Strategic Cyberwarfare

Lillethun, Erik and Sharma, Rishi

Colgate University

18 June 2024

Strategic Cyberwarfare

Erik Lillethun* Rishi Sharma*

June 18, 2024

Abstract

This paper develops a theoretical model of cyberwarfare between nations, focusing on the factors that determine the severity and outcomes of cyber conflicts. We introduce a two-country model where nations invest in offensive or defensive cyber capabilities across networked systems. We show that resource expenditure intensifies when players' effective values are similar, which can help explain the rise of cyberwarfare. We explore the implications of network structures, showing how larger attack surfaces worsen outcomes for defenders. Additionally, we investigate the impact of private cyber defence provision, and find that centralized policies may either improve or exacerbate cyber conflict.

JEL: C72, D74, D85, F5

*Department of Economics, Colgate University. We thank participants at the Stony Brook International Conference on Game Theory and the Colgate University Economics Brown Bag for useful comments on this work. We thank the Colgate University Research Council for support.

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.[1] 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 investment?

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 two-country model where each country can attack or defend a number of systems. These systems are arranged in networked groups, where access to one system can provide access to additional systems. The Attacker and Defender both make investments that determine the probability of infiltrating a particular set of systems. We model the investment process as similar to an all-pay auction, where the Attacker is able to enter a system if it makes a larger investment

[1]There have been many subsequent Russian cyberattacks against Ukraine related to the Russo-Ukrainian War, especially in 2022. These are well documented in Przetacznik [2022]. However, the 2015 attack has been the largest successful attack against the Ukrainian power grid.

than the Defender. Having entered a particular system in a network, the Attacker can then access other systems that are connected downstream.

Our model sheds light on several factors that determine the severity of cyberwarfare. As in other all-pay auction contexts, resource expenditure is most severe when the "effective values" of the two players are closer to each other. The effective value depends on the loss (for the Defender) or benefit (for the Attacker) from infiltration, the player's marginal cost of infiltration, and also on network features. Our model provides a natural explanation for the rise in cyberwarfare. If the Defender is the higher value player – which we argue is a reasonable assumption for some of the highest stakes cyberwarfare – greater progress in Attacking vs. Defending ability will lead to convergence of valuation and therefore more intensive conflict. This account is also broadly consistent with the fact that smaller and technologically less advanced countries have found it worthwhile to engage in cyberwarfare with larger countries such as the US.

An extension of our model introduces both an immediate benefit at infiltration – capturing the benefit from immediate information acquisition – and additional benefit when the Attacker is able to remain in the system longer-term. The latter captures both the benefit of receiving updated information as well as any potential option value benefits (e.g. shutting off critical infrastructure in case of a future conflict). However, the Attacker does not reap long-term benefits if they are exposed and eliminated from the systems. The Defender incurs corresponding short- and long-term losses from being infiltrated. This extension allows for a distinction between a more "patient" long-term infiltration vs. a "major" attack that seeks to extract benefit immediately at a greater risk of exposure. As expected, a patient infiltration is more likely when the Attacker's short-run benefits are low relatively to its long-run benefit. More surprisingly, we find that the Defender may actually be better off in a major attack equilibrium than in a more patient long-term infiltration equilibrium. Hence, a conflict with a patient nation state could be worse for the Defender than an attack from a more noticeably destructive adversary.

Our analysis emphasizes the role of network structure in determining cyber-

warfare outcomes. We highlight the role of the size of the "attack surface" or equivalently, the number of "attack vectors" available to the Attacker. When the Attacker is able to attack a network through many channels, this effectively implies a relatively lower Attacker cost vs. Defender cost and will hurt the Defender. It will also amplify the intensity of conflict in the case where the Defender is higher value. The increasing importance of digital connection between websites, platforms and service providers is likely to be a factor that amplifies the size of attack surfaces and therefore worsen cyberwarfare. We also find that "star" network structures, where a single central system gives access to peripheral systems that are not directly connected to each other, tend to be beneficial for the Defender. This is because such structures allow the systems to be interlinked in a way that tends to minimize the size of the attack surface.

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 cyberdefences of their own systems, which in turn provide access to other downstream networks. This type of setup is often relevant in practice, with a prominent case being the 2021 hack of Microsoft Exchange Server, which provides services to many downstream networks. We examine a variation on our model where the defensive investments are made by private entities that consider losses to their own systems but not other interlinked systems. We find two possible outcomes in this externality version of the model. The first – and more expected outcome – is that private provision leads to an underinvestment in cyber defence that hurts Defenders collectively relative to a centralized provision. This outcome takes place when the Defender is the higher effective value player. This outcome would provide support for policies that impose requirements or standards on private entities, or possibly promise retaliation when private entities are attacked. Examples of such policy initiatives in the U.S. are measures designed to impose national standards on information sharing by companies, such as the Cybersecurity Information Sharing Act of 2015.

A second – and more surprising – possibility is that public provision may

actually lead to an overprovision of cyber investment. This is likely when the Attacker is the higher value player or there are many attack vectors. In this case, the presence of a centralized Defender invites a stronger investment from the Attacker and the ensuing negative sum competition means the Defender ends up spending more resources without much corresponding improvement in actual defence. This result would suggest that centralized regulation should ideally be imposed only in situations where the systems in question have very high associated losses and there are relatively few attack vectors, and may be unnecessary or harmful if applied too broadly, including to systems that have less severe systematic losses associated to the Defender.

