Strategic Cyber Warfare

Erik Lillethun* Rishi Sharma*

July 18, 2022

Abstract

Cyber warfare has been rising in prominence as a form of international conflict in recent years. In this paper, we develop a game theory model of cyber warfare between two nation states: the Attacker and the Defender. The Attacker decides when to infiltrate one or more systems belonging to the Defender, and the Defender decides when and in what systems to monitor for infiltrators and clear them out. We analyze Markov perfect equilibria, and find that in the single system setting, the players employ mixed strategies with full support, except for extreme parameter ranges. We explicitly solve for this equilibrium, which is never Pareto efficient. In the multiple systems setting, the Attacker's infiltration costs may depend on which systems they have already infiltrated. This may arise because of networked systems, compromised multiple login credentials, or systems containing technical information that aids in infiltration. We characterize an "infiltrate all systems" equilibrium and find existence conditions. In the example where cost relationships are defined by a cluster graph, we show that the Attacker must infiltrate at most one cluster at a time, and characterize an equilibrium where the Attacker mixes over which cluster to infiltrate. In each of these models, we provide comparative statics which suggest ways in which nations can invest to improve outcomes. Finally, we apply this model to efficient non-Markovian cooperative equilibria and the problem of externalities that arise when a nation's cyber defense is controlled by private entities.

JEL: C72, D74, D85, F5

*Department of Economics, Colgate University.

1 Introduction

Cyberwarfare involving nation states is commonly described as "the next frontier for warfare," one in which countries are continuously engaged in defending their information systems while also attacking rival nations. The U.S., for example, has in recent years come under numerous notable attacks, including the 2015 Chinese attack on the Office of Personnel Management and the 2016 Russian attack on the Democratic National Committee. The U.S. is also engaged in offensive operations, with the joint operations with Israel sabotaging Iran's nuclear program in 2010 being one of the first salient examples of cyberattacks by nation states. Cyberattacks can also have substantial effects on civilian populations beyond the exposure of information as in the 2015 Russian attack on the Ukrainian power grid that led to power outages affecting over 200,000 people. This new mode of warfare raises many questions: What determines the severity of cyberwarfare? How are outcomes affected by improved offensive or defensive capacities or by the nature of network structures? What are the consequences of the private vs. public provision of cyberdefence capacity? Under what conditions are (possibly informal) agreements between countries limiting cyberattacks more or less likely? Is there hope for certain types of attacks being considered off-limits?

In this paper, we study these questions by developing a theoretical model of cyberwarfare between nation states. We emphasize the development of a framework that is tractable but rich enough to capture a range of different scenarios and one that distinctively models cyberwarfare as opposed to broader war or competitive resource expenditure. To this end, we construct a dynamic continuous time two-country model where each country can attack and defend an arbitrary number of systems. The Attacker chooses which systems to infiltrate and when, taking into account both the costs and benefits of infiltration. These systems may be interlinked as part of clusters where access to one system in a cluster can affect the costs and benefits of infiltrating additional systems. The benefits of infiltration include both a continuous stealthy stream of information as well a potential option value benefit that can be triggered by an exogeneous shock. The Defender – which incurs damages corresponding to both of these benefits for the Attacker – does not know whether the Attacker is currently infiltrating a particular system or not and therefore has to engage in costly monitoring activities to periodically clear out potential Attackers. These costly monitoring activities are meant to capture not only direct monetary costs but also factors that affect the functioning of the information system, such as the need to take systems offline or for users to follow more restrictive security procedures.

Given the Attacker's need to be unpredictable with respect to its infiltration timing and the Defender's need to be unpredictable about its monitoring behavior, the typical equilibrium will involve both the Attacker and Defender employing strategies that are mixed across time. It is useful in this context to think about how "severe" a particular equilibrium is. For example, an equilibrium where the Attacker patiently spreads the likelihood of infiltration over a

longer period and the Defender clears less frequently in expectation will be associated with relatively low cyber operations for both the Attacker and Defender as compared to a more severe equilibrium where the Attacker is more aggressive and the Defender clears more frequently. A key intuition that emerges from the mixed strategy equilibria is that the Attacker's equilibrium aggressiveness depends on the margin on the Defender's costs and benefits from clearing whereas the Defender's equilibrium proactiveness depends on the costs and benefits for the Attacker.

We find several factors that affect the severity of the equilibrium. Lower infiltration costs for the Attacker – which would capture improved offensive ability – lead to a more severe equilibrium. The Defender has to clear more frequently in expectation in order to disincentivize the Attacker, who would otherwise have an incentive to be more aggressive. The costs of improvements in offensive cyber abilities therefore may not show up in terms of increased damage but may instead show up in the increasing burden of defensive countermeasures. As noted above, this burden could in practice include both monetary costs and broader operational costs to users of information systems. Lower clearing costs for the Defender – which would capture improved defensive ability – would conversely lead to a less severe equilibrium. This is because the Attacker has to reduce its aggressiveness in order to disincentivize the Defender from clearing too frequently. Improvements in defensive ability therefore lead to lower overall costs of cyberwarfare and improved welfare.

Our model also allows us to examine how network structures can affect cyberwarfare outcomes. In particular, the order of infiltrating systems may matter to the Attacker, as earlier infiltrations reduce the cost of later infiltrations. We characterize a class of equilibria where the Attacker infiltrates a set of systems with certainty. Cost reductions allow for broader (more systems) attacks in equilibrium, and in particular, infiltration of all systems is always possible with sufficiently extreme cost reductions. In an example where cost reductions are defined by a cluster graph, we show that at most a single cluster can be infiltrated at one time, so breaking systems into smaller clusters will deter broader attacks.

We also address an externality problem that is especially common in liberal democracies with powerful private sectors. Private entities may be in charge of the cyber defenses of their own systems. We examine a variation on our model with an initial optional investment in making infiltration prohibitively costly. When systems have an upstream-downstream structure (infiltration of the former reducing the infiltration costs for the latter), and investment in the upstream system is pivotal in preventing infiltration in the downstream one, then the upstream system may not invest, even when it is beneficial for the Defenders as a whole. This tends to be a problem when downstream system infiltration losses are large relative to upstream system losses. A similar reasoning should apply to an upstream system with many downstream systems, such as a popular platforms or services, as was the case with the 2021 Microsoft Exchange Server hack. These externality considerations would lend support to policy initiatives that impose cybersecurity requirements or standards on private entities

or promise retaliation when private entities are attacked, casting doubt on the idea of a strict separation between public and private cyberdefense strategies.

While we talk about the "Attacker" and "Defender" with respect to a particular set of systems, a signature feature of cyberwarfare between countries is that typically both sides are engaged in attacking as well as defending. This raises the possibility of agreements – possibly informal – between countries that reduce the likelihood or severity of cyberwarfare. We find that a "complete" agreement is mostly likely when latent cybersecurity monitoring is effective, costs of both attacking and defending are high, and infiltration losses are large relative to gains. We also consider potential partial agreements where countries agree to avoid at least certain systems. We find that that this is the only type of agreement possible in an asymmetric situation where only one nation is capable of attacking, and this type of agreement tends to work when the infiltration benefit/loss ratios differ substantially across systems. High loss, low benefit systems would be off-limits.

Broadly, our work contributes to an existing theoretical literature studying warfare within an economic framework – a literature that goes back to Haavelmo [1954] and Schelling [1960] – by developing a theory of cyberwarfare.[1] Much of the emphasis in this literature is on the factors that make wars more or less likely to take place (e.g. Powell [1993], Yared [2010], Acemoglu et al. [2012], and Acemoglu and Wolitzky [2014]). Cyberwarfare is very different from "regular" warfare in this respect since it is continuous and unceasing, and there is no meaningful alternation between states of peace and war. Hence, in analyzing cyberwarfare, our emphasis shifts to a more continuous notion of "severity" rather than discrete questions about the presence or absence of conflict.

Another literature related to ours is focused on industrial espionage (e.g. Whitney and Gaisford [1999], Solan and Yariv [2004], Ho [2008], Barrachina et al. [2014]). This literature generally emphasizes costly information acquisition through espionage in the context of competition between firms that may lead, for example, to production cost reductions for the Attacker. The emphasis on production and market competition is quite different from our setting, where the direct damages from cyberwarfare are the predominant consideration. Furthermore, neither the warfare nor the espionage literatures focus on the the networked structures (and eventually the related externalities) that are more central to our analysis.

While there is a substantial economics literature on related topics such as warfare in general and espionage, there is practically no work on cyberwarfare in particular. One exception is Baliga et al. [2020], who study the multiple attribution problem in the context of cyberwarfare. Our work focuses on very distinct aspects of cyberwarfare to theirs and in this sense is complementary to their work. We focus on general questions related to cyberwarfare (e.g. factors determining severity and welfare consequences, role of network structures, and externalities) but abstract from the attribution problem that is central to Baliga et al. [2020] by focusing on a two-country analysis.

[1] See Garfinkel and Skaperdas [2007] for a review of the conflict literature.

In discussing potential agreements to limit cyberwarfare, our work also connects to the literature on international or interjurisdictional coordination motivated by the desire to limit harmful competition. This is especially prominent in the context of trade agreements and the trade wars they are designed to prevent (e.g. Grossman and Helpman [1995], Bagwell and Staiger [1999], Ossa [2014]). Similar considerations are present when considering international tax agreements (e.g. Keen and Konrad [2013]) or environmental agreements (e.g. Oates and Portney [2003]). None of these other settings feature the substantially unobservable nature of the underlying competition together with the network structures that are central to our cyberwar context.

Outside of the economics literature, our work connects to a large literature in computer science and information security studies that uses game theoretic models to analyze cyber attacks.[2] This literature is engineering focused. The emphasis is on using game theory to identify strategies that can improve cyber security in the context of specific types of attacks. By contrast, we approach cyberwarfare from a social science standpoint, especially understanding the factors that could improve or worsen the social costs of cyberwarfare and the potential for international agreements in this space. Our actual model is also very different from the existing work outside of economics, particularly with the emphasis on mixed strategies over time and multiple interlinked systems.

The paper proceeds as follows: Section 2 introduces our basic two country, single system model and characterizes the Markovian equilibria. Section 3 extends the model to multiple systems and provides existence conditions for equilibria without mixing across systems. Section 4 considers special examples of infiltration cost relationships between systems, making the existence conditions of Section 3 more explicit and allowing for further analysis. Section 5 characterizes certain non-Markovian equilibria where complete or partial cyber peace agreements are possible. Section 6 analyzes a positive externality problem that is possible when private entities control defensive choices, and systems have an upstream-downstream relationship. Finally, Section 7 concludes. Proofs for results presented without proof in the main text may be found in the Appendix.

2 Single System

2.1 Model

There are two players: and Attacker (A) and a Defender (D), who play a game over time $t \in [0, \infty)$. The players' payoffs depend on a state $\omega \in \{I, N\}$ ("infiltrated" or "not infiltrated") pertaining to a single system (i.e., the number of systems is $n = 1$). $\omega(t)$ is the state at the start of time t, with $\omega(0) = N$. The Attacker always knows the current state. The Defender does not, but will occasionally receive a signal informing them that the state is I.

At times t when the state has switched from $\omega = I$ to $\omega = N$ (a "clearing") in the last $\Delta > 0$ units of time, the game is on "pause." Neither the Attacker

[2]See, for example, Roy et al. [2010] or Merrick et al. [2016] for surveys of the literature.

nor the Defender may take any action. This pause represents several realistic possibilities. In part, it may reflect the Defender taking their system offline to fix vulnerabilities in their system. Once the system is back online (and may be attacked anew), the pause may reflect a minimum delay while the Attacker finds a new vulnerability in the system. In either case, infiltration is impossible and explicit monitoring is pointless, so there is a pause. Despite the name, time still progresses as normal during the pause.

At times t when $\omega(t) = N$ and the game is not on pause, the Attacker can choose whether or not to infiltrate the Defender's system. If the Attacker infiltrates, the state immediately switches to $\omega = I$. In principle, strategies could depend on the entire history of play. However, for most of the paper, we look exclusively at Nash equilibria where the strategies are functions of the time τ since $\omega = I$ most recently became possible (at $t = 0$ or after the last pause).[3] Thus, the Attacker's strategy is $A : [0, \infty) \to [0, 1]$, where $A(\tau)$ is the probability that the next infiltration occurs at a time $\leq \tau$ since $\omega = I$ most recently became possible. Note that this need not be a cumulative distribution function (CDF), because the Attacker may decide to never infiltrate ($A(\tau) \nrightarrow 1$ as $\tau \to \infty$). Each infiltration costs the Attacker $C_A > 0$ at the time the state changes from N to I.

At any time t that is not during a pause, the Defender can take a deliberate monitoring action checking their system for infiltrators and clearing them if found (i.e., if $\omega(t) = I$). We assume that the Attacker always observes this action, so the Attacker always knows what the Defender knows; we only look at one-sided asymmetric information. Clearing takes effect (switches the state to $\omega = N$) immediately. The Defender's strategy also depends only on the time τ since $\omega = I$ last became possible. Thus, it is a function $D : [0, \infty) \to [0, 1]$, where $D(\tau)$ is the probability that the next checking time is $\leq \tau$ after $\omega = I$ last became possible. Specifically, $\omega = I$ becomes possible Δ units of time after the system was most recently cleared. Again, this strategy need not be a CDF; the Defender could decide to never actively check the system. Note that if the state is currently N, then the action has no effect; we call this "checking" but not clearing. Note that checking alone does not initiate a pause, since there is no state change. Clearing has a cost of $C_D > 0$ at the time the state switches from I to N.[4] Checking without clearing also has cost C_D.

The Defender also may observe a signal. The next signal arrival time is exponentially distributed with constant rate parameter $\lambda > 0$. The signal that arrives at time t is observed and has an effect if and only if $\omega(t) = I$. When the signal arrives and $\omega(t) = I$, the Attacker receives an expected stock benefit of size B and the Defender suffers an expected stock loss of size L. When $B, L > 0$, a signal arrival may be interpreted as an opportunity for the Attacker to gain a

[3] Although we do not assume sequential rationality within the "stage game" played between each time transition from N to I becomes possible, in practice these equilibria will all be Markov perfect equilibria. This is because neither party gets signals about the other's action until the end of the stage game.

[4] The cost C_D may include not only the costs of removing the infiltrators and patching the system but also the cost of having the system offline.

strategic advantage over the Defender by damaging the system, but in doing so, the Attacker reveals that they have infiltrated the system. When $B = L = 0$, a signal arrival is like a Defender whose default cybersecurity monitoring happens to notice an infiltrator. The signal may also reflect a combination of the two: there are multiple signals, λ is the aggregate signal arrival rate, and B, L are the expected stock benefits and losses conditional on a signal arriving (not a particular signal). A signal arrival when $\omega(t) = I$ may or may not result in immediate clearing by the Defender. We assume that if the Defender adopts a strategy where they may eventually explicitly clear (i.e., $\exists \tau \geq 0$ such that $D(\tau) > 0$), then they immediately clear following a signal arrival, and if they never explicitly clear (i.e., $D(\tau) = 0, \forall \tau \geq 0$), then they never clear following a signal arrival. This assumption is for convenience; it only matters in knife-edge cases.

The timeline for a single instant of time t is as follows: First, the Attacker may infiltrate, and the state switches to $\omega = I$. Then, the Defender may check or clear, setting the state to $\omega = N$. Finally, if $\omega = I$ and there is a signal arrival at time t, the expected stock benefits and losses are realized, and the Defender may clear.[5]

The Attacker and Defender share the same discount rate $r > 0$. The Attacker reaps a flow benefit of $b > 0$ while the state is $\omega = I$ (0 while $\omega = N$) and pays a cost of C_A at the time of infiltration. The Defender suffers a flow loss of $\ell > 0$ while the state is $\omega = I$ (0 while $\omega = N$) and pays the cost of C_D at the time of checking. The Attacker and Defender both maximize their discounted expected payoffs, using the flow and stock benefits, losses, and costs just defined. The full definition of the payoff functions is cumbersome, so it has been relegated to Appendix A.

