfect attribution inherently multilateral. To see the idea, suppose attacker i becomes more aggressive. Then, whenever the defender detects an attack, her belief that attacker i was responsible increases, 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. However, despite this complementarity, our model has a unique equilibrium, which substantially simplifies the analysis.

Given that attribution problems create challenges for deterrence, a standard intuition in both academic and policy circles is that improving attribution will improve deterrence (Panetta 2012; Department of Defense 2015). We examine this idea and show that, while some types of improvement in the defender's information structure do always improve deterrence, others can backfire and actually increase attacks. For example, improving detection always reduces attacks if the perpetrator responsible for the newly detected attacks can also be unambiguously identified, or if the processes of detecting an attack and identifying the responsible party are statistically independent. Reducing false alarms also always strengthens deterrence. However, improving detection can increase attacks if those responsible for the newly detected attacks are especially difficult to identify; this follows because misidentification is "worse" than non-detection, since the defender is reluctant to retaliate against other attackers after a signal that could result from misidentification. It is also not always true that giving the defender more information in the sense of Blackwell (1951) improves deterrence: in particular, the defender need not benefit from further refinements of a signal that is already strong enough to cause retaliation. This implies that it is often a mistake to pursue too much certainty in attribution.

We consider two additional applications of our model to problems in contemporary cyber strategy. The National Cyber Strategy of the United States (2018) allows the use of conventional or cyber weapons in response to a cyberattack. When is this flexibility welfare-improving for the defender? In our model, we show that the defender always benefits from gaining access to a new retaliatory weapon, provided it is more destructive than all previously feasible means of retaliation; in contrast, gaining access to a new, less destructive weapon can sometimes undermine deterrence. We also consider the possibility that attackers can launch "false-flag" operations, attempting to mimic a different attacker. Such attacks seem particularly prevalent in the cyber domain (Bartholomew and Guerrero-Saade 2016). Here, we find that more aggressive attackers are more likely to be mimicked, as are attackers whose attacks are easier to detect and attribute. These predictions of the model are consistent with descriptions of known false-flag operations in the qualitative literature.

Finally, 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 a signal's 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 (2017) or Singer and Friedman (2014). These results also suggest a critique of the strategic shift articulated in the US Department of Defense's 2018 Cyber Strategy, which calls for a narrower focus on America's most aggressive and capable cyber adversaries, China and Russia (Department of Defense 2018). By contrast, we find that the optimal deterrence policy should target additional retaliation not against the most aggressive attackers but against those who are most deterrable—that is, those whose attacks are particularly easy to attribute and whose behavior is most responsive to changes in the likelihood of facing retaliation.

As further background for our study, we note that false alarms, detection failures, and misidentification have all arisen in major cyber incidents.

In one of the best-known and most successful cyberattacks to date, 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 Friedman 2014). Stuxnet was eventually uncovered not by the Iranians, 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 startling 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 through which the attacks had been routed. Ultimately, the attacks turned out to be the work of two sixteen-year olds in San Francisco and an eighteen-year old Israeli (Adams 2001; Kaplan 2016). 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 hack was done by Russian security agencies who had tried to cover their tracks by pretending to be Guccifer 2.0 (see the findings in ThreatConnect 2016). These are cases of *misidentification.*

Finally, in 2008, a worm on Department of Defense computers was found to have gained access to an enormous quantity 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 2010). This may, then, have been a case of a *false alarm*.[2]

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 2018 Pyeongchang Winter Olympics. In a false-flag operation, the GRU used North Korean IP addresses to deflect suspicion onto North Korea (Nakashima 2018), 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 that the I.O.C. loses on net if it expels an innocent country, but benefits if it expels the guilty one. We discuss additional examples after presenting the model.

This paper relates to several literatures. A large literature explores aspects of deterrence other than the attribution problem. Schelling (1960) explained the logic of deterrence and the importance of commitment. Jervis (1978) noted that, when the motives of a player who acquires arms for deterrence are not known to his opponent, the opponent may react by arming to protect himself from predation. This "security dilemma" equally applies to cyberweapons (Buchanan 2017). The security dilemma has been formalized using the idea that arms might be strategic complements rather than substitutes (Kydd 1997; Baliga and Sjöström 2004; Chassang and Padró i Miquel 2010). For example, Chassang and Padró i Miquel (2010) show that, in a coordination game, arms acquisition can increase preemptive incentives to go to war faster than it reduces incentives to predate. Hence, arming may cause escalation rather than deterrence. Acemoglu and Wolitzky (2014) 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.[3] Another literature explores the search for credibility, especially the role played by domestic politics (see, for example, Fearon 1997; Powell 1990; Smith 1998; Di Lonardo and Tyson 2018). We abstract from these themes in order to focus on the implications of attribution problems for deterrence with multiple attackers.

[2] Another interesting false alarm occurred in the run-up to the 2018 US midterm elections, when the Democratic National Committee notified the F.B.I. that it had detected what appeared to be an attempt by Russian hackers to infiltrate its voter database. The "attack" turned out to be the work of hackers hired by the Michigan Democratic Party to simulate a Russian incursion (Sullivan, Weiland and Conger 2018).

[3] Rohner, Thoenig and Zilibotti (2013) study the impact of trust on trade in a two period game where one player learns whether the other side is aggressive through its first period action.

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 2002, 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 that a Blackwell-improvement in information can make the defender worse off (without commitment)—this appears to be a novel result for inspection games. 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, von Stengel and Zamir 2002; Hohzaki 2007).

Inspection games appear in economics in the guise of "auditing games," where a principal tries to 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 1989). The principal also faces a commitment problem, as she may not have an incentive to monitor the agent ex post (Graetz, Reinganum and Wilde 1986; Khalil 1997). 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 (2006) and Section 8 of Polinsky and Shavell (2000)). 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 strategic complementarity cannot be analyzed. 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.[4]

There is also a literature on "crime waves" that models crime as a game of strategic complements among criminals: the more potential criminals commit crimes, the more law enforcement resources are strained, and the greater the incentive to commit additional crimes (Sah 1991; Glaeser, Sacerdote and Scheinkman 1996; Schrag and Scotchmer 1997; Bar-Gill and Harel 2001; Freeman, Grogger and Sonstelie 1996; Ferrer 2010; Bassetto and Phelan 2008; Bond and Hagerty 2010). This complementarity is related to the one in our model, if we interpret the defender's supply of "suspicion" as a fixed resource: the more one attacker attacks, the more suspect he becomes, and the less suspicion is left for other attackers. However, the crime waves literature emphasizes the possibility

[4]The one-inspectee inspection game also arises in law and economics. Tsebelis (1989) 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.

of multiple equilibria with different levels of crime, while our model has a unique equilibrium. This is because suspicion is a special kind of resource, which in particular responds to the *relative* attack probabilities of different attackers rather than the absolute attack probabilities: if all attackers double their attack probabilities, they remain equally suspicious (in fact more suspicious, because the relative probability of a false alarm has decreased), and thus face just as much retaliation. Our analysis is thus quite different from this literature, despite sharing the common theme of strategic complementarity.

Finally, repeated games with imperfect monitoring model multilateral moral hazard without commitment (Radner 1986; Green and Porter 1984; Abreu, Pearce and Stacchetti 1990). 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 (2016) 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 have taken a pro-social action. Their approach is an alternative to relational contracting models of intertemporal incentives (Baker, Gibbons and Murphy 1994).

1 A Model of Deterrence with Imperfect Attribution

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 to attack or not. 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 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 positive density on its support.