Finally, the star network is especially sensitive to the externality problem. The Defender's advantage from the star network vanishes as the externality becomes more extreme. The advantage of a star network relies on the entity controlling the central system internalizing the benefits to mount a vigorous defence. When this is not the case, the effect of the externality is similar in some respects to an increase in the number of attack vectors.

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.[2] 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.

Our work is also connected to the existing literature on network defence. Relative to this literature, we ask distinct questions about the causes and consequences of more or less severe cyberwarfare – questions that are more analogous to those asked in the economics of conflict referenced above. That

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

said, our analysis draws on and contributes to the network defence literature. Our work specifically relates to two strands of this literature.[3] Our analysis of public vs. private provision of cyberdefence relates to past work on externalities in network defence, which builds on Kunreuther and Heal [2003] and Varian [2004]. Much of this work emphasizes the possibility of reduced defensive investment due to the positive externalities from investment – an outcome that is also possible in our framework. Acemoglu et al. [2016] use a Stackelberg "Defender first" model to show that a negative externality that leads to overinvestment is also possible because private Defenders may have an incentive to divert Attackers towards other Defenders. In our analysis of the simultaneous-move game – which would capture the Attacker's uncertainty about the Defender's choices – we uncover a novel point related to over- and under-investment: under certain conditions, *public* provision can lead to overprovision of defensive investments because the presence of public provision that internalizes the network externalities can induce more aggressive behavior on part of the Attacker.

Our work also touches on themes that arise in the study of optimal defensive network structure. Goyal and Vigier [2014] study optimal network structure with a common defender and find that a star network is optimal under a wide range of circumstances. Although we define a star network somewhat differently – as a directed network with links only pointing from the center to the periphery – our results are also favorable towards a star network structure. However, our equilibrium is significantly different, and our reasons pertain to limiting the effective number of attack vectors. For related reasons, we show that the star structure is especially sensitive to the network externality problem. Insufficient defence of the central system is an especially exploitable vulnerability, akin to having more attack vectors. This latter mechanism is distinct from the tradeoffs identified in Cerdeiro et al. [2017] – who study optimal network design with private defenders – which are rooted in the pres-

[3]Fedele and Roner [2022] provide a recent review of this literature in economics. We focus our discussion here on the most connected portions of this literature. See also Roy et al. [2010] or Merrick et al. [2016] for a review of some of the related literature in computer science and information security studies, which is more engineering focused.

ence of Stackelberg-type negative externalities as in Acemoglu et al. [2016], in addition to the positive externalities.

Finally, we believe that our analysis of the case where there are both short- and long-run benefits and losses to the Attacker and Defender is new to the literature. The resulting tradeoff between a major immediate attack vs. a more patient long-term infiltration is an important aspect of cyberconflict, especially as it pertains to nation states.

The paper proceeds as follows: Section 2 introduces our model, Section 3 analyzes the cluster network structure, Section 4 examines the star network and compares it to the cluster. Section 5 analyzes externality problems that arise when private entities control defensive choices, and systems are organized in either a cluster or star network. Finally, Section 6 concludes.

2 Model

In the general model, there is a set $N = \{1, \ldots, n\}$ of systems partitioned into a set of m groups $M = \{N_1, \ldots, N_m\}$, where $n_i \equiv |N_i|$. A group is a set of systems that shares a single method of the Attacker gaining access. In the cybersecurity vernacular, this is a set of systems sharing the same single "attack vector." These attack vectors are the ways of directly gaining access to a system. Though we assume that there is only one direct attack vector per system (i.e., the set of groups is a partition), our results would not significantly change if there were multiple attack vectors per system (i.e., groups can overlap).

Initially, the Attacker and Defender simultaneously choose investments $a_i \geq 0$ and $d_i \geq 0$, respectively, for each group $i = 1, \ldots, m$. If $a_i > d_i$, the Attacker succeeds in gaining access to every system in group N_i. If $a_i < d_i$, the Attacker fails to gain access to any systems in N_i. If $a_i = d_i > 0$, the Attacker succeeds with probability $\frac{1}{2}$. If $a_i = d_i = 0$, then the tie-breaking rule must be equilibrium dependent; we will specify the tie-breaking rule in each special case we analyze. The marginal costs of investment are c_A and c_D. Thus, this first stage of the game is similar to an all-pay auction (per group of systems).

At the end of the first stage, let $I \subseteq N$ refer to the initial set of systems that the Attacker has access to.

After gaining access to an initial set of systems, it becomes easier and less costly to attack related systems. This creates an additional indirect way of gaining access to systems. The systems are networked according to directed graph $G = (N, E)$ (N is the set of vertices and E is the set of edges), where $(i, j) \in E$ if and only if there is a link from System i to System j, where $i \neq j$. The network defines which secondary systems may be attacked immediately with probability 1 and at zero cost by virtue of successfully infiltrating another system. If $i \in I$ and $j \notin I$, but $(i, j) \in E$, then the Attacker may also gain access to j for free. In principle, one could then use j's links to gain access to a third system k, but this is not relevant in the network types in this paper. $\mathcal{C} = \{C_1, \ldots, C_{|\mathcal{C}|}\}$ is the set of all components of G. In our context, a component is a maximal weakly connected subgraph. That is, if you replace the directed links with undirected links, a component is a largest possible subgraph that has a path between any two of its systems. We use $C(i) \in \mathcal{C}$ to refer to the component containing System i and C^V to the set of systems in component C. We assume that each group is a subset of a component: for all N_i, there exists $C \in \mathcal{C}$ such that $N_i \subseteq C^V$. In other words, systems in different components are unrelated enough that somewhat distinct methods are needed to directly infiltrate each. For example, although phishing attacks can be used to gain login credentials to access either system, the systems do not have any known users in common, so distinct phishing attacks must be used for each system.