2.2 Pure Strategy Equilibrium

For extreme parameter values, there will be simple pure strategy equilibria. In some instances, it may not be worth infiltrating the system, even if the Defender never clears it:

Condition 1

$$rC_A \geq b + \lambda B$$

Proposition 1
In the single system model ($n = 1$), if Condition 1 holds, then there is a pure strategy equilibrium $A(\tau) = 0, D(\tau) = 0$ (i.e., the Attacker never infiltrates, and the Defender never checks).

Condition 1 may apply when the system is not valuable to the Attacker. For example, if two nations have peaceful relations and already actively share the information contained in their systems, there is no point in engaging in

[5]The timeline is specified for technical reasons only. Unlike an analogous discrete time model, the timeline at a single moment or period will not be consequential.

cyberattacks against one another ($b+\lambda B \approx 0$). It may also apply when either the Attacker has made little investment in cyberattack capabilities compared to the Defender's investments in cyberdefense capabilities (i.e., C_A is large). However, this latter explanation seems unlikely to apply to cyberwarfare (as opposed to conventional espionage), which seems to favor the Attacker, as evidenced by the successes of small cyber criminal organizations and even individual hackers.

In other circumstances, the Defender may not feel it is worth clearing their system, even though they know that the Attacker always infiltrates as soon as possible:

Condition 2

$$rC_D \geq (1 - e^{-r\Delta})(\ell + \lambda L)$$

Proposition 2
In the single system model ($n = 1$), if Condition 1 does not hold strictly and Condition 2 holds, then there is a pure strategy equilibrium $A(\tau) = 1, D(\tau) = 0$ (i.e., the Attacker always immediately infiltrates, and the Defender never checks).

Condition 2 tends to hold when checking and clearing is costly relative to the harm done to the Defender as a result of infiltration. This could result if the system is very heavily used (and thus downtime for clearing is very costly), and the Attacker is not a major adversary. It could also occur if the Attacker is very quick at finding new vulnerabilities for reinfiltration (i.e., Δ is small), causing the Defender to be overwhelmed.

On the whole, the conditions for pure strategy equilibria seem unlikely, at least where both parties are major adversaries. Moreover, these equilibria would generate no publicity, as they involve either no attack or no public reaction to attacks. Instead, the many newsworthy instances of cyber attacks must result from a mixed strategy equilibrium.

2.3 Mixed Strategy Equilibrium

In a game such as this one, the opposing interests of the two players very intuitively lead to mixed strategies being played. If the Attacker always immediately infiltrated, the Defender would want to always immediately clear in response, making the attack unprofitable. If the Attacker never infiltrated, then the Defender would never find it worthwhile to explicitly check the system, making attacks very profitable. In either of these non-equilibrium situations, if the Attacker makes incremental adjustments in the direction of higher payoffs, with the Defender reacting to each adjustment by best responding, then this process would adjust towards the mixed strategy equilibrium of the following proposition:

Proposition 3
In the single system model ($n = 1$), if neither Condition 1 nor Condition 2

hold strictly, there exists a the following full support (i.e. on $\tau \in [0, \infty)$) mixed strategy equilibrium:

$$A(\tau) = A + (1 - A)\left(1 - e^{-\lambda A \tau}\right),$$
$$\textit{where } A \equiv \sqrt{\frac{rC_D}{(1 - e^{-r\Delta})(\ell + \lambda L)}},$$
$$D(\tau) = 1 - e^{-\rho\tau},$$
$$\textit{where } \rho \equiv \frac{b + \lambda B - (r + \lambda)C_A}{C_A}$$

Proof. The Defender is approximately indifferent across all checking times $\tau \in [0, \infty)$ when for all τ,

$$\begin{aligned}
&[p(\tau)e^{-r\Delta} + [1 - p(\tau)]]V_D(N) - C_D \approx p(\tau) \int_0^{d\tau} \{(1 - e^{-rt})[-\frac{\ell}{r}] \\
&+ e^{-rt}[-L - C_D + e^{-r\Delta}V_D(N)]\}\lambda e^{-\lambda t} \cdot dt \\
&+ (1 - e^{-\lambda \cdot d\tau})[1 - p(\tau)]e^{-r \cdot d\tau}[V_D(N) - C_D] \\
&+ e^{-\lambda \cdot d\tau}\{(1 - e^{-r \cdot d\tau})[-p(\tau)\frac{\ell}{r}] + e^{-r \cdot d\tau}[[p(\tau)e^{-r\Delta} + [1 - p(\tau)]]V_D(N) - C_D]\},
\end{aligned}$$

where $V_D(N)$ is the Defender's expected continuation payoff immediately after infiltration first becomes possible, and $p(\tau)$ is the Defender's belief that the current state is I. In other words, the Defender is always indifferent between checking immediately and checking in the next moment ($d\tau$ units of time later). Note that this equation ignores the event that infiltration occurs in the next $d\tau$ units of time, because that is a second-order effect (it is an event of vanishing probability, and affects payoffs for only a vanishing period of time). The same reasoning cannot be applied to the event that a signal arrives in the next $d\tau$ units of time, because this causes a *stock* loss $-L$ to the Defender. Taking the first-order Taylor approximation, rearranging, dividing by $d\tau$, and taking the limit as $d\tau \to 0$ yields the exact indifference condition

$$r\left[[p(\tau)e^{-r\Delta} + [1 - p(\tau)]]V_D(N) - C_D\right] = -p(\tau)(\ell + \lambda L) \tag{1}$$

Thus, the target beliefs $p(\tau)$ must be constant.

Note that we must have $A(0) > 0$, because indifference at $\tau = 0$ requires $p(\tau) > 0$. For convenience, we will just write A instead of $A(0)$ in places where confusion is unlikely. $A^+(\tau) = \frac{A(\tau) - A}{1 - A}, \tau > 0$ is the probability of infiltrating no later than τ, conditional on not infiltrating at $\tau = 0$. Lemma 1 (proved in the Appendix) shows that the Attacker's strategy must have an exponential form:

Lemma 1
In the mixed strategy equilibrium, $A^+(\tau) = 1 - e^{-\lambda A \tau}$.

For any A, this strategy yields a constant updated probability of $\omega = I$, which is equal to A. Note that $A^+(\tau)$ does turn out to be a CDF, so the

Attacker's strategy always results in infiltration in finite time. If A also satisfies the indifference condition Equation (1), then this Attacker strategy makes the Defender indifferent about checking at all times when the Attacker might be in the system. Thus, $V_D(N)$ is also their payoff if they check immediately, which yields

$$\begin{aligned} & V_D(N) = [Ae^{-r\Delta} + (1-A)]V_D(N) - C_D \\ \Leftrightarrow & V_D(N) = -\frac{C_D}{A(1-e^{-r\Delta})} \\ \Rightarrow & -\frac{rC_D}{A(1-e^{-r\Delta})} = -A(\ell + \lambda L) \\ \Leftrightarrow & A = \sqrt{\frac{rC_D}{(1-e^{-r\Delta})(\ell+\lambda L)}} \end{aligned}$$

Because indifference requires $A \leq 1$, Defender indifference is achievable if and only if Condition 2 does not hold strictly (i.e., checking is worthwhile even if the Attacker reinfiltrates at the earliest opportunity).

To justify this Attacker strategy as a best response, the Attacker must be indifferent between infiltrating as soon as possible and waiting one extra moment to infiltrate, at all times τ since infiltration most recently became possible, conditional on checking not taking place in the last τ units of time. Let $V_A(\tau)$ be the Attacker's continuation payoff conditional on state N from initiating infiltration when τ units of time have passed since it became possible to begin infiltrating (at $t = 0$, immediately after checking, or Δ after the most recent clearing). This must be a constant function. Indifference between infiltrating as soon as possible and infiltrating $d\tau$ units of time later is equivalent to

$$\begin{aligned} V_A(\tau) = & \frac{D(\tau + d\tau) - D(\tau)}{1 - D(\tau)} \mathbb{E}_D[e^{-r(t-\tau+\Delta)} | \tau \leq t < \tau + d\tau] V_A(\tau) \\ & + \frac{1 - D(\tau + d\tau)}{1 - D(\tau)} e^{-r \cdot d\tau} V_A(\tau + d\tau) \end{aligned}$$

Taking the first-order Taylor approximation of the right hand side, rearranging, and taking the limit as $d\tau \to 0$ gives the exact indifference condition:

$$\begin{aligned} & rV_A(\tau) = e^{-r\Delta} h(\tau) V_A(\tau) - h(\tau) V_A(\tau) + \lim_{d\tau \to 0} \frac{V_A(\tau + d\tau) - V_A(\tau)}{d\tau} \\ \Leftrightarrow & \lim_{d\tau \to 0} \frac{V_A(\tau + d\tau) - V_A(\tau)}{d\tau} = [r + (1 - e^{-r\Delta}) h(\tau)] V_A(\tau), \end{aligned}$$

where $h(\tau)$ is the hazard function for the Defender's strategy. The RHS is bounded below by $rV_A(\tau)$, because the hazard function takes nonnegative values. $V_A(\tau) > 0$ implies that $V_A(\tau)$ is unbounded above, which is impossible as $\frac{b+\lambda B}{r}$ is an upper bound. Thus, $V_A(\tau) = 0$. This argument applies to any τ, so $V_A(\tau) = 0, \forall \tau$.

The following Lemma (proved in the Appendix) derives a Defender strategy consistent with the Attacker's value function:

Lemma 2
Suppose that $D(\tau)$ is twice continuously differentiable. Then $V(\tau) = 0, \forall \tau \Leftrightarrow D(\tau) = 1 - e^{-\rho\tau}$, where $\rho = \frac{b+\lambda B-(r+\lambda)C_A}{C_A}$.

We must have $\rho \geq 0$, so the Attacker may be made indifferent if and only if Condition 1 does not hold strictly.

□

First, it is worth noting that since this is a mixed strategy equilibrium, each player's equilibrium strategy is determined by the parameters that influence the other player's propensity to act. The Attacker's "infiltration intensity" A (which makes the Defender more inclined to check/clear) is a decreasing function of the Defender's propensity to check/clear, represented by the benefit/cost ratio $\frac{(1-e^{-r\Delta})(\ell+\lambda L)}{rC_D}$. The Defender's clearing rate ρ (which makes the Attacker less inclined to infiltrate) is an increasing function of the Attacker's propensity to infiltrate, represented by the net benefit/cost ratio $\frac{b+\lambda B-(r+\lambda)C_A}{C_A}$.

Next, let us consider some of the comparative statics on strategies. Increasing the discount rate r makes both players less inclined to take action, since they pay costs up front and reap the benefits later, so the equilibrium strategies adjust by increasing A and decreasing ρ. Increasing ℓ, L, or Δ, or decreasing C_D all make the Defender more inclined to clear, and the Attacker adjusts by decreasing A. Increasing b or B, or decreasing C_A all make the Attacker more inclined to infiltrate, and the Defender adjusts by increasing ρ. The comparative statics with respect to signal rate λ are more subtle. The effect of increasing λ on A is ambiguous. On the one hand, the rate at which the Defender suffers losses L is increased, so they are more inclined to check/clear, so the infiltration intensity A falls. On the other hand, the absence of signal arrivals is more informative, so the infiltration rate λA must rise. The effect of λ on clearing rate ρ depends on whether signal arrivals are good or bad for the Attacker. If $B > C_A$, signal arrivals are good, because the benefit received outweighs the cost of reinfiltration, so increasing λ encourages infiltration, and ρ must rise to compensate. If $B < C_A$, then signal arrivals are bad for the Attacker, so ρ must fall.

For welfare effects, the Attacker's payoff is always 0 regardless of parameter values. Any positive payoff makes the Attacker strictly prefer infiltrating immediately, which the Defender would respond to by clearing immediately, erasing that positive payoff. The Defender's payoff (and the aggregate payoff) is $V_D(N) = -A\frac{\ell+\lambda L}{r} = -\sqrt{\frac{C_D(\ell+\lambda L)}{r(1-e^{-r\Delta})}}$. Intuitively, increasing clearing costs or infiltration losses is bad for the Defender. Increasing the discount rate is good for the Defender but only in the sense that they value their future costs and losses less. Increasing the duration of the pause period is good for the Defender, because both clearing costs and infiltration losses are avoided during the pause period. In the aggregate, welfare is negative, so this equilibrium is worse than a state of cyber peace ($A(\tau) = 0, D(\tau) = 0$). Welfare is determined entirely by the Defender's characteristics, so there is a Pareto improvement if the Defender

gets more efficient at detecting infiltrators (C_D falls) or gets more effective at closing security vulnerabilities (Δ rises). This is an argument in favor of defensive investments, which has been suggested elsewhere ("The first step is to recognize the folly of going on offense unless we have a good defense," on pg. 325 of Sanger [2019]), though the argument does not apply in the extremely asymmetric case where the Attacker has capabilities far exceeding that of the Defender (the $A(\tau) = 1, D(\tau) = 0$ equilibrium is likely to result).

As a final note, it is worth explaining why A is a concave function of the Defender's clearing cost/benefit ratio. Increasing A, the probability that the Defender has already infiltrated, has two effects on the Defender's propensity to check the system. First, if the Attacker is actually in the system, checking and clearing increases the Defender's flow payoffs from $-(\ell + \lambda L)$ (their lowest possible value) up to the equilibrium expected flow payoffs $rV_D(N) \in (-(\ell + \lambda L), 0)$. Second, if the Attacker is caught, it initiates the pause period during which the Attacker cannot infiltrate, yielding an even higher temporary flow payoff of 0. Intuitively, this is an information effect. If the Attacker is caught in the system, it informs the Defender of the security vulnerability that the Attacker exploited, so they can fix the vulnerability and make it harder for the Attacker to reinfiltrate. On the other hand, if the Defender checks but does not find any infiltrators, they do not receive any information about vulnerabilities. If instead checking without clearing did result in a pause, then the equilibrium would have $A = \frac{rC_D}{(1-e^{-r\Delta})(\ell+\lambda L)}$.

2.4 Equilibrium Uniqueness

The previous section established conditions under which a particular full support mixed strategy equilibrium exists. If either of these conditions fails, there will be a pure strategy equilibrium where either the Attacker always infiltrates immediately, or the Attacker never infiltrates (and in both cases, the Defender never clears). We now pin down the equilibrium set even further when these conditions hold. We prove that equilibrium cannot feature mass points, except for the Attacker's mass point at $\tau = 0$, and any mixed strategy equilibrium must have full support strategies.

The argument against mass points is intuitive. First, note that mass points in one's strategy (on the equilibrium path of play) are too easy for the other side to exploit. Suppose the Attacker had a mass point at infiltration time $\tau^* > 0$. Then, if the Defender were indifferent immediately before τ^*, they must strictly prefer clearing immediately at time τ^*. However, then the Attacker would gain nothing from infiltration at time τ^*, so the Attacker is not best responding. If the Defender were indifferent immediately after τ^*, then they must strictly prefer not checking at any time before τ^*. However, then the Attacker is not best responding, because the expected gains from infiltration must be even greater before τ^* than at τ^*. In the third case, the Defender is not indifferent either before or after τ^*, and the previous argument still applies. An analogous intuition applies to a Defender mass point. This is made formal below:

Proposition 4
In the single system model ($n = 1$), the equilibrium features no interior mass points: For any $\tau^* > 0$, $A(\tau^*) - \lim_{\epsilon \to 0^+} A(\tau^* - \epsilon) = 0$ and $D(\tau^*) - \lim_{\epsilon \to 0^+} D(\tau^* - \epsilon) = 0$. Furthermore, the Defender's strategy has no mass point at 0: $D(0) = 0$.

The proof of Proposition 3 had assumed that there were no interior mass points in deriving a full support mixed strategy equilibria, but Proposition 4 proves this and rules out other pure strategy equilibria and mass points in non-full-support mixed strategy equilibria.