The defender receives a payoff of $-K$ if 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 positive density on its 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, $r^s(y) = (r_i^s(y))_{i \in I} \in [0, 1]^n$.[5] 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 if 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 $y_i^{s*} \in [0, 1]$.[6] 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 $p = (p_i)_{i \in I} \in [0, 1]^n$ for the attackers and a vector of retaliation probabilities $r = (r_i^s)_{i \in I, s \in S} \in [0, 1]^{n|S|}$ 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 perceiving "business as usual". We make the following two assumptions.

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

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

[5] 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 captured by the signals.

[6] 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.

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) $F_i(1) < 1$ 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.

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 1960; Snyder 1961; Powell 1990) to isolate the new issue of how imperfect attribution affects deterrence. In reality, there are several possible benefits of successful retaliation. Retaliation can disrupt an ongoing attack. It can also provide reputational benefits vis a vis other potential attackers and thus prevent additional attacks. And it can also satisfy a "taste for vengeance," which could result from psychological or political economy concerns (Jervis 1979; McDermott, Lopez and Hatemi 2017).

Related to this point, it may seem unlikely that a victim would ever retaliate against two different countries for the same cyberattack, as our model allows. This possibility can be ruled out by assuming that $y_i < \frac{1}{2}$ for all $i \in I$ with probability 1, which, as we will see, implies that the defender retaliates against a given attacker only if she believes that he is guilty with probability at least $1 - y_i > \frac{1}{2}$—a condition that cannot be satisfied for two attackers simultaneously.

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 all other attackers become less suspect after every non-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 probability 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. If multiple attackers can attack simultaneously, our model continues to apply if the payoff consequences of each attack (and any subsequent retaliation) are additively separable and signals are independent across attacks.

Fourth, 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 used cyberattacks to disrupt the Iranian nuclear program. Cyberattacks have also been used to incapacitate an adversary's military capabilities—for instance by disrupting communications or intelligence—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 2015).

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.

Finally, a signal s should be interpreted as containing all information available to the defender concerning the origin of a potential attack. This may include, for example, the systems targeted by the attack, the location of the servers where the attack originated, and the language and style of any malicious code.

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 simple complementarity is a key factor in many of our results.

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 results because an attack by i provides a benefit of x_i, while raising the probability of facing retaliation from $\sum_s \pi_0^s r_i^s$ to $\sum_s \pi_i^s r_i^s$.

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

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

Now, at the optimum, the defender retaliates against attacker i after signal s if and only if her benefit of retaliating against him (y_i) exceeds her cost of doing so, which is $1 - \beta_i^s(p)$, the probability that he is "innocent."

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*

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

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

and

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

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.

The strategic complementarity in our model can now be seen 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.

Let us clarify a potential point of confusion. If attacker i attacks with higher probability (p_i increases) while all other attack probabilities are held fixed and the defender is allowed to respond optimally, the effect on the *total* probability that another attacker j faces retaliation, evaluated ex ante at the beginning of the game, is ambiguous: attacker j is less suspect (and therefore faces less retaliation) after any given attack, but the total probability that an attack occurs increases. However, only the former effect—the probability of facing retaliation after a given attack—matters for j's incentives, because j cannot affect the probability that he is retaliated against in error after one of i's attacks. In other words, strategic complementarity operates entirely through the "intensive" margin of the retaliation probability following a given attack, not the "extensive" margin of the total number of attacks.

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*

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

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

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', and thus equals p_i' at a unique point in the unit interval. 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 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 (1994). ■

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

Notwithstanding the strategic complementarity in the model, there is always a unique equilibrium. As discussed in the Introduction, this is in stark contrast to standard models of crime waves, which emphasize multiple equilibria. To see the intuition, suppose there are two equilibria and attacker i's attack probability increases by the greatest proportion (among all attackers) in the second equilibrium relative to the first. Then, because the defender's beliefs are determined by the attackers' relative attack probabilities, attacker i is more suspect after every signal in the second equilibrium. The defender therefore retaliates against attacker i more often in the second equilibrium. But then attacker i should attack less in the second equilibrium, not more.

Theorem 1 *There is a unique equilibrium.*

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

Fix $i \in \operatorname{argmax}_{j \in I} \frac{p_j^*}{p_j^{**}}$. As p^* is the greatest equilibrium, we have $\frac{p_i^*}{p_i^{**}} \geq 1$. Therefore, for every $s \neq 0$,

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

where the first inequality holds because $\frac{p_i^{**}}{p_i^*} \leq \frac{p_j^{**}}{p_j^*}$ for all $j \in I$ and $\pi_j^s - \pi_0^s \geq 0$ for all $j \in I$ and $s \neq 0$, and the second inequality holds because $\frac{p_i^{**}}{p_i^*} \leq 1$. Notice this implies

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

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

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.

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_i 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_i 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.

Proposition 2 *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 logic of endogenous strategic complementarity plays a role throughout the paper, including in our analysis of false-flag operations (Section 5.2) and the commitment solution (Section 6). In those sections, we discuss how this mechanism appears consistent with a variety of accounts in the qualitative literature.

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 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.

Proposition 3 *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 2, noting that increasing G in the FOSD order shifts h down. ■

The effects of defender aggressiveness are important for our subsequent discussion of changes in the defender's retaliation technology (Section 5.1) and the commitment solution (Section 6). Those sections discuss the match between these effects and descriptions in the qualitative literature.

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 subsection 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 4 *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 $s \in S\backslash\{0\}$.*

Proposition 4 is relevant for assessing aspects of the US Department of Defense's 2018 Cyber Strategy (Department of Defense 2018). This policy document shifts US strategy, emphasizing retaliation against the largest and most aggressive adversaries, especially Russia and China. Proposition 4 emphasizes that such a shift should not be taken too far. The more the US focuses on what are inherently the most aggressive adversaries, the more aggressive other actors become in response. We provide a more detailed discussion of the 2018 Cyber Strategy in the context of the commitment model in Section 6.

4 Changes in 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 the 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 2012):

> 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.

This view was also ensconced in the Department of Defense's 2015 Cyber Strategy (Department of Defense 2015), which states:

> 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.

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 some more nuanced versions of this claim.

Roughly speaking, we show that the following types of improvements in the information structure always improve deterrence:

1. Improving detection if the perpetrators of the newly detected attacks are always identified correctly.

2. Replacing misidentification with non-detection.

3. Reducing false alarms.

4. Improving detection independently of identification.

However, two types of improvements can backfire and increase equilibrium attack probabilities:

1. Refining signals that are already strong enough to cause retaliation.

2. Improving detection if the perpetrators of the newly detected attacks are especially hard to identify.

Thus, from a policy perspective, some care must be taken in investing in improved detection and attribution technologies. In particular, a defender need not benefit from further refining a signal that is already strong enough to spark retaliation, and improvements in detection technology are only valuable if the newly detected signals can also be attributed with some degree of success.[7]

We organize our results as follows. First, we present two main results—Theorems 2 and 3—that provide sufficient conditions for a change in the information structure to improve deterrence. We then show how these results imply the four "positive" claims above as corollaries. Finally, we provide examples showing that the conditions for Theorems 2 and 3 cannot be relaxed, which yield the two "negative" claims above.

[7] These results rely on the assumption that the attackers know the defender's information structure: of course, if the defender can improve her information without the attackers' knowledge, this can only make her better off. However, it is clear that the same effects would arise in a more realistic model where attackers observe the defender's information structure imperfectly. The case where attackers are completely unware of improvements in the defender's information strikes us as less reaslistic.