The Attacker reaps a benefit of $\mathcal{B}_i$ if it is able to attack System i and the Defender incurs losses of $\mathcal{L}_i$. In the baseline version of the model, the Attacker always has an incentive to attack all systems it gains access to. In Section 3.3, we examine an extension where the Attacker may choose to limit the number of systems it attacks. This implies a final set of attacked systems F, which includes all systems in I and also all systems with links coming from systems in I. F determines the players' payoffs. The Attacker's payoff is $\sum_{i \in F} \mathcal{B}_i - c_A \sum_{j=1}^{m} a_j$, and the Defender's payoff is $-\sum_{i \in F} \mathcal{L}_i - c_D \sum_{j=1}^{m} d_j$.

We note here two comparisons of our model with papers in the existing literature on network defence. First, our model is a simultaneous-move game, in line with much of the literature but in contrast to the defender-first Stackelberg-type games in Acemoglu et al. [2016] and Cerdeiro et al. [2017]. The simultaneous-move game captures the fact that the Attacker is unlikely to have perfect information about the Defender's investment choices, especially given the long history of nations keeping their capabilities secret. Second, by studying a two-player game, we abstract here from the multiple attribution problem emphasized in Baliga et al. [2020].

The timeline of the game is as follows: The Attacker and Defender simultaneously choose their investments a_i, d_i for each group. Then, Nature determines the systems I the Attacker initially gains access to according to these investments, which indirectly determines the final set of attacked systems F. Finally, players receive their payoffs. We focus on the perfect Bayesian equilibria of this game.

For each equilibrium in this paper, it is useful to think of one player as being the high value player (denoted H) and the other player as being the low value player (denoted L). For example, in the case where the Attacker is the high value player, and the Defender is the low value player, we replace all A subscripts with H and all D subscripts with L. This allows us to describe the equilibrium more concisely. The high value player is determined not by a naïve comparison of $\mathcal{B}_i$'s and $\mathcal{L}_i$'s, but rather on a comparison between two "effective values" w_A^* and w_D^* which will depend on the network structure. This lets us define a useful class of equilibria:

Definition 1

A strategy profile has the ***all-pay auction form*** *if for some effective values w_A^* and w_D^*, the following hold:*

1. *If $w_A^* < w_D^*$, then $H = D$. If $w_A^* > w_D^*$, then $H = A$.*
2. *With probability $p_L = \frac{w_H^* - w_L^*}{w_H^*}$, the low value player does not invest in any system.*

All equilibria in this paper have the all-pay auction form. The interesting distinctions lie in the determinants of w_A^* and w_D^*, the distributions determining which systems are invested in, and the investment distributions.

3 Cluster Network

Suppose that G is a cluster graph (i.e. each component is a complete graph, which we call a "cluster").[4]. To simplify and focus attention on the network structure, we also assume that systems are otherwise identical: for all pairs of systems i, j, $\mathcal{B}_i = \mathcal{B}_j = \mathcal{B}, \mathcal{L}_i = \mathcal{L}_j = \mathcal{L}$. Of particular interest are the total benefits and losses for a Cluster j: $\mathcal{B}_j^* \equiv n_j\mathcal{B}$ and $\mathcal{L}_j^* \equiv n_j\mathcal{L}$ (where we will drop the j when it is clear from the context). Let M_j be the set of indices of groups of systems in Cluster j: $M_j = \{i \in \{1, \ldots, m\} | N_i \subseteq C_j^V\}$. Let $m_j = |M_j|$. This m_j is the number of distinct methods (i.e., attack vectors) by which the Attacker may gain access to the entire cluster (or as many systems as they desire). We call the resulting game the "cluster graph game."

We now provide a formal collection of results in Proposition 2, which will be followed by a proof and then a discussion of implications.

3.1 Main Cluster Result

To sharpen the analogy to all-pay auctions, we first redefine the players' values so that the marginal cost is 1 (as in an auction where the investments are dollars). The Attacker's value for Cluster j is $v_A^* = \frac{\mathcal{B}^*}{c_A}$, and the Defender's value is $v_D^* = \frac{\mathcal{L}^*}{c_D}$. By dividing each payoff by the player's marginal cost, the payoffs now have the form "total value − total investment." For the case of $a_i = d_i = 0$ (where i is in Cluster j), we break the tie in favor of the Attacker if and only if $m_j v_A^* > v_D^*$.

Proposition 1

There exists a perfect Bayesian equilibrium of the cluster graph game where

[4]Our results in this section would continue to hold under a generalization where each component is a strongly connected subgraph. We simply make the complete graph assumption for expositional simplicity.

for every Cluster j, strategies have the all-pay auction form and the following hold:

1. $w_A^* = v_A^*, w_D^* = \frac{v_D^*}{m_j}$
2. *Conditional on investing at all, the Attacker invests in exactly one group i, where i is chosen uniformly.*
3. *Conditional on investing at all, the Defender invests i.i.d. in all groups.*
4. *Each player, conditional on investing in Group i, chooses an investment level distributed $U[0, w_L^*]$.*

Proof. First, note that the players' payoffs may be separated into terms for each cluster, where each cluster's payoff does not interact with choices regarding other clusters. Therefore, strategies may be independent across clusters, and we need consider only cluster-specific payoffs.