Now, suppose that there are no mass points, but that one party's strategy may not have full support on the equilibrium path of play. Suppose that some interval of time (τ_1, τ_2) is not in the support of the Attacker's strategy. Then, beliefs must fall over this interval of time, as signals fail to arrive. One possibility is that $D(\tau_1) = 1$, so this interval is off the path of play. The other possibility is that $p(\tau_1)$ was low enough that the Defender must not clear in (τ_1, τ_2). But then the Attacker must prefer infiltrating at any time just after τ_1 to infiltrating at any time near τ_2, so these times must also not be in the Attacker's support. This reasoning extends arbitrarily far into the future, so it must be the case that the Attacker does not infiltrate at any time in (τ_1, ∞). Again, the Defender must not clear in (τ_1, ∞), because they would prefer clearing at τ_1. But, this means that the Attacker is not best responding, as they would strictly prefer infiltration immediately after τ_1. An analogous argument can be applied to the Defender's strategy. This is made formal below:

Proposition 5
In the single system model ($n = 1$), all mixed strategy equilibrium strategies have full support, i.e., the support is $[0, \infty)$.

3 Multiple Systems

Suppose that the set of Defender systems is $N = \{1, 2, \ldots, n\}$. Each set of systems I has distinct flow benefits and losses associated with it: b_I, ℓ_I. There are signals associated with every subset $J \subseteq N$ of systems; J is the set of systems that could be relevant for the event that triggered the signal. The next signal time for each of these is independent and exponentially distributed, with rate parameters $\lambda_J \geq 0$. The total stock benefits and losses associated with signals may depend on the type of signal and the set of systems I that the Attacker is currently in: $B_J(I), L_J(I) \geq 0$ (these are aggregated over all systems in $I \cap J$). Both flow and stock benefits and losses abide by a monotonicity assumption: For all $I \subset I' \subseteq N$ and for all $J \cap I \neq \emptyset$, $b_I \leq b_{I'}$, $\ell_I \leq \ell_{I'}$, $B_J(I) \leq B_J(I')$, and $L_J(I) \leq L_J(I')$. The model allows for systems to be complements (e.g., some of the information in Systems 1 and 2 is useless to the Attacker without the corresponding information in the other system: $b_{\{1,2\}} > b_{\{1\}} + b_{\{2\}}$) or substitutes (e.g., some of the valuable information in Systems 1 and 2 is identical: $b_{\{1,2\}} < b_{\{1\}} + b_{\{2\}}$).

Checking a system without clearing costs C_D and does not initiate a pause; checking with clearing costs nC_D, all systems are cleared, and a pause affecting all systems begins. Intuitively, an Attacker being found in a system reveals a security vulnerability that has not yet been corrected in any system. If the vulnerability were fixed in only some systems, the Attacker could immediately infiltrate the other systems, so the Defender would immediately resume a low payoff. This could be justified endogenously, but in the interest of simplifying the state space, we assume that the clearing and pause occurs across all systems. In principle, the Defender may choose whether or not to clear systems after a signal arrival, but in Section 3 and Section 4, we focus only on equilibria where the Defender does clear all systems when a signal arrives.

Infiltration of the first system has total cost of C_A, as before. However, secondary infiltrations may happen instantaneously after infiltrating one system, and the costs of secondary infiltrations may be less than C_A. Getting into one system may give the Attacker easier access to other systems. Perhaps other systems have the same vulnerability. Perhaps there is some information inside of the former system giving clues for how to attack the latter. Perhaps the systems are networked in a way that there are fewer security barriers between them. Whatever the reason, the costs of infiltrating one system may depend on the systems the Attacker has already infiltrated. Moreover, these cost reductions may be asymmetric: infiltrating i makes is easier to infiltrate j but not vice versa.

Characterizing the Attacker's optimal sequence in which to infiltrate systems is a difficult problem in its own right. In fact, even in a simplified version of the problem in which all cost reductions are identical and cost dependencies are defined by a directed graph, the problem is NP-Hard. Its solution depends on finding a minimum feedback arc set, which is itself an NP-Hard problem (although an exact solution may be reasonable for a modest number of systems). For this reason, we avoid characterizing the Attacker's best response decision for the sequence of systems to infiltrate. We take as given the minimized total cost of infiltrating each set I: $C_A(I)$. We make no judgment about whether or not the Attacker exactly or only approximately optimizes in arriving at I. We do assume that the Defender best responds to I, which is certainly necessary for equilibrium.

Next, we derive conditions under which there exists an equilibrium where all of the systems in infiltration set I are treated as one, so the equilibrium resembles the single system mixed strategy equilibrium. Formally, this means that all systems in I are infiltrated at the same time, and thus the rest of the Attacker's strategy can be described by $A : [0, \infty) \to [0, 1]$, which gives the probability of next infiltrating all systems at a time $\leq \tau$ after infiltration last became possible. As before, this will have a mass point $A(0)$, and the rest of the distribution has an exponential form. For the Defender's strategy, they must check only one system in I at a time (since they believe the Attacker is always in either all systems in I or none of them). As before, it is exponential in form. There are independent "racing" exponentials: For each $i \in I$, $d_i(\tau)$ is an exponential PDF (possibly with rate depending on i), where τ is the time since the last check,

and the check occurs in the system with the lowest time realization, at that time. Thus, the next checking time is also exponentially distributed with rate equal to the sum of the individual systems' rates. This "racing exponentials" strategy is equivalent to a strategy with a single exponential distribution of the time of the next check, where the system checked is randomly determined i.i.d. for each check.

Note that the Defender must never check multiple systems at the same time in the equilibrium being studied here. Since the Attacker is always either in zero systems or in all systems of I, there is never a benefit to checking multiple systems, and it just incurs extra costs. As a result, it may be tempting for the Attacker to remove some systems from the infiltration set I, even if those systems have value outweighing the cost at the individual system level. Deviating to eliminate those systems reduces the rate at which the Defender clears *all* systems in I.

Note that the derivation of the Attacker's strategy follows the single system derivation almost exactly. There are two key differences. First, the signal rate must account for all systems in I. Let $\lambda(I) = \sum_{J \in \{J' \subseteq N | I \cap J' \neq \emptyset\}} \lambda_J$ be the overall signal rate for systems in I. Let $q_J(I) \equiv \frac{\lambda_J}{\lambda(I)}$ for $J \in \{J' \subseteq N | I \cap J' \neq \emptyset\}$ be the conditional probability of a signal being type J. Finally, let $L \equiv \sum_{J \in \{J' \subseteq N | I \cap J' \neq \emptyset\}} q_J(I) L_J(I)$ be the expected loss conditional on a signal of non-specific type, when the Attacker is in systems I. Note that many of these variables depend on I, but this is usually suppressed in the notation when I is clear from the context. Second, since checking is only done in one system, checking is less costly than clearing. The new approximate indifference condition is

$$
\begin{aligned}
&[p(\tau)e^{-r\Delta} + [1 - p(\tau)]]V_D(N) - [1 + p(\tau)(n-1)]C_D \approx p(\tau) \int_0^{d\tau} \{(1 - e^{-rt})[-\frac{\ell}{r}] \\
&+ e^{-rt}[-L - nC_D + e^{-r\Delta}V_D(N)]\}\lambda e^{-\lambda t} \cdot dt \\
&+ (1 - e^{-\lambda \cdot d\tau})[1 - p(\tau)]e^{-r \cdot d\tau}[V_D(N) - C_D] \\
&+ e^{-\lambda \cdot d\tau}\{(1 - e^{-r \cdot d\tau})[-p(\tau)\frac{\ell}{r}] \\
&+ e^{-r \cdot d\tau}[[p(\tau)e^{-r\Delta} + [1 - p(\tau)]]V_D(N) - [1 + p(\tau)(n-1)]C_D]\},
\end{aligned}
$$

Taking the first-order Taylor approximation, rearranging, dividing by $d\tau$, and taking the limit as $d\tau \to 0$ yields

$$
r\left[[p(\tau)e^{-r\Delta} + [1 - p(\tau)]]V_D(N) - C_D\right] = p(\tau)[(n-1)rC_D - \ell - \lambda L]
$$

Note that if the Attacker is in the systems, checking now costs more in the multiple system case than in the one system case (because of higher clearing costs). This accounts for adding $p(\tau)(n-1)rC_D$ to the right hand side. It is still the case that beliefs must be held constant, so $A^+(\tau) = 1 - e^{-\lambda A \tau}$. The

Defender's payoff is equal to that of checking immediately, so

$$\begin{aligned} V_D(N) &= [Ae^{-r\Delta} + (1-A)]V_D(N) - [1 + A(n-1)]C_D \\ \Leftrightarrow V_D(N) &= -\frac{[1 + A(n-1)]C_D}{A(1-e^{-r\Delta})} \end{aligned}$$

Plugging this into the indifference condition yields

$$\begin{aligned} &-\frac{[1 + A(n-1)]rC_D}{A(1-e^{-r\Delta})} = -A(\ell + \lambda L) \\ \Leftrightarrow &A^2(1-e^{-r\Delta})(\ell + \lambda L) - A(n-1)rC_D - rC_D = 0 \\ \Leftrightarrow &A = \frac{(n-1)rC_D + \sqrt{(n-1)^2r^2C_D^2 + 4(1-e^{-r\Delta})(\ell + \lambda L)rC_D}}{2(1-e^{-r\Delta})(\ell + \lambda L)} \end{aligned}$$

For this to be a strategy, we must have $\left(1-e^{-r\Delta}\right)\frac{\ell+\lambda L}{r} \geq nC_D$, i.e., the minimum losses averted by clearing exceed the total cost of clearing.

Now, we proceed to consider the Attacker's indifference condition. Once again, we have $V_A(\tau) = 0, \forall \tau$, by a straightforward extension of the single system proof. Letting $B_I \equiv \sum_{J \in \{J' \subseteq N | I \cap J' \neq \emptyset\}} q_J(I) B_J(I)$, and letting $D(\tau)$ be the probability that the Defender's next checking time in any system in I (i.e., the minimum of all checking times) is $\leq \tau$, then the Attacker's indifference condition appears the same as in the one system case except using multiple system benefits and costs, so we again derive $D(\tau) = 1 - e^{-\rho\tau}$, where $\rho = \frac{b+\lambda B-(r+\lambda)C_A(I)}{C_A(I)}$ and with the similar existence condition $b + \lambda B > (r+\lambda)C_A(I)$ (where subscript I's have been suppressed).

However, for this to be an equilibrium, one more condition must hold that is particular to the multiple systems setting: the Attacker must weakly prefer infiltrating the systems in I to infiltrating any other set of systems $I' \subseteq N$. For example, if $I' \subset I$, since the Defender will discover the Attacker at a faster rate when they are in I, they might prefer remaining in fewer systems but remaining there for longer in expectation per infiltration. As another example, if $I' \supset I$, the Defender will not explicitly check systems in $I' \setminus I$ in equilibrium, so those systems must be low enough value relative to their costs and vulnerability to signals. These conditions will constrain the sets I that may appear in equilibria as well as the individual system rates ρ_i given I, where $\rho_I = \sum_{i \in I} \rho_i$ is the signal rate associated with infiltrating only systems in I (again, I may be suppressed in the notation). For each $I' \subset N$, the I' deviation payoff, returning to the equilibrium path after clearing, must be less than or equal to the I infiltration payoff, which is 0:

$$\begin{aligned} &\int_0^\infty \Bigg\{ \int_0^t \left[\left(1-e^{-rs}\right)\frac{b_{I'}}{r} + e^{-rs}B_{I'} \right] \lambda(I') e^{-\lambda(I')s} \cdot ds \\ &+ e^{-\lambda(I')t}\left(1-e^{-rt}\right)\frac{b_{I'}}{r} \Bigg\} \rho_{I'} e^{-\rho_{I'}t} \cdot dt - C_A(I') \leq 0 \end{aligned}$$

This can be solved for a condition on ρ_i:

$$\Leftrightarrow \int_0^\infty \left\{ \left(1 - e^{-(r+\lambda(I'))t}\right) \frac{b_{I'} + \lambda(I')B_{I'}}{r + \lambda(I')} \right\} \rho_{I'} e^{-\rho_{I'}t} \cdot dt - C_A(I') \leq 0$$

$$\Leftrightarrow (b_{I'} + \lambda(I')B_{I'}) - (b_{I'} + \lambda(I')B_{I'}) \frac{\rho_{I'}}{r + \lambda(I') + \rho_{I'}} \leq (r + \lambda(I'))C_A(I')$$

$$\Leftrightarrow \rho_{I'} \geq \frac{b_{I'} + \lambda(I')B_{I'} - (r + \lambda(I'))C_A(I')}{C_A(I')} \equiv \hat{R}(I') \tag{2}$$

We define $R(I') \equiv \max\left\{\hat{R}(I'), 0\right\}$, so the conditions all take the form $\rho_{I'} \geq R(I')$ (rates must also be nonnegative). When positive, $R(I')$ is a net-benefit/cost ratio for infiltrating the systems in I'.

Let $\mathcal{N}^{max}$ be defined as the set of all $N^{max} \subseteq N$ such that $\hat{R}(N^{max}) \geq 0$ and for all disjoint covers $\{I'_k\}_{k=1}^m$ of N^{max}, $R(N^{max}) \geq \sum_{k=1}^m R(I'_k)$. In other words, N^{max} is worth infiltrating (if the clearing rates are low enough), and it has sufficiently valuable cost efficiencies. Note that the second condition implies that $[I' \cap N^{max} = \emptyset] \Rightarrow [R(I') = 0]$ (i.e., sets of systems outside of N^{max} are not worth infiltrating on their own). This is because $\{N^{max}, I'\}$ is a disjoint cover of N^{max}, and if $R(I') > 0$, then the last condition would imply that $R(N^{max}) > R(N^{max})$, which is impossible. $\mathcal{N}^{max}$ may contain multiple elements, but only as a knife-edge case. Note that for any $N^{max} \in \mathcal{N}^{max}$, $R(N^{max}) \geq R(I')$ for all $I' \subseteq N$, because all such I' belong to a disjoint cover of N^{max}. Thus, if $|\mathcal{N}^{max}| > 1$, there must be multiple maximizers of $R(I')$. $\mathcal{N}^{max}$ may also be empty, for example, if the cost efficiencies from infiltrating more systems are too small but all systems are valuable. This set $\mathcal{N}^{max}$ is precisely the set of all possible infiltrated sets I consistent with the type of equilibrium being constructed (i.e., in which the Attacker does not mix across systems):

Theorem 1
If $\mathcal{N}^{max} \neq \emptyset$, then for each $N^{max} \in \mathcal{N}^{max}$, there is an equilibrium where the Attacker infiltrates, with probability 1, all systems in N^{max} and no others, and $\rho_i = 0$ for all $i \notin N^{max}$. If $\mathcal{N}^{max} = \emptyset$, then there is no equilibrium in which the Attacker infiltrates, with probability 1, all of a set of systems (i.e., there must be mixing over the set of systems infiltrated).

Proof. If the Attacker only infiltrates systems in N^{max} in equilibrium, it is clear that $\rho_i = 0$ for all $i \notin N^{max}$, or the Defender is spending clearing costs for no reason. For any $I' \subset N, I' \cap N^{max} = \emptyset$, the relevant inequality of Equation (2) is satisfied if and only if $R(I') = 0$.

We then get a reduced system of inequalities for clearing rates. Let $n^{max} \equiv |N^{max}|$ and let M be a $(2^n - 2^{(n-n^{max})}) \times n^{max}$ matrix with each row corresponding to a non-empty subset $I' \subseteq N$ that intersects with N^{max}, and entry j in that row corresponds to some $j \in N^{max}$ and equals $-\mathbb{1}\{j \in I'\}$, unless $I' = N^{max}$, in which case it is equal to $\mathbb{1}\{j \in N^{max}\}$. Let d be the $(2^n - 2^{(n-n^{max})}) \times 1$ column vector with each entry corresponding to the same $I' \subseteq N$ intersecting with

N^{max} as the corresponding row in M and that entry equaling $-R(I')$, unless $I' = N^{max}$, in which case it is equal to $R(N^{max})$. Finally, let ρ be the $n^{max} \times 1$ column vector $(\ldots, \rho_i, \ldots)^T$ for all of the rates ρ_i with $i \in N^{max}$. Then, the reduced system of inequalities can be written as $M\rho \leq d$, and the objective is to prove that there exists a solution $\rho \geq 0$. We can apply Farkas' Lemma[6], which implies that such a solution exists if and only if there is no solution $y \geq 0$ to both $M^T y \geq 0$ and $d^T y < 0$.