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 Sufficient Conditions for a Change in the Information Structure to Improve Deterrence

This subsection presents general sufficient conditions for a change in the information structure to improve deterrence.

Let $r_i^s(p;\pi)$ be the probability that attacker i faces retaliation given signal s, prior attack probabilities p, and information structure π:

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

where $\beta_i^s(p;\pi)$ is given by equation (3), and we have made the dependence of β on π explicit. Let $x_i(p;\pi)$ be the increase in the probability that attacker i faces retaliation when he attacks given prior attack probabilities p and information structure π:

$$x_i(p;\pi) = \sum_{s\neq 0}\left(\pi_i^s - \pi_0^s\right) r_i^s(p;\pi).$$

Recall that, in equilibrium, $x_i^* = x_i(p;\pi)$.

Our first main result is that, if the information structure changes such that the defender becomes "more retaliatory," in that all cutoffs $x_i(p;\pi)$ increase *holding the attack probabilities fixed*, then in equilibrium all attack probabilities must decrease. Intuitively, this is a consequence of strategic complementarity: if π changes so that each $x_i(p;\pi)$ increases for fixed p, strategic complementarity then pushes all cutoffs even further up.

Theorem 2 *Fix two information structures π and $\tilde{\pi}$, and let p (resp. $\tilde{p}$) be the vector of equilibrium attack probabilities under π (resp. $\tilde{\pi}$). If $x_i(p;\tilde{\pi}) \geq x_i(p;\pi)$ for all $i \in I$, then $\tilde{p}_i \leq p_i$ for all $i \in I$. If in addition $x_i(p;\tilde{\pi}) > x_i(p;\pi)$ for some $i \in I$, then $\tilde{p}_i < p_i$.*

An important consequence of this result is the following: Suppose, conditional on an attack by i, weight is shifted from a signal s where i did not face retaliation to a signal s' where *no one else* faced retaliation. This always improves deterrence. The logic is that, holding the attack probabilities fixed, such a change in the information structure induces weakly more retaliation against i (at signal s', since i has become more suspect at s') and also induces weakly more retaliation against everyone else (at signal s, since everyone else has become more suspect at s). Theorem 2 then implies that all equilibrium attack probabilities must decrease.

Theorem 3 *Suppose that, with information structure π, there is a signal s where attacker i faces no retaliation (i.e. $r_i^s = 0$) and a signal s' where no other attacker j faces retaliation (i.e. $r_j^{s'} = 0$ for all $j \neq i$). Suppose also that, conditional on an attack by i, information structure $\tilde{\pi}$ shifts weight from signal s to signal 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 $(j, \hat{s}) \neq (i, s), (i, s')$. Then $\tilde{p}_j \leq p_j$ for all $j \in I$. Moreover, if either $r_i^{s'} > 0$ or $r_j^s > 0$ for some $j \neq i$, then $\tilde{p}_j < p_j$ for some $j \in I$.*

4.2 Types of Changes that Always Improve Deterrence

We can now derive the "positive" results previewed above.

4.2.1 Improving Detection without Increasing Misidentification

First, shifting mass from the null signal to a signal that never sparks mistaken retaliation always improves deterrence. For example, simultaneously improving both detection and identification—in that some attacks that previously went undetected are now both detected and unambiguously attributed—always improves deterrence.

Corollary 1 *Suppose that, with information structure π, there is a non-null signal s where all attackers $j \neq i$ face no retaliation (i.e. $r_j^s = 0$ for all $j \neq i$).*[8] *If, conditional on an attack by i, $\tilde{\pi}$ shifts weight from the null signal to signal s, then $\tilde{p}_j \leq p_j$ for all $j \in I$.*

Proof. Since $r_i^0 = 0$ and $r_j^s = 0$ for all $j \neq i$, this follows from Theorem 3. ■

4.2.2 Replacing Misidentification with Non-Detection

Second, misidentification is worse than non-detection, in the following sense: if it is possible that an attack by i is detected but is not attributed to i with enough confidence to cause retaliation, the defender would be better off if this attack were not detected at all. The intuition is that this change

[8] A trivial condition on primitives that guarantees $r_j^s = 0$ for all $j \neq i$ is $\pi_j^s = 0$ for all $j \neq i$: that is, signal s can only arise as a result of an attack by i or a false alarm.

does not affect i's incentive to attack, but it increases retaliation against everyone else because they become more suspect after signals that previously could have resulted from an attack by i.

Corollary 2 *Suppose that, with information structure π, there is a non-null signal s where attacker i faces no retaliation (i.e. $r_i^s = 0$). If, conditional on an attack by i, $\tilde{\pi}$ shifts weight from signal s to the null signal, then $\tilde{p}_j \leq p_j$ for all $j \in I$.*

Proof. Since $r_j^0 = 0$ for all $j \neq i$, this follows from Theorem 3. ■

4.2.3 Reducing False Alarms

Third, reducing false alarms (i.e., decreasing π_0^s for $s \neq 0$) always improves deterrence. When false alarms are less frequent, each non-null signal invites greater suspicion, and hence more retaliation. Also, the marginal impact of an attack on the probability of each non-null signal increases. Both of these effects increase the marginal impact of an attack on the probability of facing retaliation, and hence reduce the incentive to attack.

Corollary 3 *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 $\tilde{p}_i \leq p_i$ for all $i \in I$. Also, $\tilde{r}_i^s \geq r_i^s$ for all $s \neq 0$ and all $i \in I$.*

Proof. By Theorem 2, it suffices to show that $x_i(p; \tilde{\pi}) \geq x_i(p; \pi)$ for all i. By the definition of $x_i(p; \pi)$, since reducing false alarms increases $\pi_i^s - \pi_0^s$ for all $s \neq 0$, it suffices to show that $r_i^s(p; \tilde{\pi}) \geq r_i^s(p; \pi)$ for all $s \neq 0$. For this, it is in turn enough to show that $\beta_i^s(p; \tilde{\pi}) \geq \beta_i^s(p; \pi)$ for all $s \neq 0$. But this is immediate from equation (3). ■

4.2.4 Improving Detection Independently of Identification

Fourth, in the important special case of our model where the detection and identification processes are independent, improving detection always improves deterrence. To formulate this special case, suppose there exists a common *detection probability* $\delta \in [0, 1]$, a *false alarm probability* $\phi \in [0, 1]$, and a vector of *identification probabilities* $(\rho_i^s) \in [0, 1]^{n|S|}$ such that

$$\pi_i^0 = 1 - \delta \text{ for all } i \neq 0$$

$$\pi_i^s = \delta \rho_i^s \text{ for all } i, s \neq 0$$

$$\pi_0^0 = 1 - \phi$$

$$\pi_0^s = \phi \rho_0^s \text{ for all } s \neq 0.$$

Corollary 4 *If detection is independent of identification, improving detection decreases all equilibrium attack probabilities.*

Proof. By Theorem 2, it suffices to show that $\beta_i^s(p;\tilde{\pi}) \geq \beta_i^s(p;\pi)$ for all i and all $s \neq 0$. We have

$$\beta_i^s(p;\pi) = \frac{\gamma\delta p_i \rho_i^s}{\gamma\delta \sum_j p_j \rho_j^s + \left(n - \gamma \sum_j p_j\right) \phi \rho_0^s}.$$

Clearly, $\beta_i^s(p;\pi)$ is non-decreasing in δ. ■

Moreover, note that $\beta_i^s(p;\pi)$ depends on the detection probability and the false alarm probability only through their ratio δ/ϕ. Thus, when detection is independent of identification, improving detection is strategically equivalent to reducing false alarms.