First, suppose that $w_D^* > w_A^*$. Assuming that the Defender uses independent uniform distributions with upper bounds u_i, the Attacker's payoff in Cluster j is

$$\left[1 - \prod_{i \in M_j} \left(1 - \frac{a_i}{u_i}\right)\right] n_j \mathcal{B} - c_A \sum_{i \in M_j} a_i$$

The derivative of this with respect to any a_i is

$$\prod_{k \in M_j \setminus \{i\}} \left(1 - \frac{a_k}{u_k}\right) \frac{n_j \mathcal{B}}{u_i} - c_A$$

The Attacker is indifferent conditional on $a_k = 0, \forall k \in M_j \setminus \{i\}$ if and only if $u_i = \frac{n_j \mathcal{B}}{c_A} = w_A^*$. Then, the Attacker strictly prefers $a_i = 0$ whenever $a_k > 0$ for any $k \neq i$.

Let p_A be the probability that the Attacker invests in none of the groups in Cluster j. Now, assuming that the Attacker only ever invests in one group in the cluster, that this group is chosen uniformly, and that conditional on investing, the Attacker uses the same uniform distribution as the Defender

(the same u_i), the Defender's expected payoff in Cluster j is

$$-(1-p_A)\left(\frac{1}{m_j}\sum_{i\in M_j}\left(1-\frac{d_i}{u_i}\right)\right)n_j\mathcal{L}-c_D\sum_{i\in M_j}d_i$$

The derivative with respect to any d_i is $(1-p_A)\frac{n_j\mathcal{L}}{m_j u_i}-c_D$. Indifference is then equivalent to $p_A=\frac{w_D^*-w_A^*}{w_D^*}\in[0,1]$. Thus, the Attacker and Defender are both best responding, and this is an equilibrium.

Now, suppose that $w_A^*>w_D^*$. We now have $p_A=0$, so the Defender is indifferent if $\frac{n_j\mathcal{L}}{m_j u_i}=c_D$, so $u_i=\frac{v_D^*}{m_j}=w_D^*$. Given that p_D is the probability that the Defender invests in none of the groups in Cluster j, the Attacker's expected payoff is

$$\left\{p_D+(1-p_D)\left[1-\prod_{i\in M_j}\left(1-\frac{a_i}{u_i}\right)\right]\right\}n_j\mathcal{B}-c_A\sum_{i\in M_j}a_i$$

The derivative with respect to any a_i is

$$(1-p_D)\prod_{k\in M_j\setminus\{i\}}\left(1-\frac{a_k}{u_k}\right)\frac{n_j\mathcal{B}}{u_i}-c_A$$

Conditional on $a_k=0,\forall k\neq i$, the Attacker is indifferent if

$$(1-p_D)\frac{n_j\mathcal{B}}{u_i}=c_A\Leftrightarrow p_D=\frac{w_A^*-w_D^*}{w_A^*}$$

Whenever $a_k>0$ for some $k\neq i$, then the Attacker strictly prefers $a_i=0$.

Also, note that the Defender cannot benefit from deviating from zero investment to some small positive investment(s), even though the Attacker will invest 0 in some groups and wins ties. After such a deviation, they will still almost surely fail to prevent the Attacker from infiltrating some group in the cluster. Since infiltration of any single group leads to the same final outcome, there is no benefit from such a deviation. That is, there is no discontinuity at zero investment for the Defender. □

3.2 Discussion of Proposition 1

Proposition 1 shows that the Attacker in equilibrium will attack at most a single system, though the identity of the system is randomized. When the Attacker is the lower value player, it may not attack any system. The Defender, in general will defend all systems, though it may defend no system with positive probability when it is the low value player.

We see here that although there is only one way of directly attacking a system, there are additionally $m_j - 1$ indirect ways. Therefore, the overall equilibrium strategies for each system depend on m_j. In cybersecurity terminology, each system has an "attack surface" (set of all attack vectors) of size m_j. This larger attack surface gives the Attacker an advantage: They need only attack via one attack vector at a time, randomizing which one they attack, but the Defender must simultaneously defend all attack vectors. This multiplier allows the Attacker to effectively outspend the Defender more cheaply.

The players' payoffs are as follows:

Attacker	High Value	$(v_A^* - \frac{1}{m_j} v_D^*) c_A$
	Low Value	0
Defender	High Value	$-m_j v_A^* c_D$
	Low Value	$-v_D^* c_D$

As in other auction theory contexts, our results from Proposition 1 require thinking separately about the case where the Defender is the higher value player and where the Attacker is the higher value player, where the Attacker's value is multiplied by the number of attack vectors. A natural example for the Defender being the higher value player may be when the cluster in question has few attack vectors and corresponds to critical infrastructure. In this case, the potential losses to the Defender may be extremely large relative to the benefits for the Attacker. This would especially be the case if the Attacker in question is more interested in information gathering but not in a major escalation of hostilities. The case where the Defender is the higher value player is likely to capture many of the most prominent cases of cyberattacks, especially ones that are of broad public salience.

An example of the opposite case (i.e. the Attacker being higher value) could be one where the cluster contains a massive amount of useful information for the Attacker but where the information is not extremely sensitive for the Defender and so is not excessively damaging. We consider several implications of Proposition 1 taking both of these cases into account. The Attacker could also be considered higher value if the systems involve critical infrastructure but there are also many attack vectors (i.e., they are too well connected to other systems).