Suppose to the contrary that $M^T y \geq 0$ has a solution with $d^T y < 0$ and $y \geq 0$. We index the entries of y by the non-empty subset $I' \subseteq N$ it corresponds to. Each row of M^T corresponds to a system $i \in N^{max}$, and each column corresponds to a set of systems I'. For the i row, each column $I' \neq N^{max}$ equals $-\mathbb{1}\{i \in I'\}$, and column $I' = N^{max}$ equals $\mathbb{1}\{i \in N^{max}\}$. Thus, $M^T y \geq 0$ is equivalent to stating that for all $i \in N^{max}$,

$$\begin{aligned} &- \sum_{I' \in \mathcal{I}(i)} y_{I'} + y_{N^{max}} \geq 0 \\ \Leftrightarrow &\sum_{I' \in \mathcal{I}(i)} y_{I'} - y_{N^{max}} \leq 0, \end{aligned} \tag{3}$$

where $\mathcal{I}(i) \equiv \{I' \subseteq N | I' \neq N^{max}, I' \cap N^{max} \neq \emptyset, i \in I'\}$

The condition $d^T y < 0$ is equivalent to

$$R(N^{max}) y_{N^{max}} - \sum_{I' \in \mathcal{I}} R(I') y_{I'} < 0, \tag{4}$$

where $\mathcal{I} \equiv \{I' \subseteq N | I' \neq N^{max}, I' \cap N^{max} \neq \emptyset\}$. We can identify a solution to these systems of inequalities (if one exists) by minimizing $R(N^{max}) y_{N^{max}} - \sum_{I' \in \mathcal{I}} R(I') y_{I'}$ subject to the constraints in Equation (3), as well as nonnegativity constraints for the $y_{I'}$'s. Moreover, the value of the objective function scales with the $y_{I'}$'s, so we may take $y_{N^{max}} \geq 0$ as a given parameter of the problem (if the minimized value is negative the condition holds for some y; if it is nonnegative, it cannot hold for any y).

There may be many solutions to this linear programming problem, but there is always a solution where all of the constraints in Equation (3) bind, since each $R(I') \geq 0$. Moreover, there always exists a pseudo-corner solution where for each $i \in N^{max}$, there exists a unique $I' \in \mathcal{I}(i)$ such that $y^*_{I'} > 0$ (and thus, $y^*_{I'} = y_{N^{max}}$). This is because the overall marginal effect of increasing any $y_{I'}$, including the direct effect via $R(I')$ and the indirect effects of tightening the constraints for all $i \in I'$, is constant. Therefore, there always exists a disjoint cover $\{I'_k\}_{k=1}^m$ of N^{max} for which there is a solution $y^*_{I'_k} = y_{N^{max}}$ for all $k \in \{1, \ldots, m\}$ and $y^*_J = 0$ for all other $J \neq N^{max}$. For this choice of disjoint cover, the minimized value equals $\left[R(N^{max}) - \sum_{k=1}^m R(I'_k)\right] y_{N^{max}}$. Since $N^{max} \in \mathcal{N}^{max}$, this must always be ≥ 0, i.e., Equation (4) cannot hold.

[6]The version of Farkas' Lemma being used states the following: Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then exactly one of the following assertions is true: $Ax \leq b$ has a solution $x \geq 0$, or $A^T y \geq 0$ has a solution $y \geq 0$ with $b^T y < 0$.

Thus, there exists a solution $\rho \geq 0$ to $M\rho \leq d$, which implies that there exist individual system clearing frequencies ρ_i that make Attacker deviations to infiltrating a subset of systems unprofitable and that sum to the correct aggregate frequency. That is, the candidate equilibrium is an equilibrium.

Now, consider the $\mathcal{N}^{max} = \emptyset$ case and any $N^{max} \subseteq N$. If $\hat{R}(N^{max}) < 0$, then there is no $\rho_{N^{max}}$ that gives the Attacker a nonnegative continuation payoff, whereas a payoff of 0 can be guaranteed by never infiltrating. If $R(I') > 0$ for some I' that does not intersect N^{max}, then the Attacker would receive a positive continuation payoff by deviating to infiltrate I', as they would not be explicitly cleared by the Defender. Finally, note that if $R(N^{max}) < \sum_{k=1}^{m} R(I'_k)$ for some disjoint cover $\{I'_k\}_{k=1}^{m}$ of N^{max}, the vector with $y_{I'_k} = y_{N^{max}}$ for all $k \in \{1, \ldots, m\}$, and $y_J = 0$ for all other $J \subseteq N$ solves both Equation (3) and Equation (4). Thus, there is no equilibrium where the Attacker infiltrates with probability 1 all of the systems in N^{max} and no others.

□

Note that an equilibrium that is symmetric across systems ($\rho_i = \rho_j, \forall i, j \in N$) may not be possible due to asymmetries in the parameters. For example, if System 1 is unusually high value (high b_1 and B_1) compared to System 2, the Defender will have to clear System 1 more frequently than System 2 in any of these equilibria. If the systems are identical ($b_i = b_j$, $B_{\{i\}} = B_{\{j\}}$, $\lambda_i = \lambda_j, \forall i, j \in N$, and $C_A(I)$ depends only on $|I|$), then they may have identical clearing rates in this type of equilibrium.

The definition of $\mathcal{N}^{max}$ shows a tension that may lead to the non-existence of this type of equilibrium. Recall that the Defender must never check multiple systems at the same time due to the desire to save on costs. They cannot, for example, have perfectly correlated checking times. This implies that removing a system from the equilibrium set may be tempting, because it reduces the overall rate at which the Attacker is cleared. However, if that System i is also valuable enough ($R(\{i\}) > 0$), then it must be in the equilibrium set, because if it is not, it will never be cleared, which makes it too tempting to leave out of the set. Thus, systems of low but positive net value may rule out equilibria with a pure strategy across the systems dimension, and these situations may be candidates for an equilibrium with a mixed strategy across systems.

Theorem 1 immediately implies a couple corollaries about the existence of an "infiltrate all" (i.e. $I = N$) equilibrium. The model has a "global" signal if $\lambda_N = \lambda > 0$, and $\lambda_I = 0, \forall I \subset N$. The model is one of "perfect substitutes" (given a global signal) if for all $I \subseteq N$, $b_I + \lambda B_I = \sum_{i \in I} b_i + \lambda \sum_{i \in I} B_i$. The model is one of "positive net values" if for every set of systems I, $b_I + \lambda(I) B_I > (r + \lambda(I)) C_A(I)$.

Corollary 1
If the model has a global signal, perfect substitutes, positive net values, and there are perfect cost reductions (i.e., $C_A(I) = C_A, \forall I \subseteq N$), then the "infiltrate all" equilibrium exists.

Proof. The relevant equilibrium condition is that for any partition $\{I_k\}_{k=1}^m$,

$$\frac{\sum_{i\in N} b_i + \lambda \sum_{i\in N} B_i - (r+\lambda)C_A}{C_A} \geq \sum_{k=1}^{m} \frac{\sum_{i\in I_k} b_i + \lambda \sum_{i\in I_k} B_i - (r+\lambda)C_A}{C_A}$$
$$= \frac{\sum_{i\in N} b_i + \lambda \sum_{i\in N} B_i - m(r+\lambda)C_A}{C_A},$$

which is clearly true. □

Corollary 1 uses a very extreme assumption about infiltration costs, but there is almost always some flexibility to have higher costs from infiltrating more systems. However, in the opposite extreme, the "infiltrate all" equilibrium does not exist:

Corollary 2
If the model has a global signal, perfect substitutes, positive net values, and there are no cost reductions (i.e., $C_A(I) = |I|C_A, \forall I \subseteq N$), then the "infiltrate all" equilibrium does not exist. Moreover, there is no equilibrium in which a set of systems is infiltrated with certainty.

Proof. Equilibrium requires that the clearing rates can deter all deviations to infiltrate only a single system:

$$\frac{\sum_{i\in N} b_i + \lambda \sum_{i\in N} B_i - n(r+\lambda)C_A}{nC_A} \geq \sum_{i\in N} \frac{b_i + \lambda B_i - (r+\lambda)C_A}{C_A}$$
$$= \frac{\sum_{i\in N} b_i + \lambda \sum_{i\in N} B_i - n(r+\lambda)C_A}{C_A},$$

which is never true for $n > 1$. By the definition of $\mathcal{N}^{max}$, positive net values implies that either $\mathcal{N}^{max} = \{N\}$ (which has just been ruled out) or $\mathcal{N}^{max} = \emptyset$. □

These corollaries show that cost reductions tend to support the existence of an "infiltrate all" equilibrium, and that both cases are possible in the extremes. Section 4 will address what happens in between in some special cases. Here, we have taken a cost-focused view. In the absence of cost reductions, the "infiltrate all" equilibrium could instead be supported by complementarities between systems from the perspective of Attacker benefits.

As one final general result, we consider the support of the Attacker's strategy for what sets of systems to infiltrate. The following result shows that in equilibrium, the Attacker must not infiltrate multiple sets of systems simultaneously if these sets are completely unrelated to one another or have a negative relationship:

Proposition 6
For any two disjoint sets of systems I and J, if the following hold:

1. $b_{I\cup J} + \lambda(I\cup J)B_{I\cup J} \leq b_I + b_J + \lambda(I)B_I + \lambda(J)B_J$ *(i.e., they are substitutes)*

2. $C_A(I \cup J) \geq C_A(I) + C_A(J)$ *(i.e., there are no costs savings from infiltrating both),*

then in equilibrium, the Attacker must never simultaneously infiltrate set $I \cup J$ ($I \cup J$ is not in the support of the Attacker's strategy).

Proof. The Attacker's indifference across time implies an expected payoff of 0 for every set of systems infiltrated in equilibrium, so following Equation (2)

$$\begin{aligned}
&b_{I\cup J} + \lambda(I \cup J)B_{I\cup J} - (r + \lambda(I \cup J) + \rho_{I\cup J})C_A(I \cup J) = 0 \\
\Rightarrow &b_I + b_J + \lambda(I)B_I + \lambda(J)B_J - (r + \lambda(I \cup J) + \rho_I + \rho_J)[C_A(I) + C_A(J)] \geq 0 \\
\Rightarrow &b_I + b_J + \lambda(I)B_I + \lambda(J)B_J - r[C_A(I) + C_A(J)] + \lambda(I)C_A(I) + \lambda(J)C_A(J) \\
&+ \rho_I C_A(I) + \rho_J C_A(J) > 0,
\end{aligned}$$

where the last inequality follows from $\lambda(I \cup J) \geq \max\{\lambda(I), \lambda(J)\}$. This inequality implies that at least one of I or J must yield a positive expected payoff if the Attacker deviated to infiltrating that set in isolation. This contradicts the equilibrium conditions. □

Infiltrating a smaller subset of systems not only reduces the costs per infiltration, but it also reduces the frequency at which the costs must be paid (less exposure to both signals and endogenous clearing). Consequently, if such a deviation does not harm the average benefit net of costs (guaranteed for one of the two sets by the two conditions), it will be profitable.

Note that Proposition 6 still allows for I and J to be simultaneously infiltrated along with some third set K. This third set K could have significant complementarities or costs savings when combined with both I and J. However, there are many leading cases where this is not possible, such as the cluster graphs example of Section 4.1.

4 Linear Costs Example

Suppose that the cost dependencies between systems are defined by a directed graph (with no self-links) with $n \times n$ adjacency matrix G. A link from system i to system j ($g_{ij} = 1$) means that infiltrating system i before system j reduces the cost of infiltrating system j by $d > 0$ (from the default cost of C_A). This is what is meant by "linear;" each cost reduction is of the same magnitude d. We assume that the maximum column sum $\max_{j\in N} \sum_{i\in N} g_{ij} \leq \frac{C_D}{d}$, so that infiltration costs are never negative.

Some comparative statics may be derived without considering the particular graph structure. We look at how difficult it is to satisfy the existence condition for the "infiltrate all" equilibrium (i.e., how hard it is to "support" this equilibrium). A few simplifying assumptions will allow us to separate graph structure from the other parameters. Suppose that for all $I \subseteq N$, $b_I + \lambda(I)B_I = |I|(b + \lambda B)$ for some $b, B, \lambda > 0$ and that for any partition $\{I_k\}_{k=1}^m$ of N, $\sum_{k=1}^m \lambda(I_k) - \lambda(N) = (m-1)\Lambda$ for some $\Lambda \geq 0$. This is true if all

of the systems are independent ($b_I = |I|b$, $\lambda(I) = |I|\lambda$, $B_I = B, \forall I$), or if there is a single global signal ($b_I = |I|b$, $\lambda(I) = \lambda$, $B_I = |I|B, \forall I$), or both ($b_I = |I|b$, $\lambda(I) = (|I| + 1)\lambda$, $B_I = \frac{|I|}{|I|+1}B, \forall I$, the last of these being justified by either the individual signals or the global signal conveying no benefit). Other cases are also allowed but are harder to justify. For example, this assumption does not allow each nonempty subset of systems to have its own signal, each with the same rate. Then, you would have $\lambda(I) = \left(2^{|I|} - 1\right)\lambda$, which would require $B_I = \frac{1}{2^{|I|}-1}B$. This implies that $B_{\{i\}} = B, \forall i$. If $|I| = 2$, then $B_I = \frac{1}{3}B$. However, B_I is a conditional expected value equal to $\frac{2}{3}B + \frac{1}{3}B_2 > \frac{1}{3}B$, where B_2 is the benefit from the two system signal.

First, we assume that any single system is worth infiltrating on its own if it is never explicitly cleared: $b+\lambda B-(r+\lambda)C_A > 0$. Let $\mu_\Lambda \equiv \frac{(r+\Lambda)C_A}{b+\lambda B}$. Intuitively, Λ is a measure of signal overlap: the per system average redundant signal rate counted by naïvely adding up the $\lambda(I_k)$'s, which double counts some signals. In the examples given, Λ always equals either 0 or λ. However, it could exceed λ if there were other signals in between independent "per system" signals and the global signal.

Consider any partition of N, $\{I_k\}_{k=1}^m$. Note that the relevant condition holds trivially for $m = 1$, so we only consider $m \geq 2$. The condition to be satisfied for all partitions is the following (from Theorem 1):

$$\frac{n(b+\lambda B) - (r+\lambda(N))C_A(N)}{C_A(N)} \geq \sum_{k=1}^{m} \frac{|I_k|(b+\lambda B) - (r+\lambda(I_k))C_A(I_k)}{C_A(I_k)}$$

$$\Leftrightarrow \mu_\Lambda \geq \frac{1}{m-1}\left[\sum_{k=1}^{m} \frac{|I_k|C_A}{C_A(I_k)} - \frac{nC_A}{C_A(N)}\right] \tag{5}$$

Equation (5) lets us derive some comparative statics about the basic parameters of the model. First, note that changes in the default infiltration costs C_A and the magnitude of cost reductions d both interact with the graph structure (the $C_A(I)$'s), so these comparative statics must be derived for specific classes of graphs.

Increasing the discount rate r or decreasing b or B all make the equilibrium easier to support by increasing μ_Λ (the no clearing cost-benefit ratio). Checking rates deter the Attacker by forcing them to pay more infiltration costs over time. Therefore, the constraints on clearing rates pertain to an average benefit to average cost ratio. Under our assumptions, the average benefits are identical across all subsets of systems. So, reducing b or B has a greater effect on the deviation side of the inequality, as the number of deviations multiplies the effect. Similarly, increasing r reduces the average benefit net of average costs, as infiltration costs are paid up front.