4.3 Types of Changes that Can Degrade Deterrence

We now give our "negative" results. We can organize these results by showing why the conclusion of Theorem 3 can fail if either $r_i^s > 0$ or $r_j^{s'} > 0$ for some $j \neq i$.

4.3.1 Improving Detection while Worsening Identification

We first show how deterrence can be undermined by improving detection but simultaneously worsening identification. That is, shifting weight from the null signal to a signal where someone other than the attacker faces retaliation can reduce retaliation against both attackers and increase attacks. This is a partial converse to the result that replacing misidentification with non-detection improves deterrence (Corollary 2).

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 $y_1 = y_2 = \frac{1}{4}$ with probability 1.[9]

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

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, but it may be "confused" with an attack by attacker 2. In equilibrium, this causes the defender to stop retaliating after $s = 2$, which leads the less aggressive type of attacker 2 to start attacking, which in turn causes the defender to stop retaliating after $s = 1$ as well.

Claim 2 *In the unique equilibrium with information structure $\tilde{\pi}$, both attackers attack whenever they have the opportunity, and the defender never retaliates. Thus, $p_1 = p_2 = 1$.*

4.3.2 Refining Signals that Already Cause Retaliation

Deterrence can also be undermined by refining a signal that is already strong enough to cause retaliation. This can occur even if the signal refinement corresponds to a strict improvement in the information structure in the sense of Blackwell (1951), and even if there is only one attacker, so that the model is a classical inspection game (Avenhaus, von Stengel and Zamir 2002). That is, a Blackwell-improvement in the defender's information can reduce her payoff in an inspection game.[10]

To see the intuition, suppose there is a single attacker and there are three possible signals: null, imperfectly informative, and perfectly informative. Suppose the perfect signal is rare, so that even certain retaliation following the perfect signal is not enough to deter an attack on its own. The defender must then also be willing to retaliate following an imperfect signal. Moreover,

[9] 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.

[10] As far as we know, this is a novel observation. One somewhat related result is due to Crèmer (1995), who shows that, in a principal-agent model, the principal may benefit from having less information about the agent's performance, because this makes it credible to carry out certain threats, such as failing to renegotiate the contract.

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 is higher 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

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

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

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

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. The intuition is that, when weight shifts from π_1^1 to π_1^2, the attacker must attack with higher probability to keep the defender willing to retaliate after signal 1.

More generally, deterrence is undermined by extra information in regions of the defender's belief space where the probability of retaliating against a given attacker is concave in the defender's posterior belief about whether that attacker attacked. Since this is typically the case when the defender is almost certain the attacker attacked (as then she retaliates with probability close to 1), this implies that pursuing too much certainty in attribution is usually a mistake.

Of course, for any fixed attack probabilities, the defender benefit from having additional information, as this improves the accuracy of her retaliation decisions. Thus, in general, if the effect of improving the defender's information on deterrence is positive, the overall effect on the defender's payoff is positive; while if the effect on deterrence is negative, the overall effect can go either way.

5 Applications

We now explore two applications of particular relevance to contemporary discussions surrounding cyber strategy.

Section 5.1 considers the possibility that the defender may have multiple ways to retaliate, for example with a less destructive weapon (like a reciprocal cyberattack) or a more destructive one (like a conventional military, or even nuclear, attack). Our main result is that adding a more destructive weapon to the defender's arsenal always improves deterrence, while adding a less destructive weapon can undermine deterrence.

Section 5.2 asks what happens when one attacker can attempt to mimic another attacker via a false-flag operation. Here we show that more aggressive attackers are more likely to be mimicked, as are (more surprisingly) attackers who are themselves easy to detect and identify when they attack.

5.1 Different Kinds of Retaliation

A central debate in cyber strategy concerns what weapons should be available for retaliation against a cyber attack. This question was raised with new urgency by the 2018 United States Nuclear Posture Review, which for the first time allowed the possibility of first-use of nuclear weapons in response to devastating but non-nuclear attacks, including cyberattacks (Sanger and Broad 2018). Less dramatically, the 2018 National Cyber Strategy puts both kinetic and cyber retaliation on the table as possible responses to cyber activity (United States 2018).

Our model can capture many aspects of this debate, but not all of them. In particular, we do model the fact that a more destructive form of retaliation is likely more costly to use in error, but we cannot capture all possible objections to the Nuclear Posture Review, such as the potential consequences of "normalizing" first-use of nuclear weapons. Nonetheless, in the context of our model, we provide some support for the spirit of the Nuclear Posture Review by showing that adding a more destructive weapon to the defender's arsenal always improves deterrence. By contrast, adding a less destructive weapon to the defender's arsenal has competing effects and, as such, can either weaken or strengthen deterrence.

We model introducing a new retaliation weapon into the defender's arsenal as follows: There is the original, legacy weapon ℓ, and a new weapon, n. Each weapon $a \in \{\ell, n\}$ is characterized by three numbers: the damage it does to an attacker, w^a (previously normalized to 1), the benefit using it provides to a type-y defender, y^a, and the cost to the defender of using it on an innocent attacker, z^a (previously normalized to 1). Thus, when the defender observes signal s and forms belief β_i^s that attacker i is "guilty," she retaliates using the weapon $a \in \{0, \ell, n\}$ that maximizes

$$y^a - (1 - \beta_i^s)\, z^a,$$

where $a = 0$ corresponds to not retaliating, with $y^0 = z^0 = w^0 = 0$. We continue to assume that $K > y^a$ for all $y \in \operatorname{supp} G$ and all a, so that deterring an attack is preferred to being attacked and retaliating.

A couple points are worth noting. All else equal, a defender prefers to retaliate with a weapon that provides higher retaliatory benefits (higher y^a) and lower costs for mistaken retaliation (lower z^a). It seems reasonable to assume that these two features of a weapon may co-vary positively—stronger weapons provide both better retaliatory benefits and are more costly when misused. So the defender may face a trade-off, and she will balance this trade-off differently following different signals. In particular, her willingness to use a weapon with larger costs for mistakes depends on her uncertainty about the identity of the perpetrator: When attribution is quite certain, the defender is more willing to opt for a powerful response. When attribution is less certain, the defender will want to respond in a way that limits costs in case of a mistake.

In light of this tradeoff, we ask when introducing the new weapon into the arsenal improves the defender's payoff.

First, it is easy to construct examples where introducing a weaker weapon (i.e., one with $w^n < w^\ell$) into the defender's arsenal makes her worse-off. For example, suppose that the new weapon also imposes lower costs when used in error ($z^n < z^\ell$). Then there could be signals where the defender would have used the legacy weapon, but now switches to the new weapon. (Indeed, if $y^n > y^\ell$ then the defender never uses the legacy weapon.) If $w^\ell - w^n$ is sufficiently large this undermines deterrence, and the defender is made worse-off overall if the cost of being attacked (K) is sufficiently large. The intuition is that, when a weaker weapon is available, ex post the defender is sometimes tempted to use it rather than the stronger weapon (in particular, when she is uncertain of the identify of the perpetrator). This is bad for ex ante deterrence. The defender can thus benefit from committing in advance to never retaliate with a less destructive weapon.

By contrast, introducing a new weapon that imposes greater costs on attackers (i.e., $w^n \geq w^\ell$) always benefits the defender.[11] The intuition is that, holding the attack probabilities fixed, making a new, more destructive weapon available weakly increases the expected disutility inflicted on every attacker: this follows because, for each signal, the defender's optimal response either remains unchanged or switches to the new, more damaging weapon. This reduces everyone's incentive to attack, and strategic complementarity then reduces the equilibrium attack probabilities even more.