There are two natural metrics capturing the nature of cyberwarfare in this context. The first is the intensity of conflict, i.e. the amount of investment by both the Attacker and the Defender. This reflects the total resources being devoted to cyber conflict. Second, we are also interested in the welfare effect of cyberwarfare. The overall welfare effect can be somewhat difficult to interpret in this context because the Attacker's welfare obviously includes the benefit it derives from attacking systems, which we might see as a less "legitimate" benefit on societal grounds, at least in many potential scenarios.

A major question of interest in connection to cyberwarfare in practice is the consequence of improvements in defensive vs. offensive cyberwarfare capabilities. From a policy perspective, it is often suggested that countries should prioritize improvements in their defensive rather than offensive capabilities. A natural expectation might be that offensive improvements will lead to more severe cyberwarfare whereas defensive improvements will reduce the intensity of conflict. Our results imply that this is indeed the case when the Defender is the higher value player. In this case, the offensive improvements – captured by lower c_A (implying higher w_A^*) cause both sides to invest more. Improvements in defensive capability – lower c_D – make the Attacker less likely to invest.

The case where the Defender is higher value than the Attacker is probably a natural assumption for some of the highest stakes cyberwarfare applications (e.g. critical infrastructure, highly classified state secrets). In these cases, faster improvements in offensive vs. defensive capabilities globally would provide a natural explanation for the rise of cyber conflict over time in the context of our model. This account would also explain why smaller and less technolog-

ically advanced countries are able to engage more aggressively in cyberconflict with countries such as the U.S., i.e. improvements in offensive relative to defensive capabilities would bring the effective valuations of Attackers closer to Defenders' valuations.

These patterns are reversed in the case where the Attacker is the higher value player. An improvement in defensive capability now causes *more* investment by both the Attacker and the Defender. Improvement in offensive capabilities reduce the investment by the Defender. As w_L^* rises, the upper bound of the investment range rises, increasing both players' expected investments. When the Attacker is the higher value player, improvements in offensive capabilities (or worsening of defensive capabilities) actually reduce the intensity of conflict. The key intuition for the outcomes in both the higher value Attacker and higher value Defender cases is that cyberwarfare is most intense when the values of the Attacker and Defender are closer to each other.

The welfare effects of changes in offensive and defensive capability track closely the investment effects. When the Defender is higher value, total welfare improves when defensive capability improves and worsens when offensive capability improvements. In both cases, this is driven by decreases or increases in wasteful conflict investments. Conversely, in the Attacker high value case, total welfare improves when offensive capability improves or when defensive capability worsens.

3.3 Extension with Short vs. Long-Run Benefits and Losses

We now consider an extension where the Attacker does not necessarily choose to attack each system in a cluster even though it has gained access. In practice, an Attacker may face a tradeoff between immediately maximizing its short-term benefit by attacking a system vs. maintaining a more patient long-term presence that allows greater benefit to be reaped over time. The Attacker now makes a non-trivial choice about the final number of systems to attack in Cluster j, which we denote f_j. In the baseline version of the model, this

choice would be trivial, i.e. $f_j = |C_j^V|$. For this extension, we specifically assume Cluster j immediately gives the Attacker guaranteed short run benefit $f_j b$ and the Defender immediately suffers losses $f_j \ell$, where $b \geq 0$ and $\ell \geq 0$. These represent actions that may be taken immediately on attacking a system, such as stealing data that already resides there. Note that we have assumed here equal benefits and losses for each system. Based on short-run benefits alone, there is no reason to attack less than the entire cluster.

After the Attacker has chosen f_j and reaped the short run benefits, there is a passive detection process that may lead to the discovery of the Attacker in a system. For example, IT staff may notice system usage at unusual times of day and investigate. For each attacked system $i \in F$, there is an independent probability $q \in (0, 1)$ that the Attacker's presence will be discovered in that system. Being discovered in System i also reveals the Attacker's presence in any other system in the same component $C(i)$. This prevents the Attacker from reaping any of the long-run benefits of remaining in the systems in $C(i)$.

If the Attacker is *not* discovered in any system in Cluster j, then they reap additional benefits $f_j B$ and the Defender incurs additional losses $f_j L$, where $B \geq 0$ and $L \geq 0$. These could represent the effect of the Attacker maintaining a long term presence in the systems, such as passive intelligence gathering or the option value of being able to damage the functioning of the systems should the need arise. To ensure that all systems are desirable to attack and defend in isolation, we assume that $b + B > 0$ and $\ell + L > 0$. If on the other hand the Attacker is discovered in one system in a cluster through the passive monitoring process, the Defender can infer their presence in other systems in that cluster. Therefore, if the Attacker is discovered in any system in a cluster, they will be quickly eliminated from that entire cluster and there will be no long-run gains or losses within that cluster.

For any number f_j of attacked systems in Cluster j, the probability that the Attacker is not discovered in Cluster j is $(1 - q)^{f_j}$.[5] If the Attacker is

[5]Note that the Attacker only puts themselves at risk for detection by actually attacking a system and not by merely gaining access to it. Therefore, f_j could be less than the number of intitally infiltrated systems in Cluster j.

not discovered in Cluster j, then the Attacker gets Cluster j payoff $f_j(b+B)$ (ignoring investment costs), and the Defender gets Cluster j payoff $-f_j(\ell+L)$. If the Attacker is detected in Cluster j, these payoffs are instead just $f_j b$ and $f_j \ell$, respectively.

The timeline of the game is now as follows: The Attacker and Defender simultaneously choose their investments a_i, d_i for each group. Then, Nature determines the systems I the Attacker initially gains access to according to these investments. The Attacker observes this initial set of systems and then chooses a final set of systems F to attack (implying f_j for each j). Finally, Nature determines whether the Attacker is discovered in each cluster according to probability q, and then payoffs are realized. We focus on the perfect Bayesian equilibria of this game.