The effect of increasing λ depends on Λ. If $\Lambda = 0$, as with independent signals, then increasing λ is similar to increasing the infiltration benefits, thus making the equilibrium harder to support. This is because the infiltration costs due to signals is proportionate to the number of systems infiltrated, so deviations save on these costs. If $\Lambda = \lambda$, as is the case with a global signal, then $\frac{d\mu_\Lambda}{d\lambda} =$

$\frac{C_A(b+\lambda B)-B(r+\lambda)C_A}{(b+\lambda B)^2}$. This is positive, making the equilibrium easier to support, if $b > rB$. With a global signal, there is no savings on infiltration costs due to signals from deviating to infiltrate a smaller set of systems. So, more frequent signals will make deviations less appealing if signal arrivals are not sufficiently valuable to the Attacker.

Finally, we consider an increase in Λ, a measure of signal overlap, while λ is held fixed. An example of such a change is starting with independent signals for each system and then adding a global signal. This makes the equilibrium easier to support. The Λ in the numerator of μ_Λ has entirely to do with the costs of reinfiltration (signals' effect on benefits comes from the λ in the denominator). If there is more overlap, then for any deviation to a subset of systems, the Attacker avoids exposure to a smaller proportion of signals and the associated reinfiltration costs. For example, for any such deviation, the Attacker is still exposed to a global signal, but they now only get a proportion of the benefits if that signal arrives.

None of the above comparative statics depend on an expression for $C_A(I)$ or a characterization of the worst-case partition. As mentioned before, finding $C_A(I)$ for arbitrary graphs is an NP-Hard problem and will admit no simple characterization. However, we are able to answer some questions about graph structure by considering important classes of graphs for which there are simple expressions for costs.

4.1 Cluster Graphs

We consider any graph G on N, each component of which is a complete graph. This is known as a cluster graph. We call each component (i.e., maximal connected subgraph) of a cluster graph a "cluster." A cluster graph is a reasonable model of groups of computer systems that are physically networked together or that have special lines of communication open between them but which have no such connections to systems outside of the cluster. Another example of a cluster is a group of systems, each of which may contain distinct technical information about the other systems in the cluster, which may aid in infiltrating other systems in the cluster.

Let $\mathcal{C}$ be the set of components, which is a partition of N. We will sometimes refer to components sorted by size, in ascending order: $I^1, I^2, \ldots, I^{|\mathcal{C}|}$. We maintain all other assumptions of Section 4. We also assume that any single system is worth infiltrating on its own if it is never explicitly cleared: $b + \lambda B - (r + \lambda)C_A > 0$. Let $\mu \equiv \frac{(r+\lambda)C_A}{b+\lambda B} < 1$ be the single system cost/benefit ratio. $\mu < 1$ implies that the only candidate equilibrium that infiltrates a set of systems with certainty is the equilibrium that infiltrates all systems.

Any infiltration set I is split into complete subgraphs by intersection with components. For each of these subgraphs containing k systems, the last infiltrated system gets $k-1$ cost reductions, the next to last gets $k-2$, and so on, for a total of $\frac{k(k-1)}{2}$ cost reductions. Therefore, the condition to be satisfied for

all partitions is the following (from Equation (5)):

$$\mu_\Lambda \geq \frac{1}{m-1}\left[\sum_{k=1}^{m} \frac{|I_k|C_A}{|I_k|C_A - \sum_{I\in\mathcal{C}} \frac{|I_k\cap I|(|I_k\cap I|-1)}{2}d} - \frac{C_A}{C_A - \sum_{I\in\mathcal{C}} \frac{|I|}{n}\frac{|I|-1}{2}d}\right] \tag{6}$$

The RHS of Equation (6) pertains to average "cost reduction quotients." These quotients measure the magnitude of the cost reductions for various infiltration sets. For any component I, let $Q(I) \equiv \frac{C_A}{C_A - \frac{|I|-1}{2}d}$ be the associated cost reduction quotient. Also relevant are weighted average cost reduction quotients. Let $W(\mathcal{I}) \equiv \frac{C_A}{C_A - \sum_{I\in\mathcal{I}} \frac{|I|}{\sum_{I'\in\mathcal{I}}|I'|}\frac{|I|-1}{2}d}$. This is a weighted cost reduction quotient, using a weighted sum of cost reductions across the components in $\mathcal{I}$, and where the weights are the proportions of systems belonging to the components. $W \equiv W(\mathcal{C})$ is the weighted cost reduction quotient corresponding to infiltrating all systems (the last term in Equation (6)).

Result 1
If G is a cluster graph, then the "infiltrate all" equilibrium exists if and only if $|\mathcal{C}| = 1$, i.e., G is the complete graph, and $\mu_\Lambda \geq \frac{1}{n-1}[n - W]$.

To understand why the "infiltrate all" equilibrium never exists when there are multiple clusters, it is worth reviewing where the clearing rate bounds for existence come from. In order to maintain indifference in the "infiltrate all" equilibrium, the Attacker's costs due to explicit clearing, $\rho_N C_A(N)$, must exactly equal the Attacker's net benefit in the absence of explicit clearing, $n(b + \lambda B) - (r + \lambda(N))C_A(N)$. This requires an aggregate checking rate ρ_N equal to the ratio of average net benefit (in absence of explicit clearing) to average infiltration cost. For all deviations to subsets I, the clearing rate ρ_I must exceed the ratio of average net benefit to average cost (averaging over I). Under our assumptions, the average benefits $b + \lambda B$ are identical across all infiltration sets. Infiltrating the largest cluster $I^{|\mathcal{C}|}$ must have an average cost no greater than that of infiltrating all systems. Moreover, it cannot result in exposure to more signals: $\lambda(I^{|\mathcal{C}|}) \leq \lambda(N)$. Therefore, existence requires $\rho_{I^{|\mathcal{C}|}} \geq \rho_N$. But this requires that some other systems are not explicitly cleared, which allows positive payoffs from deviation to infiltrate those systems.

If there is only one cluster, i.e., the graph is complete, then the "infiltrate all" equilibrium may exist. The issues raised in the previous paragraph are not decisive, as deviating to infiltrate a subset of systems must "break" links. This causes all subsets of systems to have higher average infiltration costs compared to infiltrating all systems. For comparative statics, note that increasing the cost reduction quotient from infiltrating all systems, W, while keeping the number of systems fixed, must make the equilibrium easier to support. This may result from an increase in the magnitude of cost reductions, d, or a decrease in the default infiltration cost, C_A. These changes both increase the relative cost savings from infiltrating more systems, making the "infiltrate all" equilibrium relatively more attractive. Increasing the number of systems also makes the

"infiltrate all" equilibrium easier to support, as the RHS of the equilibrium condition is

$$\frac{1}{n-1}[n-W] = \frac{1}{n-1}\frac{n[C_A - \frac{n-1}{2}d] - C_A}{C_A - \frac{n-1}{2}d}$$
$$= \frac{C_A - \frac{n}{2}d}{C_A - \frac{n-1}{2}d},$$

which strictly decreases as a function of n. Intuitively, the potential cost savings increase superlinearly with the total number of systems, improving the average cost savings of infiltrating more systems.

The non-existence of the "infiltrate all" equilibrium in the case of multiple clusters raises the question of what equilibrium may arise in its place. Theorem 1 shows that it cannot involve infiltration of some set of systems with certainty; there must be mixing over systems. There may instead exist an equilibrium in which the Attacker always infiltrates an entire cluster but randomizes over which cluster to infiltrate. Let $\mu_\Lambda(I) \equiv \frac{(r+\Lambda(I))C_A}{b_I+\lambda(I)B_I}$, where $\Lambda(I) \equiv \frac{\lambda(I)-\sum_{k=1}^m \lambda(I_k)}{m-1}$ for all partitions $\{I_k\}_{k=1}^m$ of I (assumed to be the same across all such partitions). Then, the following result holds:

Proposition 7
If G is a cluster graph, then there exists an equilibrium with support $\mathcal{C}$ (in the systems dimension) if and only if for all $I \in \mathcal{C}$, $\mu_\Lambda(I) \geq \frac{1}{|I|-1}[|I| - W(\{I\})]$.

Although we presented Result 1 independently because of its simplicity, Proposition 7 generalizes it as a special case when $|\mathcal{C}| = 1$. For each cluster, the probability that it has been infiltrated (in the absence of signals) is held constant at A. As a result, infiltration occurs at a faster rate, but it affects fewer systems at a time. The Defender is held indifferent about which cluster to check at any given time, checking only a single system at a time. After checking, conditional on not having infiltrated, the Attacker may immediately infiltrate the checked cluster, bringing the Defender back to the point of indifference. Checking occurs at a faster rate than before, but clearing does not (this still keeps the Attacker indifferent), because checking the wrong cluster will fail to discover and clear the Attacker. Finally, infiltration is uniform across clusters. At first glance, this seems unlikely. Won't the Defender then strictly prefer checking high loss clusters? No, because the Defender gets the same payoff from checking a cluster immediately and waiting, and the "check immediately" payoff is the same across clusters. The comparative statics are much the same as those of Result 1, but they hold at the cluster level for each cluster.

5 Long Run Agreements

The Markovian equilibria we have focused on thus far have the simplification that short run play (i.e., between clearing events) simply repeats itself forever.

As a result, there was no possibility of the parties cooperating on better outcomes with the threat of penalties if they defect from cooperation. In this section, we view the setting as a repeated game with the short run game as the stage game and look at some equilibria with cooperation. We assume that attribution is perfect; when a signal arrives, the Defender nation always knows that the Attacker nation is responsible. Imperfect attribution may make cooperative agreements less valuable and less permanent, but they are still possible.[7]

5.1 Symmetric Case

First, we consider the symmetric case, where both nations (1 and 2) act as both Attacker and Defender, choosing both infiltration and clearing times. We focus on equilibria where strategies do not interact across the two games (the 1 infiltrating 2 game and the 2 infiltrating 1 game). This allows us to apply the "aggressive" equilibrium from the previous section to both cases. All parameters may be nation specific, so Nation i's parameters will have a superscript of i. In the interaction of Nation i attacking Nation j, in the aggressive equilibrium, Nation i gets payoff 0, and Nation j gets payoff $-\sqrt{\frac{C_D^j(\ell^j+\lambda^j L^j)}{r^j(1-e^{-r^j\Delta})}}$. Summing up each nation's Attacker and Defender payoffs, each Nation i gets expected payoff $-\sqrt{\frac{C_D^i(\ell^i+\lambda^i L^i)}{r^i(1-e^{-r^i\Delta})}} \equiv -\pi_D^i$. The following result (a corollary of Proposition 3) shows conditions under which cyber peace (no attacks or explicit clearing) is possible, using a grim trigger punishment:

Corollary 3
There exists an equilibrium with cyber peace on the path of play if the following holds for all $i \neq j$:

$$b^i + \lambda^j(B^i - e^{-r^i\Delta}\pi_D^i) - (r^i + \lambda^j)C_A^i \leq 0$$

Proof. There is no beneficial Defender deviation, as clearing is not expected to have any effect. The best Attacker deviation is to infiltrate immediately. We construct an equilibrium where as soon as the signal process reveals an infiltration from either nation, the Defender clears that system, and they revert to the aggressive equilibrium forever. Attacker i will not be tempted to deviate

[7] If imperfect attribution were only a matter that some cyber attacks cannot be attributed to anyone, and this prevents triggering a punishment, the agreement equilibrium conditions would be tighter but would otherwise be similar. On the other hand, if there were some chance of attributing a cyber attack to the other nation despite their innocence, then the agreement equilibrium becomes more complicated and/or lower payoff. However, agreement is still possible, and may feature temporary punishments that are triggered on the equilibrium path of play. Baliga et al. [2020] focuses on the deterrence problem in cyberwarfare, albeit in a model where retaliation is valued for its own sake, not just as an off-path punishment.

if the following holds:

$$\int_0^\infty \left[(1-e^{-r^i t})\frac{b^i}{r^i} + e^{-r^i t}\left(B^i - e^{-r^i \Delta}\pi_D^i\right)\right]\lambda^j e^{-\lambda^j t}\cdot dt - C_A^i \leq 0$$
$$\Leftrightarrow b^i + \lambda^j(B^i - e^{-r^i\Delta}\pi_D^i) - (r^i+\lambda^j)C_A^i \leq 0$$

□

First, note that this equilibrium cannot exist if B^i is too large. Since punishments are enforced entirely via the signal process, if the expected benefits to the Attacker of a signal arrival outweighs the punishment, then they actually want the signal arrival to happen, despite the punishment. Second, note that if $B^i < e^{-r^i\Delta}\pi^i$ but λ^j is low, the signal process may not catch the infiltrator quickly enough for the punishment to be a sufficient deterrent, so again the equilibrium does not exist. Finally, note that as $r^i \to 0$, then $\pi_D^i \to \infty$. Thus, we get the usual condition that the equilibrium exists if r^i is sufficiently low for all $i \in \{1,2\}$.

This also shows that, counterintuitively, a nation may wind up worse off from a public improvement in their own defensive capabilities. If the nations have an agreement of cyber peace between them, then reduction in one's own defensive costs may make violating the agreement overly tempting by reducing the punishment costs, thereby switching to the aggressive equilibrium. This type of improvement is always valuable if there is no hope of coming to an agreement. However, defensive improvements that increase the other nation's infiltration costs have the opposite effect of making agreements easier to support. This type of investment is valuable if there is currently no agreement and making the investment encourages the formation of an agreement, but it is useless otherwise. This leads to the following rule of thumb for defensive investment: invest in the cost efficiency of monitoring and patching when cooperation is hopeless, and invest in closing security vulnerabilities in order to encourage cooperation. Note that realistically, there will be some overlap. For example, closing i's security vulnerabilities may raise both C_A^j and Δ, and the overall effect on cooperation is ambiguous.

Now, consider the case where each nation has two systems (subscript 1 and 2) with a single global signal (independent per nation). We consider the existence of an equilibrium where each nation infiltrates only System 1 of the other nation, and there is no clearing, and where deviations are punished with the aggressive "infiltrate all" equilibrium.[8] We call this a "partial infiltration" equilibrium. Is such an equilibrium easier to support than the peaceful equilibrium? The change in the inequality for attacking deviations is simple. For the "no infiltration" equilibrium, the equilibrium payoff of 0 must exceed the "infiltrate both" deviation. For the "partial infiltration" equilibrium, the equilibrium payoff of $\frac{b_1^i+\lambda^j B_1^i-\ell_1^i-\lambda^i L_1^i}{r^i}$ must exceed the "infiltrate both" deviation. Thus, the "partial infiltration" equilibrium is easier to support if and only if

[8] In light of Corollary 2, we assume that cost reductions are sufficiently large.

$b_1^i + \lambda^j B_1^i > \ell_1^i + \lambda^i L_1^i, \forall i$, i.e., it is a Pareto improvement to have both nations infiltrating the systems labeled 1, relative to cyber peace. Note that in the "partial infiltration" equilibrium, there is no temptation for a nation to deviate to clear System 1, for this immediately triggers the punishment for a negative continuation payoff. The "partial infiltration" equilibrium may not exist even when it is joint welfare maximizing and the cyber peace equilibrium does exist. This may occur due to asymmetry with one nation losing as a result of the partial infiltration, although this obstacle could be overcome with conditional transfers. Because in the symmetric case they exist only if partial infiltration is efficient, we believe these "partial infiltration" equilibria are far more interesting and plausible in the asymmetric case, discussed next.