Proposition 5 *Assume $w^n \geq w^\ell$.*

Let p (resp. $\tilde{p}$) denote the equilibrium attack probabilities when the new weapon is unavailable (resp., available). Then $p \geq \tilde{p}$.

Let u (resp. $\tilde{u}$) denote the defender's equilibrium payoff when the new weapon is unavailable (resp., available). Then $u \leq \tilde{u}$.

5.2 False Flags

The attribution problem creates the possibility for false-flag operations, where one attacker poses as another to evade responsibility. False-flag operations are common in the cyber context (see Bartholomew and Guerrero-Saade 2016). For instance, we have already discussed Russia's attempt to mask their attack on the Pyeongchang Olympics by routing the attack through North Korean servers.

A false-flag operation amounts to one attacker attempting to attack in a way that mimics, or is likely to be attributed to, another attacker. If multiple attackers can mimic each other, there will naturally be multiple equilibria, where different attackers are mimicked most often, due to a coordination motive in mimicking. As our main question of interest here is who is mostly likely to be mimicked, we rule out this effect by assuming that only attacker 1 has the ability to mimic other attackers.

For simplicity, in this subsection we consider a version of the "independent detection and identification" model of Section 4.2.4, while allowing the detection probability to vary across attackers.

[11] It is straightforward to generalize this result to the case where there are many legacy weapons. In this case, the required condition is that the new weapon is more destructive than any of them.

In particular, we assume the information structure is

$$\pi_i^0 = 1 - \delta_i \text{ for all } i \neq 0$$

$$\pi_i^i = \delta_i \rho_i \text{ for all } i \neq 0$$

$$\pi_i^j = \delta_i \frac{1-\rho_i}{n-1} \text{ for all } i \neq j \neq 0$$

$$\pi_0^0 = 1 - \phi$$

$$\pi_0^s = \frac{\phi}{n} \text{ for all } s \neq 0.$$

Thus, attackers differ in how detectable they are (δ_i) and how identifiable they are (ρ_i), but the information structure is otherwise symmetric.

The "mimic" (attacker 1) chooses an attack probability p_1 and, conditional on attacking, a probability distribution over whom to mimic, $\alpha \in \Delta(I)$. Given α, if the mimic attacks, signal $s = 0$ realizes with probability $1 - \delta_1$ and each signal $i \neq 0$ realizes with probability

$$\pi_1^i(\alpha) := \delta_1 \left(\alpha_i \chi_i + \sum_{j \neq i} \alpha_j \frac{1 - \chi_j}{n-1} \right),$$

where $\chi_i \in (0,1)$ measures 1's ability to successfully mimic attacker i. For example, an attacker with a less sophisticated arsenal of cyber weapons may be easier to mimic.

If the mimic chooses strategy α, for $i \neq 1$, we have

$$\beta_1^i(\alpha) = \frac{\gamma p_1 \pi_1^i(\alpha)}{\gamma \left[p_1 \pi_1^i(\alpha) + \delta_i p_i \rho_i + \sum_{j \neq 1,i} \delta_j p_j \frac{1-\rho_j}{n-1} \right] + \left(1 - \frac{\gamma}{n} \sum_j p_j \right) \frac{\phi}{n}}.$$

Denote the probability with which the mimic faces retaliation at signal s by

$$r_1^s(\alpha) = 1 - G_1\left(1 - \beta_1^s(\alpha)\right)$$

Given the vector of attack probabilities p (including p_1), the mimic chooses α to solve

$$\min_{\alpha' \in \Delta(I)} \sum_{s \in I} \pi_1^s(\alpha') \, r_1^s(\alpha).$$

(Note that α is fixed here by equilibrium expectations.) The derivative with respect to α_i' is

$$\delta_1 \left(\chi_i r_1^i(\alpha) + \sum_{j \neq i} \frac{1-\chi_i}{n-1} r_1^j(\alpha) \right).$$

Thus, at the optimum, this derivative must be equal for all $i \in \operatorname{supp} \alpha$, and must be weakly greater for all $i \notin \operatorname{supp} \alpha$. In particular, if $i, i' \in \operatorname{supp} \alpha$, we have

$$\chi_i \left(\frac{1}{n} \sum_{j \in I} r_1^j(\alpha) - r_1^i(\alpha) \right) = \chi_{i'} \left(\frac{1}{n} \sum_{j \in I} r_1^j(\alpha) - r_1^{i'}(\alpha) \right),$$

where both terms in parentheses are non-negative. Note that $r_1^i(\alpha)$ is increasing in $\beta_1^i(\alpha)$, which in turn is increasing in $\pi_1^i(\alpha)$ and decreasing in δ_i, p_i, and ρ_i. We obtain the following result:

Proposition 6 *Ceteris paribus, an attacker is mimicked more in equilibrium if he is more aggressive, easier to identify, easier to detect, or easier to mimic: for any two attackers $i, j \neq 1$, if $p_i \geq p_j$, $\rho_i \geq \rho_j$, $\delta_i \geq \delta_j$, and $\chi_i \geq \chi_j$, then $\alpha_i \geq \alpha_j$.*

The intuition for why more aggressive attackers are more like to be the victim of false-flag operations is that such attackers are more suspect when the signal points to them, which makes the mimic less suspect. The same intuition also explains the apparently more subtle result that attackers that are easier to identify or detect are mimicked more: When such an attacker attacks, the signal is especially likely to point to him, rather than to a different attacker. This makes an easily identified or detected attacker especially suspect when the signal points to him, which makes him an attractive target for false-flag operations.

The above analysis coheres with the descriptive literature on false-flag operations. The Russians, for instance, chose to mimic the North Koreans, who had a pre-existing reputation for aggressiveness in cyber space and might have been particularly suspect following any attack targeting South Korea.

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 Information Warfare Monitor 2009, 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 2010). 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 attribution problem and, especially, concerns about false-flags. 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 (2014) 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. And China is indeed 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 and because signals are particularly likely to point to China (as in Proposition 6). 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.

6 Optimal Deterrence with 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 matters 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.[12] 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 2017). For instance, recent changes to US policy delegate many decisions over cyber retaliation to the commander of US Cyber Command, requiring only minimal consultation with other government agencies (Sanger 2018).

We show that, as one might expect, with commitment the defender retaliates more often after some signals. Interestingly, this always leads *all* attackers to 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. We also characterize which attackers should be the focus of increased retaliation under commitment. After establishing each result, we discuss its implications for contemporary policy debates.

6.1 The Commitment Model

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-

[12] 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 2002).

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 $x_i^*=\sum_s\left(\pi_i^s-\pi_0^s\right)\left(1-G_i\left(y_i^s\right)\right)$, so attacker i attacks with probability $1-F_i\left(\sum_s\left(\pi_i^s-\pi_0^s\right)\left(1-G_i\left(y_i^s\right)\right)\right)$. In the event attacker i attacks, the defender suffers a loss consisting of the sum of several terms (the terms in brackets above). 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

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

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

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

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

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 first-order condition 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 first-order condition.) 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 first-order conditions 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.

Theorem 4 *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 (cf. Definition 1) is shifted up under commitment, due to the defender's additional deterrence benefit from committing to retaliate aggressively.

Theorem 4 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 1960; Snyder 1961; Powell 1990). 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.