For expositional simplicity, we also adopt an approximation of the model where f_j may be any real number in the interval $[0, |C_j^V|]$. The probability of not being discovered still has the same form $(1-q)^{f_j}$, and the payoff functions are otherwise unchanged.[6] This is not an essential assumption, but when f_j would optimally not be an integer, the rounding conditions would be quite complicated and not very insightful.

Under these assumptions, Proposition 1 still holds as long as we slightly redefine the expected benefits and losses: The Attacker's Cluster j expected payoff (excluding investment costs) is $\mathcal{B}_j^* \equiv [b + (1-q)^{f_j^*} B] f_j^*$, and the Defender's Cluster j expected payoff is $\mathcal{L}_j^* \equiv [\ell + (1-q)^{f_j^*} L] f_j^*$, where f_j^* is the equilibrium choice of f_j. Now, the only difference is the decision about the number of systems attacked in equilibrium f_j^*, the main results for which are provided in Proposition 2. We call this game the extended cluster graph game.

Let $\sigma_A \equiv \frac{b}{B}$ be the Attacker's short-run/long-run benefit ratio. We use $W(\cdot)$ to refer to the principal branch of the Lambert W (product logarithm) function, which is always an increasing function in the relevant domain. In Proposition 2, e always refers to Euler's number, which appears because of the

[6] We could obtain essentially identical results without this continuous approximation if we specified the model with a continuum of systems within each cluster. However, thinking about linkages between systems in such a model is much less intuitive and does not provide additional insight.

exponential form of the "no detection" probability $(1-q)^{f_j}$.

Proposition 2

In every perfect Bayesian equilibrium of the extended cluster graph game and for each Cluster j, then conditional on successfully gaining access to a system in the cluster ($I \cap C_j^V \neq \emptyset$), the equilibrium number of systems attacked, f_j^*, satisfies the following:

1. *If $\sigma_A \geq \frac{1}{e^2}$, then $f_j^* = |C_j^V|$, i.e., "attack all."*
2. *Otherwise,*
   (a) *If $\left[\frac{W(-\sigma_A e)-2}{\ln(1-q)} - |C_j^V|\right]\sigma_A - \left[\frac{e^{W(-\sigma_A e)-1}}{\ln(1-q)} + (1-q)^{|C_j^V|}|C_j^V|\right] < 0$, then $f_j^* = |C_j^V|$.*
   (b) *If $\left[\frac{W(-\sigma_A e)-2}{\ln(1-q)} - |C_j^V|\right]\sigma_A - \left[\frac{e^{W(-\sigma_A e)-1}}{\ln(1-q)} + (1-q)^{|C_j^V|}|C_j^V|\right] > 0$, then $f_j^* = \frac{W(-\sigma_A e)-1}{\ln(1-q)}$.*

Proof. See Appendix A.1. □

With this modification to the model, another potential determinant of conflict intensity is the probability of detection, q. A lower q means that more patient long-term infiltration is more feasible whereas a very high value for q means that the Attack is unlikely to persist beyond the initial infiltration. It is straightforward to show that both $\mathcal{B}_j^*$ and $\mathcal{L}_j^*$ shrink when q rises (in all cases of Proposition 2). The same is then true of w_A^* and w_D^* when applying this to Proposition 1. The result is smaller investment sizes. The impact on the extensive margin (probability of investment) is ambiguous, as it will interact with the number of attack vectors. In both the high value and low value cases, the Defender's payoff rises. The Attacker is neither better nor worse off in their low value case. The effect is ambiguous in their high value case, as both their benefits and costs decrease, and this also interacts with the number of attack vectors.

In addition to the levels of investment and welfare outcomes, we can now think about the number of systems infiltrated, f_j^*, which depends on various factors. First, note that there is a potential for both a continuous change (if

it is the interior solution) and a discontinuous change (if it switches to the "attack all" solution). If every system in the cluster is attacked, we call it a "major attack," and otherwise it is called a "minor attack." The size of the attack increases as the Attacker's short run/long run benefit ratio σ_A increases:

Corollary 1
In the perfect Bayesian equilibrium of the extended cluster graph game described in Proposition 2, for every Cluster j, f_j^* is a weakly increasing function of σ_A, and it is strictly increasing in the parameter range that results in a minor attack.

Proof. See Appendix A.2. □

A large σ_A Attacker, such as a cyber criminal or hacktivist, will tend to attack more systems. Their attitude is "get in, steal/wreck what we can, and get out." A nation state primarily interested in the option value of destroying an enemy's critical systems would have a small σ_A and tend to attack fewer systems. This is likely in nations preparing for possible war. However, if the nations are already at war, immediate damage may be more valuable (high σ_A), leading the Attacker to attack more systems. The effect for a nation state primarily interested in intelligence gathering is more ambiguous, and depends on how much intelligence is obtainable immediately versus over time.

Second, there is a non-monotonic relationship between cluster size and major attacks:

Corollary 2
In the perfect Bayesian equilibrium of the extended cluster graph game described in Proposition 2, there exists a cutoff size $\overline{N}$ such that any Cluster j of size $|C_j^V|$ may have a minor attack if any only if $\frac{W(-\sigma_A e)-1}{\ln(1-q)} \leq |C_j^V| \leq \overline{N}$.