5.2 Asymmetric Case

Now, suppose that only one of the nations is capable of attacking (Nation A, with Nation D as the Defender). Since the Attacker can never receive a negative payoff, cooperation involving no infiltration is impossible, because this gives the Attacker zero payoff. For this reason, we look at a setting with two systems with perfect cost reductions ($d = C_A$, so the second infiltration costs nothing), where the systems share a global signal with rate λ but are otherwise independent, system-level parameters have subscripts $i \in \{1, 2\}$, and the "infiltrate all" equilibrium serves as the punishment outcome. We construct a "partial infiltration" equilibrium where on the path of play, the Attacker infiltrates only System 1, and the punishment is triggered if the Attacker is ever discovered (by the signal) to be in System 2 or the Defender ever explicitly checks System 1.[9] In this equilibrium, the Defender never checks either system on the path of play. This includes signals revealing the Attacker to be in System 1. Signals revealing the Attacker to be in System 2 result in both systems being cleared, and the punishment is triggered. Because System 1 is never cleared, the Attacker is not indifferent, and will immediately infiltrate System 1 on the path of play. The existence conditions for a partial infiltration equilibrium are given in the following result (a corollary of Theorem 1):

Corollary 4
A partial infiltration equilibrium exists if the following two conditions hold:

$$b_1 + \lambda B_1 - (r^A + \lambda)C_A \geq \frac{r^A}{\lambda}[b_2 + \lambda B_2] \tag{7}$$

$$\ell_1 + \lambda L_1 \leq e^{-r^D \Delta} \frac{C_D + \sqrt{C_D^2 + 4(1 - e^{-r^D \Delta})\frac{\ell + \lambda L}{r^D} C_D}}{2(1 - e^{-r^D \Delta})} + C_D \tag{8}$$

[9] In some circumstances, it might be impossible for the Attacker to distinguish between a Defender clearing System 1 explicitly and clearing it as a result of a signal. This would be true if the signal were due to some passive cybersecurity monitoring. However, this is not a problem if the signal causes the Attacker to inflict damage on the Defender, thereby revealing their presence. This latter case is consistent with our global signal assumption.

Proof. On the equilibrium path of play, the Attacker receives payoff

$$\begin{aligned} V^A &= \int_0^\infty [(1-e^{-r^A t})\frac{b_1}{r^A} + e^{-r^A t}(B_1+V^A)]\lambda e^{-\lambda t}\cdot dt - C_A \\ \Leftrightarrow V^A &= \frac{b_1}{r^A+\lambda} + \frac{\lambda}{r^A+\lambda}(B_1+V^A) - C_A \\ \Leftrightarrow V^A &= \frac{b_1+\lambda B_1 - (r^A+\lambda)C_A}{r^A} \end{aligned}$$

If the Attacker deviates to infiltrate System 2 as well as System 1, they get payoff $\frac{b_1+b_2+\lambda(B_1+B_2)}{r^A+\lambda} - C_A$. The Attacker will not deviate if the following holds:

$$\begin{aligned} &\frac{b_1+\lambda B_1-(r^A+\lambda)C_A}{r^A} \geq \frac{b_1+b_2+\lambda(B_1+B_2)}{r^A+\lambda} - C_A \\ \Leftrightarrow &\frac{b_1+\lambda B_1-\lambda C_A}{r^A} \geq \frac{b_1+b_2+\lambda(B_1+B_2)}{r^A+\lambda} \\ \Leftrightarrow &b_1+\lambda B_1-(r^A+\lambda)C_A \geq \frac{r^A}{\lambda}[b_2+\lambda B_2] \end{aligned}$$

The Defender receives the equilibrium payoff $-\frac{\ell_1+\lambda L_1}{r^D}$. The Defender's most tempting deviation is to immediately clear system 1, which leads to the aggressive "infiltrate all" equilibrium as a punishment, so they get the "clear immediately" payoff followed by the punishment equilibrium: $e^{-r^D\Delta}V_D(N) - C_D = -e^{-r^D\Delta}A\frac{\ell+\lambda L}{r^D} - C_D$. Therefore, the Defender will not be tempted to deviate if and only if

$$\ell_1+\lambda L_1 \leq e^{-r^D\Delta}\frac{C_D+\sqrt{C_D^2+4(1-e^{-r^D\Delta})\frac{\ell+\lambda L}{r^D}C_D}}{2(1-e^{-r^D\Delta})} + C_D$$

□

Deviation for the Attacker is unappealing (Equation (7)) if System 1 is valuable relative to System 2 ($b_1+\lambda B_1$ is large relative to $b_2+\lambda B_2$), the Attacker is patient (r^A is small), or the signal which catches deviations is frequent (λ is large) as long as $B_1 > C_A$. Deviation for the Defender is unappealing (Equation (8)) when infiltration of System 2 entails heavy losses relative to System 1 and the Defender is patient (r^D is small).

Overall, this partial infiltration equilibrium tends to exist when the infiltrated system(s) are high benefit to the Attacker and low loss to the Defender (though infiltrating System 1 need not be joint welfare maximizing), while the uninfiltrated system(s) are low benefit to the Attacker and high loss to the Defender. This agrees with the common perception that equilibrium might allow low grade cyberwar in which information systems are regularly infiltrated without escalation (e.g., pure information gathering), but critical systems are considered off limits (e.g., power and water systems). This is similar to the idea of setting rules of cyberwarfare, such as in a "Digital Geneva Convention" (pg.

329 of Sanger [2019]), and our conditions might be thought of as conditions for countries to abide by such a convention.

The existence of the partial infiltration equilibrium begs the question, why not simply have an information sharing agreement and circumvent infiltration costs? This may reasonably exist as a policy between allies, however, it seems unlikely between adversaries. When pragmatists wield power but are publicly held in check by ideologues, then the unofficial and secretive policy of refusing to publicize and punish information gathering cyber operations may be preferred. Alternatively, it might be desirable to keep these agreements secret and unofficial, because the fact that two nations are able to reach an agreement may reveal something about their capabilities to third parties.

6 Investment in Defense: Private vs. Public

In many nations (e.g., liberal democracies), cybersecurity decisions are not made by a social planner, who internalizes all costs and losses for all systems, but rather by private stakeholders in individual systems. There have been several cyberattacks on private systems (e.g., Sony Pictures hack, WannaCry, NotPetya, and the 2021 Microsoft Exchange hack). This has been cited as a cybersecurity vulnerability of liberal democracies. As David Sanger notes about the Obama administration's reasoning on cyber defense in corporate America, "Clearly, the government could not protect against every cyberattack, just as it could not protect against every car theft or house burglary," (pg. 146 of Sanger [2019]). Here, we formalize this in a variant of our model as a positive externality of cybersecurity that upstream firms have on downstream firms. This is perhaps exemplified by the 2021 attack on Microsoft Exchange Servers, which though nominally an attack on Microsoft, gave Attackers access to the systems of many other organizations, a plurality of them U.S. based ("Victims of Microsoft hack" 2021).

Here, we consider a model with two systems ($n = 2$) and a single nation-state Attacker. System 1 is the "upstream" system, and it confers an infiltration cost discount of $d > 0$ on "downstream" System 2. System 2 confers no cost reduction on System 1. For simplicity, we assume that there is only a global signal with rate λ. Default infiltration costs may differ between the two systems, and they are denoted by C_A^i. Finally, the Defender or Defenders have an initial investment stage prior to playing the subsequent cyberwarfare game. They may pay a protection cost $C_P > 0$ (per system), which makes infiltrating that system impossible or sufficiently costly that the Attacker will never infiltrate in equilibrium, as in Condition 1. These choices are observed by all players before the cyberwarfare stage begins. Thus, backward induction may apply, and we look only at equilibria where the cyberwarfare stage subgames feature equilibria of the type already discussed in this paper. We call this the "investment game," and we consider two cases: a single public Defender, and two independent private Defenders.

We assume that System 1 is worth infiltrating on its own if possible ($(r +$

$\lambda)C_A^1 < b_1 + \lambda B_1)$, but $(r+\lambda)C_A^2 > b_2 + \lambda B_2$ and $(r+\lambda)(C_A^2 - d) < b_2 + \lambda B_2$, so that the Attacker is willing to infiltrate System 2 if and only if they have infiltrated System 1 (conditional on no investment to protect System 2). This creates the possibility of an extensive margin externality: the public Defender protects System 1 (and indirectly System 2), but the private Defender 1 does not. We also assume that the private Defenders share information on the results of checking: if the Attacker is discovered or not discovered, this is shared with the other Defender (this is irrelevant for the signal, which is global).

Proposition 8
In the investment game, the public Defender would protect System 1 *whereas the private Defender* 1 *would not if either of the following holds:*

1. $rC_D < (1 - e^{-r\Delta})(\ell_i + \lambda L_i)$ *for all* i,
   and $C_P \in \left(\sqrt{\frac{C_D(\ell_i + \lambda L_i)}{r(1-e^{-r\Delta})}}, \frac{C_D + \sqrt{C_D^2 + 4(1-e^{-r\Delta})\frac{\ell+\lambda L}{r} C_D}}{2(1-e^{-r\Delta})} \right)$
2. $(1 - e^{-r\Delta})(\ell_1 + \lambda L_1) < rC_D < (1 - e^{-r\Delta})(\ell_2 + \lambda L_2)$,
   and $C_P \in \left(\frac{\ell_1 + \lambda L_1}{r}, \frac{\ell_1 + \lambda L_1}{r} + \sqrt{\frac{C_D(\ell_2 + \lambda L_2)}{r(1-e^{-r\Delta})}} \right)$

Proof. First, suppose that both systems are worth clearing: $rC_D < (1-e^{-r\Delta})(\ell_i + \lambda L_i)$ for all i. A public Defender will either protect System 1 or protect nothing (resulting in the usual "infiltrate all" equilibrium, which we assume to exist from d being sufficiently large), because protecting System 1 effectively protects both systems. There is an equilibrium where the public Defender protects System 1 whenever

$$C_P \leq \frac{C_D + \sqrt{C_D^2 + 4(1 - e^{-r\Delta})\frac{\ell+\lambda L}{r} C_D}}{2(1 - e^{-r\Delta})} \tag{9}$$

In the case with private Defenders, Defender i will not protect its system when[10]

$$C_P > \sqrt{\frac{C_D(\ell_i + \lambda L_i)}{r(1 - e^{-r\Delta})}} \tag{10}$$

There are always intermediate values of C_P satisfying both Equation (9) and Equation (10) for each i.

Next, suppose that System 1 is not worth clearing, but System 2 is: $(1 - e^{-r\Delta})(\ell_1 + \lambda L_1) < rC_D < (1 - e^{-r\Delta})(\ell_2 + \lambda L_2)$. For both the public and private cases, the equilibrium is like a single system equilibrium (for System 2), where the Attacker always immediately infiltrates System 1 and this system is

[10] In the case of private Defenders, the "infiltrate all" mixed strategy equilibrium will usually look a bit different from in Section 3, because the two Defenders will have different indifference conditions, so the systems' infiltration rates must be different. We assume that such an equilibrium exists. However, the Defenders' indifference must result in the payoff given, as they get the same payoff as the "clear immediately" strategy yields.

never cleared, even when the signal reveals the Attacker's presence. The public Defender will protect System 1 if

$$C_P \leq \frac{\ell_1 + \lambda L_1}{r} + \sqrt{\frac{C_D(\ell_2 + \lambda L_2)}{r(1 - e^{-r\Delta})}}, \tag{11}$$

where the first term on the RHS corresponds to the System 1 payoff and the second term corresponds to the System 2 payoff. In the private case, Defender 1 will not protect System 1 if

$$C_P > \frac{\ell_1 + \lambda L_1}{r} \tag{12}$$

□

In either of the cases of Proposition 8, the Defenders in the private model suffer a loss from a positive externality. Both systems are infiltrated despite it being worth the cost to collectively protect them. In Case 1, the externality derives entirely from the public Defender considering the costs of clearing *both* systems in the case where one is found to be infiltrated. A private Defender only considers paying to clear their own system, even though both systems will be cleared if the Attacker has infiltrated. The magnitude of Defender welfare lost from this externality is $\sqrt{\frac{C_D(\ell_1+\lambda L_1)}{r(1-e^{-r\Delta})}} + \sqrt{\frac{C_D(\ell_2+\lambda L_2)}{r(1-e^{-r\Delta})}} - C_P$. The comparative statics on the size of this externality are very natural. The externality increases with C_D and decreases with C_P, as clearing costs harm the unprotected Defender(s), and protection costs make the public Defender's protection action less valuable. Increasing the magnitudes of the losses $(\ell_i + \lambda L_i)$ makes protection more desirable, increasing the size of the externality. Finally, any increase of the discount factor applied to the pause period (i.e., $1-e^{-r\Delta}$) reduces the size of the externality, because it makes clearing more beneficial for an unprotected system. Increasing r also has another effect of reducing the magnitude of future losses compared to the up front costs of protection, which also reduces the size of the externality.

In Case 2, the externality derives not from the public Defender considering the costs of clearing both systems at the same time (as it clears only System 2) but rather from the entire payoff from System 2, which is not considered by the private Defender 1. If we assume that Defender 2 also does not protect ($C_P > \sqrt{\frac{C_D(\ell_2+\lambda L_2)}{r(1-e^{-r\Delta})}}$), the magnitude of the externality is equal to the slack in Equation (11): $\frac{\ell_1+\lambda L_1}{r} + \sqrt{\frac{C_D(\ell_2+\lambda L_2)}{r(1-e^{-r\Delta})}} - C_P$. The comparative statics are similar to the previous case, except the the clearing costs and pause period do not matter for System 1, because it is not cleared. If instead Defender 2 does protect ($C_P < \sqrt{\frac{C_D(\ell_2+\lambda L_2)}{r(1-e^{-r\Delta})}}$), then the size of the externality is the loss that is still suffered by Defender 1: $\frac{\ell_1+\lambda L_1}{r}$.

Note that inter-system externalities do not appear without adding the investment stage. Consider the model of Section 3. On the extensive margin, the

Attacker's conditions which determine whether or not there is a pure strategy equilibrium depend only on the Attacker's parameters, and do not depend on the public vs. private nature of the Defender. On the intensive margin (i.e., comparing the public and private mixed strategy equilibria), there is no externality for two reasons. First, the Attacker always infiltrates System 2 immediately after infiltrating System 1, so there is no opportunity for the Defender to prevent the attack from spreading from System 1 to System 2. Second, even if there were a minimum delay between infiltrating each system so clearing System 1 would protect System 2, the clearing rates in equilibrium make the Attacker indifferent, so they depend on the Attacker's parameters and not the public vs. private nature of the Defender.

7 Conclusion

Cyber attacks by nation states have become very common in recent years, and more nations have become involved in launching attacks. This has raised many urgent questions. How can a nation improve its cyberwarfare outcomes? Is investment in offensive or defensive capabilities more effective? Which systems should be attacked (defended) and which should be left alone? How can we achieve a state of cyber peace? What may cause externalities between Defenders that are private entities? In this paper, we addressed these questions with an Attacker-Defender game theory model in which the Attacker decides when to infiltrate one or more systems, and the Defender when to monitor, clear, and patch them.

We find that except for extreme cases, in Markovian equilibrium, nations play full support mixed strategies with respect to timing of their actions. This implies that Attackers reap no net benefit from their attacks in equilibrium, so the outcome is not Pareto efficient. Payoffs are entirely determined by Defender characteristics, so investment in the cost efficiency of defense and in loss mitigation are the only ways to improve outcomes (in this particular equilibrium).

The picture is more complicated when the Defender has multiple systems to attack. From the Attacker's perspective, there may be interactions between systems. We focused on infiltration cost reduction relationships between systems, and showed that positive net value systems are always infiltrated in equilibrium, but negative net value systems may still be infiltrated because of the cost reductions they generate for infiltrating later systems. High magnitude of cost reductions tends to support infiltrating more systems in equilibrium. In the case where cost reductions are defined by a cluster graph, we showed that at most one cluster may be infiltrated at a time, and even this may fail. In general, more systems may be simultaneously infiltrated if cost reductions are a larger proportion of infiltration costs and if clusters are larger. If Defender nations wish to avoid widespread simultaneous infiltration, they should work to further compartmentalize systems or groups of systems and eliminate vulnerabilities shared in common between systems, perhaps by using different platforms and requiring individuals to use distinct credentials.

We then considered a problem that is common in liberal democracies: externalities that arise when private entities are in charge of defending their own systems. We supposed that these entities may invest in making infiltration of their systems prohibitively costly. When one system is upstream of (confers an infiltration cost reduction on) another downstream system, then infiltration of the former may effectively be a prerequisite for infiltration of the latter. Yet, this is a benefit that is external to the upstream Defender, so they may not make efficient defensive investments. This tends to be a problem when the infiltration losses suffered by the downstream Defender are large relative to the losses suffered by the upstream Defender. Therefore, this tends to be a serious problem with an upstream platform that is used by many downstream users, as was the case with the Microsoft Exchange hack.