The US Department of Defense 2018 Cyber Strategy (Department of Defense 2018) differs from the Obama-era approach articulated in the 2015 Cyber Strategy (Department of Defense 2015) by focussing fairly narrowly on threats from Russia and China, rather than from a broad range of major and minor powers and even non-state actors (see Kollars and Schenieder 2018, for a comparison). One interpretation of the new strategy is that it ranks attackers in terms of ex ante aggressiveness (i.e. the distributions F_i of the benefits of attack) and mainly threatens retaliation against the most aggressiveness attackers. But this misses the main impact of deterrence in influencing *marginal* decisions. The marginal deterrence benefit to the defender from becoming more aggressive against attacker i after signal s is given by the $f_i\left(x_i^*\right)\left(\pi_i^s - \pi_0^s\right) l_i\left(y^*\right)$ term in the equation for y_i^{s*}. This benefit is larger if signal s is more informative that i attacked or if i's aggressiveness is likely to be close to the threshold. It has little to do with i's overall aggressiveness.

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^*.

6.2 Signal Informativeness and Retaliation

Finally, we analyze *which* signals the defender is likely to respond to more aggressively under commitment, relative to the no-commitment equilibrium.

We start with an example showing that the optimal commitment strategy does not necessarily involve retaliating more aggressively after all signals. Suppose 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

$$\frac{\pi_{i^*}^s}{\pi_0^s} \geq \frac{\pi_{i^*}^{s'}}{\pi_0^{s'}}.$$

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 7 *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 $\min \left\{ y_i^s, y_i^{s'}, \tilde{y}_i^s, \tilde{y}_i^{s'} \right\} > 0$. 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 $\tilde{y}_i^{s'} \geq y_i^{s'}$.*

Theorem 4 is in broad agreement with recent arguments calling for more aggressive cyberdeterrence (e.g., Hennessy 2017). 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.

Our analysis can also be contrasted with the suggestion by Clarke and Knake (2010) 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.

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 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. Reducing false alarms also always reduces attacks, as does replacing misattribution with non-detection.

Third, deterrence is unequivocally enhanced by introducing a more destructive weapon to the defender's arsenal, but adding a less destructive weapon can undermine deterrence. Further, attribution problems create incentives for false-flag operations. Consistent with qualitative observations, attackers are more likely to be the victim of false-flag operations if they are themselves more aggressive or if their attacks are easier to detect, identify, or mimic.

Fourth, 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. The defender also should not commit to retaliate more against the most aggressive attackers, but rather against the most deterrable attackers.

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 attack others and face attacks themselves. 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, 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 4. Fix a permutation ρ on I mapping i to j and a corresponding permutation ρ' on $S\backslash\{0\}$. Then

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

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 Theorem 2. Suppose towards a contradiction that $\tilde{p}_i > p_i$ for some i. Let $i \in \operatorname{argmax} \frac{\tilde{p}_i}{p_i}$. Since $\tilde{p}_i > p_i$, we must have $x_i(\tilde{p};\tilde{\pi}) < x_i(p;\pi)$. Combined with the assumption that $x_i(p;\tilde{\pi}) \geq x_i(p;\pi)$, we have $x_i(\tilde{p};\tilde{\pi}) < x_i(p;\tilde{\pi})$. But, for every $s \neq 0$, we have

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

where the inequality follows because $\frac{p_i}{\tilde{p}_i} \leq \frac{p_j}{\tilde{p}_j}$ for all $j \in I$ and $\frac{p_i}{\tilde{p}_i} < 1$. This implies $r_i^s(\tilde{p};\tilde{\pi}) \geq r_i^s(p;\tilde{\pi})$, and hence (since $\tilde{\pi}_i^s \geq \tilde{\pi}_0^s$ for all $s \neq 0$) $x_i(\tilde{p};\tilde{\pi}) \geq x_i(p;\tilde{\pi})$. Contradiction.

The proof of the strict inequality is almost identical: Now $\tilde{p}_i \geq p_i$ implies $x_i(\tilde{p};\tilde{\pi}) \leq x_i(p;\pi)$, which combined with the assumption that $x_i(p;\tilde{\pi}) > x_i(p;\pi)$ again implies $x_i(\tilde{p};\tilde{\pi}) < x_i(p;\tilde{\pi})$. The same argument now gives a contradiction. ■

Proof of Theorem 3. By Theorem 2, it suffices to show that $x_j(p;\tilde{\pi}) \geq x_j(p;\pi)$ for all j. Note that, for all j,

$$\begin{aligned}
x_j(p;\tilde{\pi}) - x_j(p;\pi) &= \sum_{s\neq 0} \left(\tilde{\pi}_j^s - \tilde{\pi}_0^s\right) r_j^s(p;\tilde{\pi}) - \sum_{s\neq 0} \left(\pi_j^s - \pi_0^s\right) r_j^s(p;\pi) \\
&= \left(\tilde{\pi}_j^s - \tilde{\pi}_0^s\right) r_j^s(p;\tilde{\pi}) + \left(\tilde{\pi}_j^{s'} - \tilde{\pi}_0^{s'}\right) r_j^{s'}(p;\tilde{\pi}) \\
&\quad - \left(\pi_j^s - \pi_0^s\right) r_j^s(p;\pi) - \left(\pi_j^{s'} - \pi_0^{s'}\right) r_j^{s'}(p;\pi),
\end{aligned}$$

and $\tilde{\pi}_0^s = \pi_0^s$ and $\tilde{\pi}_0^{s'} = \pi_0^{s'}$.

For $j = i$, note that $\beta_i^s(p;\tilde{\pi}) \leq \beta_i^s(p;\pi)$, and hence $r_i^s(p;\tilde{\pi}) \leq r_i^s(p;\pi) = 0$, so $r_i^s(p;\tilde{\pi}) = 0$. Conversely, $\beta_i^{s'}(p;\tilde{\pi}) \geq \beta_i^s(p;\pi)$, and hence $r_i^{s'}(p;\tilde{\pi}) \geq r_i^{s'}(p;\pi)$. Therefore,

$$\begin{aligned}
x_i(p;\tilde{\pi}) - x_i(p;\pi) &= \left(\tilde{\pi}_i^{s'} - \tilde{\pi}_0^{s'}\right) r_i^{s'}(p;\tilde{\pi}) - \left(\pi_i^{s'} - \pi_0^{s'}\right) r_i^{s'}(p;\pi) \\
&\geq \left(\tilde{\pi}_i^{s'} - \tilde{\pi}_0^{s'} - \pi_i^{s'} + \pi_0^{s'}\right) r_i^{s'}(p;\pi) \\
&\geq 0,
\end{aligned}$$

where the last inequality uses $\tilde{\pi}_i^{s'} > \pi_i^{s'}$ and $\tilde{\pi}_0^{s'} = \pi_0^{s'}$.

For $j \neq i$, note that $\beta_j^s(p;\tilde{\pi}) \geq \beta_j^s(p;\pi)$, and hence $r_j^s(p;\tilde{\pi}) \geq r_j^s(p;\pi)$. Conversely, $\beta_j^{s'}(p;\tilde{\pi}) \leq \beta_j^{s'}(p;\pi)$, and hence $r_j^{s'}(p;\tilde{\pi}) \leq r_j^{s'}(p;\pi) = 0$, so $r_j^{s'}(p;\tilde{\pi}) = 0$. Therefore,

$$\begin{aligned}
x_j(p;\tilde{\pi}) - x_j(p;\pi) &= \left(\tilde{\pi}_j^s - \tilde{\pi}_0^s\right) r_j^s(p;\tilde{\pi}) - \left(\pi_j^s - \pi_0^s\right) r_j^s(p;\pi) \\
&= \left(\pi_j^s - \pi_0^s\right) \left(r_j^s(p;\tilde{\pi}) - r_j^s(p;\pi)\right) \\
&\geq 0,
\end{aligned}$$

where the second equality uses $\tilde{\pi}_j^s = \pi_j^s$ and $\tilde{\pi}_0^s = \pi_0^s$.