Proof. See Appendix A.3 □

For small clusters, the size of the cluster is a binding constraint. When that constraint relaxes enough, the Attacker may settle into an interior number of systems. However, the short run gains from infiltrating all systems eventually

becomes so large that it is worth mostly sacrificing the long run gains to maximize the short run gains. This shows that the number of systems attacked is a weakly increasing function of the cluster size, as shown in Figure 1. First, the number of systems attacked increases, as the cluster size is binding. Then, it remains constant at the local maximum. Finally, it jumps up to the "attack all" corner solution and proceeds to increase more from there.

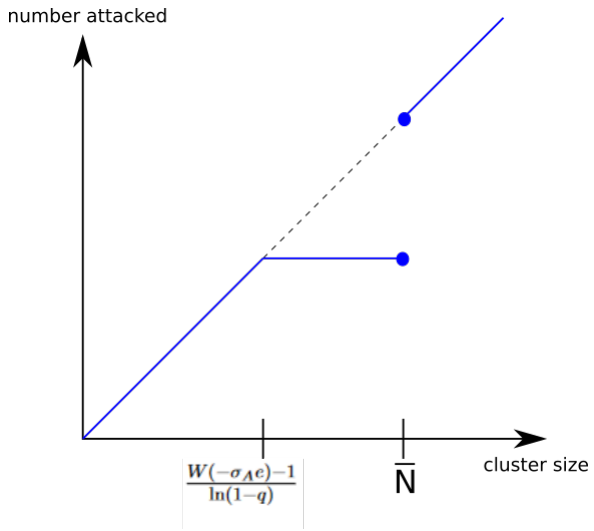

Figure 1: The Attacker's optimal number of systems to attack in a cluster (vertical axis) for each cluster size (horizontal axis).

Finally, we compare both investment magnitude and Defender welfare between major attacks and minor attacks:

Corollary 3

Consider any very small parameter change that causes the equilibrium to change from a minor attack to a major attack in Cluster j (so the Attacker's second stage payoff is nearly unchanged). Then,

1. *If $\sigma_A < \sigma_D$, then w_D^* jumps upward. If the Defender was initially the low value player, then the Defender's payoff drops and both players tend to invest more (otherwise there is almost no change).*

2. *If $\sigma_A > \sigma_D$, then w_D^* jumps downward. If the Defender was initially the*

low value player, then the Defender's payoff rises and both players tend to invest less (otherwise there is almost no change).

The proof of Corollary 3 is trivial. The Defender's condition for preferring the minor attack to the major attack in Cluster j is the exact opposite of the Attacker's condition, but with b and B replaced with ℓ and L. Therefore, if the Attacker is nearly indifferent, the Defender's preference is strict if and only if $\sigma_A \neq \sigma_D$.

Part 1 of Corollary 3 is fairly obvious, but Part 2 is rather surprising: the Defender may prefer for the Attacker to attack more systems. When $\sigma_A > \sigma_D$, the Defender cares more about the long run, relative to the Attacker. Thus, when the Attacker decides they are willing to put their long-run benefits at risk by infiltrating all systems, this benefits the Defender, because detection will be more likely, preventing the long-run losses. If the Defender is a patient actor like a nation state, and the Attacker is an organization that is less likely to persist long into the future, like a terrorist organization, then the Defender might actually benefit from major attacks. Combining Corollary 2 with Corollary 3 suggests that this type of Defender may benefit from connecting more systems together in the same cluster, even though it enables more indirect attacks from the Attacker.

4 Star Network

Section 3 studied a setting where each system is symmetric in terms of its position in the network. In this section, we assume that systems are organized into a star network, with a single central system that provides access to all peripheral systems. This model is especially relevant when there is a single, centralized service used by many other systems. A good example of this type of attack is the 2015 attack on the U.S. Office of Personnel Management, which gave the attackers sensitive data on millions of government employees, potentially creating vulnerabilities at other government agencies. In this model, System 1 is a central system, and Systems $2, \ldots, n$ are peripheral systems. The set of edges is then $E = \{(1, i) | i \in \{2, \ldots, n\}\}$.

We assume that there are n groups: $N_i = \{i\}$ for all $i \in \{1, 2, \ldots, n\}$. That is, every system may be targeted independently from one another. For simplicity, we assume that the central system has no intrinsic value, and all of the peripheral systems have the same positive values: $\mathcal{B}_1 = \mathcal{L}_1 = 0$, and $\mathcal{B}_i = \mathcal{B}, \mathcal{L}_i = \mathcal{L}$ for all $i \geq 2$.

The players' per system value/cost ratios are $v_A = \frac{\mathcal{B}}{c_A}$ and $v_D = \frac{\mathcal{L}}{c_D}$. In the $a_i = d_i = 0$ case, the Attacker wins the tie if and only if $2v_A > v_D$. Although we model only one star, we expect the results would be the same if there were multiple stars, on a star by star basis (as we found to be true on a cluster by cluster basis in the previous section). We call the resulting game the "star graph game." We omit the proof of the following proposition, because it is essentially a special case of Proposition 5 in Section 5.