Finally, we considered non-Markovian equilibria with efficient outcomes, i.e., tacit cyber peace agreements. First, we noted that cyber peace is reliant on nations possessing sufficiently effective latent monitoring processes, as no one would deliberately monitor if they expect to find no infiltrators. Second, we noted that the same investment in cost efficiency of defense and in loss mitigation that helps in a state of cyber war will also make cyber peace less likely, as the punishment for violating the agreement is less severe. On the other hand, making your systems harder to infiltrate could both improve cyber war outcomes and make cyber peace more likely (by increasing Attacker costs). Finally, we show that in any asymmetric situation where only one nation can mount an attack, any "peaceful" equilibrium must allow the Attacker to infiltrate some systems, and this will tend to work when the infiltrated systems are high benefit to the Attacker, and low loss to the Defender, while the reverse is true for uninfiltrated systems.

References

Daron Acemoglu and Alexander Wolitzky. Cycles of conflict: An economic model. *American Economic Review*, 104(4):1350–67, 2014.

Daron Acemoglu, Mikhail Golosov, Aleh Tsyvinski, and Pierre Yared. A dynamic theory of resource wars. *The Quarterly Journal of Economics*, 127(1): 283–331, 2012.

Kyle Bagwell and Robert W Staiger. An economic theory of gatt. *American Economic Review*, 89(1):215–248, 1999.

Sandeep Baliga, Ethan Bueno De Mesquita, and Alexander Wolitzky. Deterrence with imperfect attribution. *American Political Science Review*, 114(4): 1155–1178, 2020.

Alex Barrachina, Yair Tauman, and Amparo Urbano. Entry and espionage with noisy signals. *Games and economic behavior*, 83:127–146, 2014.

Michelle R Garfinkel and Stergios Skaperdas. Economics of conflict: An overview. *Handbook of defense economics*, 2:649–709, 2007.

Gene M Grossman and Elhanan Helpman. Trade wars and trade talks. *Journal of Political Economy*, 103(4):675–708, 1995.

Trygve Haavelmo. A study in the theory of economic evolution. Technical report, 1954.

SJ Ho. Extracting the information: espionage with double crossing. *Journal of Economics*, 93(1):31–58, 2008.

Michael Keen and Kai A Konrad. The theory of international tax competition and coordination. *Handbook of public economics*, 5:257–328, 2013.

Kathryn Merrick, Medria Hardhienata, Kamran Shafi, and Jiankun Hu. A survey of game theoretic approaches to modelling decision-making in information warfare scenarios. *Future Internet*, 8(3):34, 2016.

Wallace E Oates and Paul R Portney. The political economy of environmental policy. In *Handbook of environmental economics*, volume 1, pages 325–354. Elsevier, 2003.

Ralph Ossa. Trade wars and trade talks with data. *American Economic Review*, 104(12):4104–46, 2014.

Robert Powell. Guns, butter, and anarchy. *American Political Science Review*, 87(1):115–132, 1993.

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

David E Sanger. *The perfect weapon: War, sabotage, and fear in the cyber age*. Broadway Books, 2019.

Thomas C Schelling. The strategy of conflict. 1960.

Eilon Solan and Leeat Yariv. Games with espionage. *Games and Economic Behavior*, 47(1):172–199, 2004.

"Victims of Microsoft hack" 2021. Victims of Microsoft hack scramble to plug security holes. CBS News, March 2021. URL `https://www.cbsnews.com/news/microsoft-hack-victims-plug-security-holes/`.

Merrill E Whitney and James D Gaisford. An inquiry into the rationale for economic espionage. *International Economic Journal*, 13(2):103–123, 1999.

Pierre Yared. A dynamic theory of war and peace. *Journal of Economic Theory*, 145(5):1921–1950, 2010.

A Full Definition of Payoff Functions

Given strategies A and D, the Attacker's expected payoff $V_A(A, D)$ immediately after infiltration has most recently become possible satisfies[11]

$$V_A(A,D) = \left[\lim_{\tau\to\infty} A(\tau)\right] \mathbb{E}_{A(\tau)|\tau<\infty}\Bigg[e^{-r\tau}\Bigg\{-C_A + \left[\lim_{\tau'\to\infty} D(\tau')\right]$$
$$\mathbb{E}_{D(\tau')|\tau'<\infty}\Bigg[\int_0^{\tau'}\left[\left(1-e^{-r\tau''}\right)\frac{b}{r} + e^{-r\tau''}[B + e^{-r\Delta}V_A(A,D)]\right]\lambda e^{-\lambda\tau''}\cdot d\tau''$$
$$+ e^{-\lambda\tau'}\left[\left(1-e^{-r\tau'}\right)\frac{b}{r} + e^{-r(\tau'+\Delta)}V_A(A,D)\right]\Bigg]$$
$$+\left[1-\lim_{\tau'\to\infty} D(\tau')\right]\int_0^{\infty}\left[\left(1-e^{-r\tau''}\right)\frac{b}{r} + e^{-r\tau''}[B + e^{-r\Delta}V_A(A,D)]\right]\lambda e^{-\lambda\tau''}\cdot d\tau''\Bigg\}\Bigg]$$

The Defender's expected payoff $V_D(A, D)$ immediately after infiltration has most recently become possible satisfies

$$V_D(A,D) = \left[\lim_{\tau\to\infty} D(\tau)\right]\mathbb{E}_{D(\tau)|\tau<\infty}\Bigg\{A(\tau)\mathbb{E}_{A(\tau')|\tau'<\tau}\Bigg[e^{-r\tau'}\Bigg(\int_0^{(\tau-\tau')}\Bigg[(1-e^{-rs})-\frac{\ell}{r}$$
$$+ e^{-rs}[-L - C_D + e^{-r\Delta}V_D(A,D)]\Bigg]\lambda e^{-\lambda s}\cdot ds$$
$$+ e^{-\lambda(\tau-\tau')}\left[(1-e^{-r(\tau-\tau')}) - \frac{\ell}{r} + e^{-r(\tau-\tau')}[-C_D + e^{-r\Delta}V_D(A,D)]\right]\Bigg)\Bigg]$$
$$+ [1-A(\tau)]e^{-r\tau}[-C_D + V_D(A,D)]\Bigg\}$$
$$+\left[1-\lim_{\tau\to\infty} D(\tau)\right]\left[\lim_{\tau'\to\infty} A(\tau')\right]\mathbb{E}_{A(\tau')|\tau'<\infty}\Bigg[e^{-r\tau'}\int_0^{\infty}\Bigg[(1-e^{-rs})-\frac{\ell}{r}$$
$$+ e^{-rs}[-L - C_D + e^{-r\Delta}V_D(A,D)]\Bigg]\lambda e^{-\lambda s}\cdot ds\Bigg]$$

B Omitted Proofs

B.1 Proof of Proposition 1

Proof. Given that the Defender never clears ($D(\tau) = 0$), if the Attacker's expected payoff $V_A(\tau)$ from infiltrating at any time τ is ≥ 0, then $V_A(0) \geq 0$ as

[11] In these expressions, we assume that the Defender always clears after a signal arrival.

well (because the signal process is memoryless, waiting must reduce payoffs due to discounting). This payoff is

$$\begin{aligned} V_A(0) &= \int_0^\infty \left[\left(1 - e^{-r\tau}\right) \frac{b}{r} + e^{-r\tau}[B + V_A(0) + C_A] \right] \lambda e^{-\lambda\tau} \cdot d\tau - C_A \\ &= \frac{b}{r} + \frac{\lambda}{r+\lambda} \left(-\frac{b}{r} + B + V_A(0) + C_A \right) - C_A \\ \Leftrightarrow V_A(0) &= \frac{b + \lambda B - rC_A}{r} \leq 0 \end{aligned}$$

Since the payoff of never infiltrating ($A(\tau) = 0$) is 0, this is a best response. Given that the Attacker never infiltrates, clearing always gives the Defender a negative payoff $-C_D$, whereas not clearing gives payoff 0, so the Defender is best responding with $D(\tau) = 0$. □

B.2 Proof of Proposition 2

Proof. Using the same payoff derived in the proof of Proposition 1, Condition 1 not holding strictly implies that $V_A(0) \geq 0$. Note again that delay only reduces payoffs, so always infiltrating immediately ($A(\tau) = 1$) is a best response.

Since the Attacker is always known by the Defender to be in the system (because $A(\tau) = 1$), if clearing is ever payoff increasing relative to not clearing, then clearing immediately must also be payoff increasing. The Defender's payoff $V_D(0)$ from a strategy of clearing immediately is

$$\begin{aligned} V_D(0) &= e^{-r\Delta} V_D(0) - C_D \\ \Leftrightarrow V_D(0) &= -\frac{C_D}{1 - e^{-r\Delta}} \end{aligned}$$

The expected payoff from never clearing is

$$V_D = \int_0^\infty \left[(1 - e^{-r\tau}) - \frac{\ell}{r} + e^{-r\tau}[-L + V_D] \right] \lambda e^{-\lambda\tau} \cdot d\tau = -\frac{\ell + \lambda L}{r}$$

Since $V_D \geq V_D(0)$ is equivalent to Condition 2, $D(\tau) = 0$ is a best response. □

B.3 Proof of Lemma 1

Proof. $p(\tau)$ derives from the Attacker's strategy and the signal distribution via Bayes' Rule:

$$p(\tau) = \frac{Ae^{-\lambda\tau} + [1 - A]A^+(\tau) - [1 - A]\int_0^1 A^+(\tau + \frac{\ln(\epsilon)}{\lambda}) \cdot d\epsilon}{(Ae^{-\lambda\tau} + [1 - A]A^+(\tau) - [1 - A]\int_0^1 A^+(\tau + \frac{\ln(\epsilon)}{\lambda}) \cdot d\epsilon) + [1 - A][1 - A^+(\tau)]}$$

The final two terms of the numerator were derived as follows: Since the probability of a signal not arriving between times t and τ is $e^{-\lambda[\tau - t]} = \int_0^\infty \mathbb{1}\{t > \tau + \frac{\ln(\epsilon)}{\lambda}\} \cdot d\epsilon$, the second two terms of the numerator come from $t(\omega), \omega \in \Omega$

being the induced random variable of next infiltration time with sample space Ω and probability measure P

$$
\begin{aligned}
&(1-A)A^+(\tau)\mathbb{E}[e^{-\lambda[\tau-t(\omega)]}|0<t(\omega)\leq\tau] \\
&=(1-A)A^+(\tau)\int_{\{\omega\in\Omega|0<t(\omega)\leq\tau\}} e^{-\lambda[\tau-t(\omega)]}dP(\omega|0<t(\omega)\leq\tau) \\
&=(1-A)A^+(\tau)\int_{\{\omega\in\Omega|0<t(\omega)\leq\tau\}}\int_0^\infty \mathbb{1}\{t(\omega)>\tau+\frac{\ln(\epsilon)}{\lambda}\}\cdot d\epsilon\cdot dP(\omega|0<t(\omega)\leq\tau) \\
&=(1-A)A^+(\tau)\int_0^\infty\int_{\{\omega\in\Omega|0<t(\omega)\leq\tau\}} \mathbb{1}\{t(\omega)>\tau+\frac{\ln(\epsilon)}{\lambda}\}\cdot dP(\omega|0<t(\omega)\leq\tau)\cdot d\epsilon \\
&=(1-A)A^+(\tau)\int_0^1\left[\frac{A^+(\tau)-A^+(\tau+\frac{\ln(\epsilon)}{\lambda})}{A^+(\tau)}\right]\cdot d\epsilon \\
&=(1-A)A^+(\tau)-(1-A)\int_0^1 A^+(\tau+\frac{\ln(\epsilon)}{\lambda})\cdot d\epsilon
\end{aligned}
$$

For the indifference condition to hold, $p(\tau)$ must be a constant function. In particular, since $p(0)=A$, $p(\tau)=A$ for all τ. Thus,

$$A(1-A^+(\tau))=Ae^{-\lambda\tau}+[1-A]A^+(\tau)-[1-A]\int_0^1 A^+(\tau+\frac{\ln(\epsilon)}{\lambda})\cdot d\epsilon$$

Taking the Laplace transform ($\mathcal{L}(\cdot)$) of both sides yields

$$
\begin{aligned}
&\Leftrightarrow\frac{A}{s}-A\mathcal{L}(A^+(\tau))=\frac{A}{s+\lambda}+(1-A)\mathcal{L}(A^+(\tau))-(1-A)\int_0^1\mathcal{L}(A^+(\tau+\frac{\ln(\epsilon)}{\lambda}))\cdot d\epsilon \\
&\Leftrightarrow\frac{A}{s}-A\mathcal{L}(A^+(\tau))=\frac{A}{s+\lambda}+(1-A)\mathcal{L}(A^+(\tau))-(1-A)\mathcal{L}(A^+(\tau))\int_0^1 e^{\frac{\ln(\epsilon)}{\lambda}s}\cdot d\epsilon \\
&\Leftrightarrow\frac{A}{s}-A\mathcal{L}(A^+(\tau))=\frac{A}{s+\lambda}+(1-A)\mathcal{L}(A^+(\tau))-(1-A)\frac{\lambda}{s+\lambda}\mathcal{L}(A^+(\tau)) \\
&\Leftrightarrow A\lambda=s(s+\lambda)\mathcal{L}(A^+(\tau))-(1-A)s\lambda\mathcal{L}(A^+(\tau)) \\
&\Leftrightarrow\lambda A=s(s+\lambda A)\mathcal{L}(A^+(\tau)) \\
&\Leftrightarrow\mathcal{L}(A^+(\tau))=\frac{1}{s}\frac{\lambda A}{s+\lambda A}
\end{aligned}
$$

Taking the inverse Laplace transform yields

$$A^+(\tau)=1-e^{-\lambda A\tau}, \text{a.e.}$$

Since this is continuous, and the solution is required to be non-decreasing, we can dispense with the "almost everywhere" and conclude that the strategy is an exponential CDF. □

B.4 Proof of Lemma 2

Proof. By definition, $\forall \tau$,

$$V_A(\tau) = \int_0^\infty \left\{ \int_0^t \left[(1-e^{-rs})\frac{b}{r} + e^{-rs}(B + e^{-r\Delta}V_A(0)) \right] \lambda e^{-\lambda s} \cdot ds \right.$$
$$\left. + e^{-\lambda t}\left[(1-e^{-rt})\frac{b}{r} + e^{-r(t+\Delta)}V_A(0) \right] \right\} \frac{d(\tau+t)}{1-D(\tau)} \cdot dt$$
$$+ \frac{1-\lim_{\tau\to\infty} D(\tau)}{1-D(\tau)} \int_0^\infty \left[(1-e^{-rt})\frac{b}{r} + e^{-rt}(B + e^{-r\Delta}V_A(0)) \right] \lambda e^{-\lambda t} \cdot dt - C_A$$

Substituting $V_A(0) = 0$ and $V_A(\tau) = 0$ yields

$$0 = \int_0^\infty \left\{ \int_0^t \left[(1-e^{-rs})\frac{b}{r} + e^{-rs}B \right] \lambda e^{-\lambda s} \cdot ds + e^{-\lambda t}\left[(1-e^{-rt})\frac{b}{r} \right] \right\} \frac{d(\tau+t)}{1-D(\tau)} \cdot dt$$
$$+ \frac{1-\lim_{\tau\to\infty} D(\tau)}{1-D(\tau)} \int_0^\infty \left[(1-e^{-rt})\frac{b}{r} + e^{-rt}B \right] \lambda e^{-\lambda t} \cdot dt - C_A$$
$$\Leftrightarrow 0 = \left[\frac{r}{r+\lambda}\frac{b}{r} + \frac{\lambda}{r+\lambda}B \right] \left[\int_0^\infty \left[1 - e^{-(r+\lambda)t} \right] \frac{d(\tau+t)}{1-D(\tau)} \cdot dt + \frac{1-\lim_{\tau\to\infty} D(\tau)}{1-D(\tau)} \right] - C_A$$
$$\Leftrightarrow \int_0^\infty e^{-(r+\lambda)t} \frac{d(\tau+t)}{1-D(\tau)} \cdot dt - \frac{1-\lim_{\tau\to\infty} D(\tau)}{1-D(\tau)} = \frac{b + \lambda B - (r+\lambda)C_A}{b+\lambda B}$$