For the strict inequality, note that $r_i^{s'} > 0$ implies $p_i > 0$ because, if attacker i never attacks, the defender's payoff from retaliation $y_i - 1$ is negative as $y_i < 1$ with probability one. Next, $p_i > 0$ implies $\beta_i^{s'}(p;\tilde{\pi}) > \beta_i^s(p;\pi)$ as $\tilde{\pi}_i^{s'} > \pi_i^{s'}$. Finally, since G has positive density on its support, $r_i^{s'} > 0$ and $\beta_i^{s'}(p;\tilde{\pi}) > \beta_i^s(p;\pi)$ imply $r_i^{s'}(p;\tilde{\pi}) > r_i^{s'}(p;\pi)$ and hence $x_i(p;\tilde{\pi}) > x_i(p;\pi)$. Similarly, $r_j^s > 0$ implies $p_j > 0$ and $\beta_j^s(p;\tilde{\pi}) > \beta_j^s(p;\pi)$ so $r_j^s(p;\tilde{\pi}) > r_j^s(p;\pi)$ and hence $x_j(p;\tilde{\pi}) > x_j(p;\pi)$. ■

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. 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 Proposition 5. Let $r_i\left(\beta_i^s\right)$ (resp., $\tilde{r}_i\left(\beta_i^s\right)$) denote the expected disutility inflicted on the attacker from the defender's ex post optimal retaliation strategy at belief β_i^s, when the new weapon is unavailable (resp., available). We claim that $r_i\left(\beta_i^s\right) \leq \tilde{r}_i\left(\beta_i^s\right)$ for every β_i^s. To see this, let $\Pr\left(a|A\right)$ denote the probability that the defender retaliates with weapon a given arsenal A, and note that

$$r_i\left(\beta_i^s\right) = \Pr\left(a = o | A = \{0, o\}\right) w^o = w^o - \Pr\left(a = 0 | A = \{0, o\}\right) w^o,$$

while

$$\begin{aligned}\tilde{r}_i\left(\beta_i^s\right) &= \Pr\left(a = o | A = \{0, o, n\}\right) w^o + \Pr\left(a = n | A = \{0, o, n\}\right) w^n \\ &\geq w^o - \Pr\left(a = 0 | A = \{0, o, n\}\right) w^o,\end{aligned}$$

and $\Pr\left(a = 0 | A = \{0, o, n\}\right) \leq \Pr\left(a = 0 | A = \{0, o\}\right)$ by revealed preference.

Now, as in the proof of Proposition 1, for every i we have

$$x_i^* = \sum_{s \neq 0} \left(\pi_i^s - \pi_i^0\right) r_i\left(\beta_i^s\right).$$

Hence, shifting up $r_i(\cdot)$ is analogous to shifting down $G_i(\cdot)$, so by the same argument as in the proof of Proposition 3, this decreases p_i for all i.

It remains to show that $u \leq \tilde{u}$. This can be seen in two steps. First, holding the attack probabilities fixed at $\tilde{p}$, the defender is weakly better off when the new weapon is added to her arsenal (by revealed preference). Next, holding the arsenal fixed, the defender is weakly better off when she best-responds to attack probabilities $\tilde{p}$ rather than p. This follows because, since $K > y^a$, for any fixed retaliation strategy the defender receives a higher payoff when facing $\tilde{p}$ rather than p. Combining these observations, the defender is better off best-responding to $\tilde{p}$ with a larger arsenal rather than best-responding to p with a smaller arsenal. ■

Proof of Claim 5. We check that these strategies form an equilibrium. Note that the defender's posterior beliefs $\left(\beta_i^s\right)$ 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

$$\begin{array}{cc} \beta_0^0 = \frac{9}{10} & \beta_1^0 = \frac{1}{10} \\ \beta_0^1 = \frac{3}{4} & \beta_1^1 = \frac{1}{4} \\ \beta_0^2 = \frac{1}{2} & \beta_1^2 = \frac{1}{2} \end{array}$$

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 Theorem 4. By the defender's FOC with commitment, for all $i \in I$,

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

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

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

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, $p = \lim_{m \to \infty} h^m((1, \ldots, 1)) \geq \lim_{m \to \infty} \tilde{h}^m((1, \ldots, 1)) = \tilde{p}$. ■

Proof of Proposition 7. Under the assumption $\min\left\{y_i^s, y_i^{s'}, \tilde{y}_i^s, \tilde{y}_i^{s'}\right\} > 0$, the defender's FOC is necessary and sufficient for optimality. Under the FOC,

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

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

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

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 4) 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

Abreu, Dilip, David Pearce and Ennio Stacchetti. 1990. "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring." *Econometrica* 58(5):1041–1063.

Acemoglu, Daron and Alexander Wolitzky. 2014. "Cycles of Conflict: An Economic Model." *American Economic Review* 104(4):1350–1367.

Adams, James. 2001. "Virtual Defense." *Foreign Affairs* 80(3):98–112.

Avenhaus, Rudolf, Bernhard von Stengel and Shmuel Zamir. 2002. Inspection Games. In *Handbook of Game Theory with Economic Applications, Volume 3*, ed. Robert Aumann and Sergiu Hart. North-Holland pp. 1947–1987.

Baker, George, Robert Gibbons and Kevin Murphy. 1994. "Subjective Performance Measures in Optimal Incentive Contracts." *Quarterly Journal of Economics* 109(4):1125–1156.

Baliga, Sandeep and Tomas Sjöström. 2004. "Arms Races and Negotiations." *Review of Economic Studies* 71(2):351–369.

Bar-Gill, Oren and Alon Harel. 2001. "Crime Rates and Expected Sanctions: The Economics of Deterrence Revisited." *The Journal of Legal Studies* 30(2):485–501.

Bartholomew, Brian and Juan Andres Guerrero-Saade. 2016. "Wave Your False Flags! Deception Tactics Muddying Attribution in Targeted Attacks." *Virus Bulletin Conference* .

Bassetto, Marco and Christopher Phelan. 2008. "Tax Riots." *Review of Economic Studies* 75(3):649–669.

Berman, Eli, Jacob N. Shapiro and Joseph H. Felter. 2011. "Can Hearts and Minds be Bought? The Economics of Counterinsurgency in Iraq." *Journal of Political Economy* 119(4):766–819.

Blackwell, David. 1951. The Comparison of Experiments. In *Proceedings, Second Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press pp. 93–102.

Bond, Philip and Kathleen Hagerty. 2010. "Preventing Crime Waves." *American Economic Journal: Microeconomics* 2(3):138–159.

Buchanan, Ben. 2017. *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations.*

Chassang, Sylvain and Christian Zehnder. 2016. "Rewards and Punishments: Informal Contracting through Social Preferences." *Theoretical Economics* 11(3):1145–1179.

Chassang, Sylvain and Gerard Padró i Miquel. 2010. "Conflict and Deterrence under Strategic Risk." *Quarterly Journal of Economics* 125(4):1821–1858.

Clark, David D. and Susan Landau. 2010. Untangling Attribution. In *Proceedings of a Workshop on Deterring Cyberattacks: Informing Strategies and Developing Options for U.S. Policy.* Washington, DC: National Academies Press.

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

Crèmer, Jacques. 1995. "Arm's Length Relationships." *Quarterly Journal of Economics* 110(2):275–295.