The right hand side does not depend on τ, so it is necessary for the left hand side to not depend on τ. Let κ be defined as the right hand side above. Setting the derivative of the LHS equal to 0, we get

$$0 = \int_0^\infty e^{-(r+\lambda)t} \left[\frac{d'(\tau+t)[1-D(\tau)] + d(\tau)d(\tau+t)}{[1-D(\tau)]^2} \right] \cdot dt - \frac{1-\lim_{\tau\to\infty} D(\tau)}{1-D(\tau)} \frac{d(\tau)}{1-D(\tau)}$$
$$\Leftrightarrow 0 = \int_0^\infty e^{-(r+\lambda)t} d'(\tau+t) \cdot dt + h(\tau) \left[\int_0^\infty e^{-(r+\lambda)t} d(\tau+t) \cdot dt - [1 - \lim_{\tau\to\infty} D(\tau)] \right]$$

Integrating the first term by parts yields

$$\Leftrightarrow 0 = -d(\tau) + (r+\lambda) \int_0^\infty e^{-(r+\lambda)t} d(\tau+t) \cdot dt + h(\tau) \left[\int_0^\infty e^{-(r+\lambda)t} d(\tau+t) \cdot dt - [1 - \lim_{\tau\to\infty} D(\tau)] \right]$$
$$\Leftrightarrow \frac{h(\tau) - (r+\lambda)[1-\lim_{\tau\to\infty} D(\tau)]}{(r+\lambda) + h(\tau)} = \int_0^\infty e^{-(r+\lambda)t} \frac{d(\tau+t)}{1-D(\tau)} \cdot dt - \frac{1-\lim_{\tau\to\infty} D(\tau)}{1-D(\tau)}$$

The RHS is κ and thus does not depend on τ. As a result, the hazard rate does not depend on τ. Given that $D(0) = 0$, D must be an exponential CDF (and thus, $\lim_{\tau\to\infty} D(\tau) = 1$). Solving for the rate parameter $\rho = h(\tau)$ yields

$$\Leftrightarrow \rho = (r+\lambda)\frac{\kappa}{1-\kappa}$$
$$= \frac{b + \lambda B - (r+\lambda)C_A}{C_A}$$

□

B.5 Proof of Proposition 4

Proof. **Attacker's Strategy:**

Suppose that there exists $\tau^* > 0$ such that $A(\tau^*) - \lim_{\epsilon \to 0^+} A(\tau^* - \epsilon) > 0$ (i.e., a mass point). This implies that the Defender's beliefs satisfy $p(\tau^*) - \lim_{\epsilon \to 0^+} p(\tau^* - \epsilon) > 0$, i.e., they jump upwards at τ^*. By Equation (1), this implies that either the Defender must clear by no later than τ^* (i.e., $D(\tau^*) = 1$) or beliefs have not yet risen high enough to justify clearing (i.e., $D(\tau^*) = 0$). In the former case, either τ^* is off the equilibrium path (i.e., there exists $\epsilon > 0$ such that $D(\tau^* - \epsilon) = 1$), or the Defender has a mass point at τ^*, and the Attacker cannot be best responding, because infiltrating at τ^* must give a lower payoff than infiltrating at $\tau^* + \delta$ for some small $\delta > 0$. In the latter case, the Attacker cannot be best responding, as there is a profitable deviation to $A(\tau) = 1, \forall \tau$ (i.e., infiltrate immediately).

Defender's Strategy:

Suppose that there exists $\tau^* \geq 0$ such that $D(\tau^*) - \lim_{\epsilon \to 0^+} D(\tau^* - \epsilon) > 0$ (when $\tau^* = 0$, take $\lim_{\epsilon \to 0^+} D(\tau^* - \epsilon)$ to be $= 0$). Then, we must also have $\lim_{\epsilon \to 0^+} V_A(\tau^* + \epsilon) - V_A(\tau^*) > 0$ (recall that $V_A(\tau)$ is the Attacker's continuation payoff conditional on state N from infiltrating at τ). This implies that for some $\delta > 0$, every $\tau \in (\tau^* - \delta, \tau^*]$ is not in the support of the Attacker's strategy. If $\tau^* > 0$, then the Defender would profit from deviating to the pure strategy of clearing at time $\tau^* - \delta$, because $p(\tau^* - \delta) > p(\tau^*)$ (beliefs drop as signals fail to arrive, absent any new infiltrations). If $\tau^* = 0$, then $\tau = 0$ is not in the support of the Attacker's strategy $\Rightarrow p(\tau^*) = 0$, so the Defender is not best responding by clearing at τ^*. □

B.6 Proof of Proposition 5

Proof. **Attacker's Strategy:**

Suppose that there exists $[\tau_1, \tau_2)$ that is not in the support of the Attacker's strategy and is on the equilibrium path of play. Then, $p(\tau_2) < p(\tau_1)$. If $D(\tau_1) = 1$, this interval would be off the path of play, so the Defender must not strictly prefer clearing to waiting at τ_1. Thus, the Defender prefers not clearing at any time in $[\tau_1, \tau_2)$. Then, there exists $\epsilon > 0$ such that $[\tau_1, \tau_2 + \epsilon)$ is not in the support of the Attacker's strategy, because infiltrating just after τ_1 (say, at $\tau_1 + \delta$ for small enough $\delta > 0$) must then yield a higher payoff than infiltrating at a time in $[\tau_2, \tau_2 + \epsilon)$. This argument repeats ad infinitum by replacing τ_2 with $\tau_2 + \epsilon$. Since $\tau_2 - \tau_1$ increases with each repetition, ϵ need not decrease with between repetitions (the gains from infiltrating at the start of the interval rather than at the end are even greater). Thus, this implies that $[\tau_1, \infty)$ is not in the support of the Attacker's or Defender's strategies. However, then the Attacker is not best responding, as infiltrating at some time $\tau_1 + \epsilon$ for small $\epsilon > 0$ must yield the maximum possible payoff.

Defender's Strategy:

Suppose that there exists $[\tau_1, \tau_2)$ that is not in the support of the Defender's strategy and is on the equilibrium path of play (i.e., $D(\tau_1) < 1$). Then, (τ_1, τ_2)

must not be in the support of the Attacker's strategy, because for any infiltration time $\tau \in (\tau_1, \tau_2)$, the Attacker would receive a higher payoff from infiltrating at time $\tau' = \frac{\tau_1+\tau}{2} \in (\tau_1, \tau)$, as the Defender does not clear between τ' and τ. However, (τ_1, τ_2) not being in the support of the Attacker's strategy was already ruled out in the previous part of the proof.

These two parts combined also imply that the equilibrium path of play is $\tau \in [0, \infty)$. Suppose instead that there exists finite τ^* such that $D(\tau^*) = 1$. In fact, let τ^* be the greatest lower bound of τ for which $D(\tau) = 1$. Then, conditional on no clearing by $\tau^* - \delta$ for small $\delta > 0$, the Attacker is certain that clearing must occur within the next δ units of time. If δ is small enough, then the Attacker must not infiltrate in $[\tau^* - \delta, \tau^*)$ (the brief infiltration is not worth the costs). But this interval is on the equilibrium path of play, contradicting first part of the proof. □

B.7 Proof of Result 1

Proof. We begin with the non-existence part of the proposition. Suppose $|\mathcal{C}| > 1$. Consider the partition into components. Then, Equation (6) becomes

$$\mu_\Lambda \geq \frac{1}{|\mathcal{C}| - 1} \left[\sum_{I \in \mathcal{C}} Q(I) - W \right]$$

For the largest component $I^{|\mathcal{C}|}$, $Q(I^{|\mathcal{C}|}) \geq W$. Therefore,

$$\Rightarrow \mu_\Lambda \geq \frac{1}{|\mathcal{C}| - 1} \sum_{I \in [\mathcal{C} \setminus \{I^{|\mathcal{C}|}\}]} Q(I)$$

However, each $Q(I) \geq 1$, so this implies $\mu_\Lambda \geq 1$, i.e., the systems are not worth infiltrating. We have assumed that this is false, so the "infiltrate all" equilibrium does not exist.

Now, suppose that $|\mathcal{C}| = 1$. We prove existence in two steps, which together show that the RHS of this inequality is at its largest when the partition is into singleton sets, i.e., $m = n$ and $|I_k| = 1$ for all $k = 1, \ldots m$.

First, we prove that for any fixed m, the RHS of the inequality is maximized when $m-1$ of the partition elements are singletons. Consider the strictly convex, strictly decreasing function $f(x) = \frac{1}{x}$. The LHS equals $\sum_{k=1}^{m} f(C_A - \frac{|I_k|-1}{2}d)$. WLOG, let I_1 be a largest partition element, and consider any other partition element $I_k, k \neq 1$. Removing one element from I_k and adding it to I_1 increases the first term by $f(C_A - \frac{|I_1|}{2}d) - f(C_A - \frac{|I_1|-1}{2}d)$ and decreases the k'th term by $f(C_A - \frac{|I_k|-1}{2}d) - f(C_A - \frac{|I_k|-2}{2}d)$. Since these changes both arise from the same magnitude change in the argument ($\frac{1}{2}d$), and $f(x)$ is strictly convex and strictly decreasing, the increase in the first term must exceed the decrease in the k'th term, resulting in an increase in the LHS. These changes can repeatedly be applied for all k until all but I_1 is a singleton.

Next, we prove that $m = n$ maximizes the RHS of the inequality. For any $m \in \{1, \ldots, n-1\}$, increasing m by 1 raises the RHS if the following holds (note

that all but one of the partition elements is a singleton)

$$\frac{1}{C_A} + \frac{1}{m-1}\left[\frac{1}{C_A - \frac{n-m}{2}d} - \frac{1}{C_A - \frac{n-1}{2}d}\right] \leq \frac{1}{C_A} + \frac{1}{m}\left[\frac{1}{C_A - \frac{n-m-1}{2}d} - \frac{1}{C_A - \frac{n-1}{2}d}\right]$$
$$\Leftrightarrow \frac{m-1}{m}\frac{1}{C_A - \frac{n-m-1}{2}d} + \frac{1}{m}\frac{1}{C_A - \frac{n-1}{2}d} - \frac{1}{C_A - \frac{n-m}{2}d} \geq 0$$

The first two terms form a convex combination of values of a convex function, so Jensen's Inequality implies that they are no smaller than $\frac{1}{C_A - \frac{(m-1)(n-m-1)+(n-1)}{2m}d} = \frac{1}{C_A - \frac{n-m}{2}d}$. The inequality holds, so $m = n$ must maximize the RHS of the inequality.

Combining these two results gives us the largest possible value of the RHS of the inequality: $\frac{1}{n-1}[n - W]$. □

B.8 Proof of Proposition 7

Proof. For every $I \in \mathcal{C}$, let p_I be a probability of infiltrating that cluster, conditional on infiltration (i.i.d. across infiltrations). We will try to construct an equilibrium where an Attacker's strategy consists of $A(\tau)$ and these p_I's.

This class of equilibrium raises an additional consideration for the Defender's strategy. If the Defender checks a system and finds that the Attacker has not infiltrated that system, the Attacker may still have already infiltrated other systems. Thus, a checking but not clearing event cannot "reset the clock" at $\tau = 0$. However, we will focus on an equilibrium where it is as if the clock resets after a check. After a system $i \in I$ has been checked (where I is in the support of the Attacker's strategy), the Attacker will immediately infiltrate I with probability Ap_I. This will bring the Defender back to the point of indifference with regards to both timing of checking and which set of systems to check.

The Defender is approximately indifferent between checking immediately and waiting in systems I if

$$\begin{aligned}
&[Ap_I e^{-r\Delta} + [1 - Ap_I]]V_D(N) - [1 + Ap_I(n-1)]C_D \\
&\approx A\sum_{J\in\mathcal{C}} p_J \int_0^{d\tau} \{(1 - e^{-rt})[-\frac{\ell_J}{r}] \\
&+ e^{-rt}[-L(J) - nC_D + e^{-r\Delta}V_D(N)]\}\lambda(J)e^{-\lambda(J)t} \cdot dt \\
&+ (1 - e^{-\lambda\cdot d\tau})[1 - A]e^{-r\cdot d\tau}[V_D(N) - C_D] \\
&+ e^{-\lambda\cdot d\tau}\{(1 - e^{-r\cdot d\tau})[-A\sum_{J\in\mathcal{C}} p_J\frac{\ell_J}{r}] \\
&+ e^{-r\cdot d\tau}[[Ap_I e^{-r\Delta} + [1 - Ap_I]]V_D(N) - [1 + Ap_I(n-1)]C_D]\},
\end{aligned}$$

Taking the first-order Taylor approximation, rearranging, dividing by $d\tau$, and

taking the limit as $d\tau \to 0$ yields

$$\begin{aligned} & r\left[[Ap_I e^{-r\Delta} + [1 - Ap_I]]V_D(N) - C_D\right] \\ & = A[p_I(n-1)rC_D - \sum_{J \in \mathcal{C}} p_J[\ell_J + \lambda(J)L(J)]] \end{aligned}$$

This may be solved for p_I as a function that does not depend on I, thus $p_I = \frac{1}{|\mathcal{C}|}$ for all $I \in \mathcal{C}$.

As before, beliefs are held constant at A with $A^+(\tau) = 1 - e^{-\frac{1}{|\mathcal{C}|}\sum_{\mathcal{I} \in \mathcal{C}} \lambda(I) A \tau}$. Let $\overline{A} \equiv \frac{A}{|\mathcal{C}|}$ be the infiltration intensity for a single cluster. The Defender's payoff is equal to that of checking a cluster immediately, so

$$\begin{aligned} & V_D(N) = [\overline{A}e^{-r\Delta} + (1 - \overline{A})]V_D(N) - [1 + \overline{A}(n-1)]C_D \\ \Leftrightarrow & V_D(N) = -\frac{[1 + \overline{A}(n-1)]C_D}{\overline{A}(1 - e^{-r\Delta})} \end{aligned}$$

Let $\mathcal{L} \equiv \sum_{I \in \mathcal{C}}[\ell_I + \lambda(I)L(I)]$ be the sum of losses across the sets of systems in the support. Plugging this into the indifference condition yields

$$\begin{aligned} & -\frac{[1 + \overline{A}(n-1)]rC_D}{\overline{A}(1 - e^{-r\Delta})} = -\overline{A}\mathcal{L} \\ \Leftrightarrow & \overline{A}^2(1 - e^{-r\Delta})\mathcal{L} - \overline{A}(n-1)rC_D - rC_D = 0 \\ \Leftrightarrow & \overline{A} = \frac{(n-1)rC_D + \sqrt{(n-1)^2 r^2 C_D^2 + 4(1 - e^{-r\Delta})\mathcal{L}rC_D}}{2(1 - e^{-r\Delta})\mathcal{L}} \end{aligned}$$

The Defender's strategy consists of independent "racing exponentials," exactly as in Section 3. Indifference for the Attacker requires expected payoff of 0 for every cluster. Thus, for each $I \in \mathcal{C}$, $\rho_I = \hat{R}(I)$. This is an equilibrium if it is also possible to drive below zero the payoff from infiltrating every other set of systems I': $\rho_{I'} \geq \hat{R}(I'), \forall I'$.

At a minimum, we must have for all $I \in \mathcal{C}$, $\mu_\Lambda(I) \geq \frac{1}{|I|-1}[|I| - W(\{I\})]$. This comes from applying Result 1 to each component. If this condition fails, then for some $I \in \mathcal{C}$, there exists a partition $\{I_k\}_{k=1}^m$ of I such that $\sum_{k=1}^m \hat{R}(I_k) > \hat{R}(I)$, so there are no clearing rates consistent with equilibrium. Furthermore, there are no other equilibrium conditions. We need not consider deviations I' containing systems from multiple components. The reasoning is similar to the negative multiple components result of Result 1. Breaking I' up into its intersections with components must result in at least one subset $I'' \subset I'$ such that $Q(I'') \geq Q(I')$, thus creating even tighter equilibrium conditions. □