Department of Defense. 2015. "The DoD Cyber Strategy." Available at: http://archive.defense.gov/home/features/2015/0415_cyber-strategy/final_2015_dod_cyber_strategy_for_web.pdf.

Department of Defense. 2018. "Summary: DoD Cyber Strategy 2018." Available at: https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF.

Di Lonardo, Livio and Scott A. Tyson. 2018. "Political Instability and the Failure of Deterrence." University of Rochester typescript.

Edwards, Benjamin, Alexander Furnas, Stephanie Forrest and Robert Axelrod. 2017. "Strategic Aspects of Cyberattack, Attribution, and Blame." *Proceedings of the National Academy of Sciences* 114(11):2825–2830.

Fearon, James D. 1997. "Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs." *Journal of Conflict Resolution* 41(1):68–90.

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

Ferrer, Rosa. 2010. "Breaking the Law when Others Do: A Model of Law Enforcement with Neighborhood Externalities." *European Economic Review* 54(2):163–180.

Freeman, Scott, Jeffrey Grogger and Jon Sonstelie. 1996. "The Spatial Concentration of Crime." *Journal of Urban Economics* 40(2):216–231.

Glaeser, Edward L., Bruce Sacerdote and Jose A. Scheinkman. 1996. "Crime and Social Interactions." *Quarterly Journal of Economics* 111(2):507–548.

Goldsmith, Jack. 2013. "How Cyber Changes the Laws of War." *European Journal of International Law* 24(1):129–138.

Graetz, Michael J., Jennifer F. Reinganum and Louis L. Wilde. 1986. "The Tax Compliance Game: Toward an Interactive Theory of Law Enforcement." *Journal of Law, Economics and Organization* 2:1–32.

Green, Edward J. and Robert H. Porter. 1984. "Noncooperative Collusion under Imperfect Price Information." *Econometrica* 52(1):87–100.

Hayden, Michael. 2011. "Statement for the Record, House Permanent Select Committee on Intelligence, The Cyber Threat." Available at https://www.hsdl.org/?view&did=689629.

Hennessy, Susan. 2017. "Deterring Cyberattacks: How to Reduce Vulnerability." *Foreign Affairs* November/December.

Hohzaki, Ryusuke. 2007. "An Inspection Game with Multiple Inspectees." *European Journal of Operational Research* 178(3):894–906.

Information Warfare Monitor. 2009. "Tracking GhostNet: Investing a Cyber Espionage Network.".

Jervis, Robert. 1978. "Cooperation Under the Security Dilemma." *World Politics* 30(2):167–214.

Jervis, Robert. 1979. "Deterrence Theory Revisited." *World Politics* 31:289–324.

Kaplan, Fred. 2016. *Dark Territory: The Secret History of Cyber War.* Simon & Schuster.

Kello, Lucas. 2017. *The Virtual Weapon.* New Haven: Yale University Press.

Khalil, Fahad. 1997. "Auditing without Commitment." *RAND Journal of Economics* 28(4):629–640.

Kollars, Nina and Jacquelyn Schenieder. 2018. "Defending Forward: The 2018 Cyber Strategy is Here." *War on the Rocks* September 20.

Kydd, Andrew. 1997. "Game Theory and the Spiral Model." *World Politics* 49(3):371–400.

Lando, Henrik. 2006. "Does Wrongful Conviction Lower Deterrence?" *Journal of Legal Studies* 35(2):327–337.

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

Lindsay, Jon R. 2015. "Tipping the Scales: The Attribution Problem and the Feasibility of Deterrence Against Cyberattack." 1(1):53–67.

McDermott, Rose, Anthony C. Lopez and Peter K. Hatemi. 2017. "'Blunt Not the Heart, Enrage It': The Psychology of Revenge and Deterrence." *Texas National Security Review* 1(1):69–89.

Milgrom, Paul and John Roberts. 1994. "Comparing Equilibria." *American Economic Review* 84(3):441–459.

Mookherjee, Dilip and Ivan Png. 1989. "Optimal Auditing, Insurance, and Redistribution." 104(2):399–415.

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

Nakashima, Ellen. 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.

Nye, Jr., Joseph S. 2011. "Nuclear Lessons for Cyber Security?" *Strategic Studies Quarterly* 5(4):18–38.

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

Png, Ivan. 1986. "Optimal Subsidies and Damages in the Presence of Judicial Error." *International Review of Law and Economics* 6(1):101–105.

Polinsky, A. Mitchell and Steven Shavell. 2000. "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature* 38(1):45–76.

Powell, Robert. 1990. *Nuclear Deterrence Theory: The Search for Credibility.* Cambridge University Press.

Radner, Roy. 1986. "Repeated Principal-Agent Games with Discounting." *Econometrica* 53(5):1173–1198.

Rogin, Josh. 2010. "The Top 10 Chinese Cyber Attacks (That We Know Of)." *Foreign Policy* January 22.

Rohner, Dominic, Mathias Thoenig and Fabrizio Zilibotti. 2013. "War Signals: A Theory of Trade, Trust and Conflict." *Review of Economic Studies* 80(3):1114–1147.

Sah, Raaj K. 1991. "Social Osmosis and Patterns of Crime." *Journal of Political Economy* 99:1272–1295.

Sanger, David. 2018. "Trump Loosens Secretive Restraints on Ordering Cyberattacks." *New York Times* September 20.

Sanger, David and William Broad. 2018. "Pentagon Suggests Countering Devastating Cyberattacks With Nuclear Arms." *New York Times* January 16.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge: Harvard University Press.

Schrag, Joel and Suzanne Scotchmer. 1997. "The Self-Reinforcing Nature of Crime." *International Review of Law and Economics* 17(3):325–335.

Segerson, Kathleen. 1988. "Uncertainty and Incentives for Nonpoint Pollution Control." *Journal of Environmental Economics and Management* 15(1):87–98.

Shachtman, Noah. 2010. "Insiders Doubt 2008 Pentagon Hack was Foreign Spy Attack." *Wired* August 25.

Shavell, Steven. 1985. "Uncertainty Over Causation and the Determination of Civil Liability." *Journal of Law and Economics* 28(3):587–609.

Shaver, Andrew and Jacob N. Shapiro. Forthcoming. "The Effect of Civilian Casualties on Wartime Informing: Evidence from the Iraq War." *Journal of Conflict Resolution* .

Shevchenko, Vitaly. 2014. "'Little Green Men' or 'Russian Invaders'?" *BBC* March 11.

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

Smith, Alastair. 1998. "International Crises and Domestic Politics." *American Political Science Review* 92(3):623–638.

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

Sullivan, Eileen, Noah Weiland and Kate Conger. 2018. "Attempted Hacking of Voter Database Was a False Alarm, Democratic Party Says." *New York Times* August 23.

ThreatConnect. 2016. "Guccifer 2.0: All Roads Lead to Russia." Available at `https://threatconnect.com/blog/guccifer-2-all-roads-lead-russia/`.

Trager, Robert F. and Dessislava P. Zagorcheva. 2006. "Deterring Terrorism: It Can Be Done." *International Security* 30(3):87–123.

Tsebelis, George. 1989. "The Abuse of Probability In Political Analysis: The Robinson Crusoe Fallacy." *American Political Science Review* 83(1):77–91.

United States. 2018. "National Cyber Strategy." Available at: `https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf`.

Weissing, Franz J. and Elinor Ostrom. 1991. Irrigation Institutions and the Games Irrigators Play: Rule Enforcement without Guards. In *Game Equilibrium Models II: Methods, Morals, and Markets*, ed. Reinhard Selten. Springer-Verlag pp. 188–262.