Proposition 3

There exists a perfect Bayesian equilibrium of the star graph game with the all-pay auction form, where the following hold:

1. *Conditional on any I, the Attacker always infiltrates the maximum set of peripheral systems: $1 \in I \Rightarrow (N \setminus \{1\}) \subseteq F$ and $1 \notin I \Rightarrow (I \setminus \{1\}) \subseteq F$*

2. *$w_A^* = (n-1)v_A, w_D^* = \frac{1}{2}(n-1)v_D$*

3. *Conditional on investing at all, the Attacker invests in only the central system or only the peripheral systems (each case with equal probabilities), in the latter case investing i.i.d. in all peripheral systems.*

4. *Conditional on investing at all, the Defender invests in all systems, independently across systems.*

5. *Each player, conditional on investing in the central system, chooses an investment level distributed $U[0, w_L^*]$. Conditional on investing in a peripheral system, their investment level is distributed $U[0, \frac{w_L^*}{n-1}]$.*

4.1 Discussion of Proposition 3

All-Pay Auction Logic As with the cluster results, in both cases (Parts (2a) and (2b)), there is a low value player (L) and a high value player (H), determined by the inequality $w_A^* = (n-1)v_A \lesseqgtr \frac{1}{2}(n-1)v_D = w_D^*$. Raising v_L or decreasing v_H causes an increase in investment on the extensive margin (reducing the probability of no investment by L). Investment increases on the intensive margin (increasing the upper bounds of the supports of the investment distributions) may be caused by an increase in v_L or, in the case of the central system, an increase in the number of peripheral systems.

There is always higher expected investment (for each player) in the central system compared with each individual peripheral system. However, the aggregate expected investment in the periphery equals the expected investment in the central system. This is because there is nothing inherently more costly or less effective about infiltrating the peripheral systems directly. If the Defender made it more costly to infiltrate peripheral systems directly (via high investments there), then the Attacker would respond by only attacking the central system, and the Defender would not be best responding (wasting investments on the ignored periphery).

The payoffs in the star example are as follows:

Attacker	High Value	$(n-1)(v_A - \frac{1}{2}v_D)c_A$
	Low Value	0
Defender	High Value	$-(n-1)2v_Ac_D$
	Low Value	$-(n-1)v_Dc_D$

First, note that the Defender never benefits from an increase in their value (v_D), whereas the Attacker may benefit from an increase in theirs (v_A). For the Defender, an increase in v_D has two possible effects: larger potential losses and greater competitiveness in bidding. In the Defender high value case, these are exactly countervailing, so increasing v_D has no payoff effect. In the Defender low value case, there is no value to the greater competitiveness, since the Defender is still indifferent between investing and giving up. However, the Attacker benefits from an increase in v_A in the high value case, because this

corresponds to larger potential benefits, not losses. On the other hand, both players find that increases in the other player's value have a neutral or negative effect on payoffs, because that other player is increasing their investments.

Both players find that increases in their own costs cause a reduction in payoff, but only in their high value case. In the low value case, they are still held indifferent between investing and giving up, and the payoff at which they give up has nothing to do with their own costs (which are zero).

Comparison with the Cluster Results Perhaps the most interesting distinction between the star results and the cluster results from Section 3 is that the star is effectively a special case of the cluster. Consider the following payoff table for the analogous cluster, which has $n-1$ systems:

Attacker	High Value	$(n-1)(v_A - \frac{1}{m_j} v_D) c_A$
	Low Value	0
Defender	High Value	$-(n-1) m_j v_A c_D$
	Low Value	$-(n-1) v_D c_D$

The payoffs in the star are the same as the payoffs in the cluster when there are two attack vectors ($m_j = 2$). This is not immediately obvious, since there are n groups in the star example, not 2. However, attacking a peripheral system directly only yields $\frac{1}{n-1}$ of the value of the whole star. The Attacker would need to attack all $n-1$ peripheral systems directly to achieve the same effect as attacking the central system. So for attacking the star as a whole, there are effectively only two attack vectors: attacking the central system or attacking peripheral systems directly. Alternatively, one could look at it from the perspective of any single peripheral system. The single system has two attack vectors: being attacked directly or indirectly via the central system. In conclusion, it is a mistake to think of the Attacker's advantage in terms of only the aggregate number of attack vectors when these different attacks achieve different results. The cluster had a special symmetry property, causing all attacks to achieve the same result.

These results illustrate the benefit of a star system from the Defender's

perspective. Compared to other network structures, the star effectively limits the number of attack vectors while still interlinking the systems. It thereby effectively minimizes the size of the attack surface. As discussed already in the relation to the cluster results, a smaller attack surface allows the Defender to expend less defensive investment.

5 Private Defence Investments

In many nations – especially liberal democracies – cybersecurity decisions are not made by a central planner, who internalizes all the 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 cyberdefence 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 type of situation is perhaps best 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).

There is only one difference between the baseline model and the model used in this section: there is more than one Defender. Let K_i be the set of systems controlled by Defender i. Defender i chooses only d_j for each $j \in K_i$ and has payoff depending only on systems in K_i:

$$-\sum_{K \subseteq K_i} \left[Pr(K = F \cap K_i) \sum_{j \in K} \ell_j \right] - \sum_{j \in K_i} c_D d_j$$

We examine in turn both the cluster and star network structures in this private

defender setup.

For this section only, we assume that the Defenders may correlate their investment choices based the outcomes of a signal (public to the Defenders, not observed by the Attacker). That is, we allow for correlated equilibrium (Aumann [1974]) as a slight generalization of Nash equilibrium. In the cluster case, this helps smooth out discontinuities in the Defenders' payoff functions at zero investment. It also allows us to focus on the role of externalities in isolation, as the correlated strategies will rule out any loss to the Defenders due to a failure to coordinate. Since the correlating signal will be a simple "invest/do not invest" signal observed only by the Defenders, we omit the details of the signal and treat the Defenders as if they were one player, but with different payoff functions for different investments. Treating the Defenders as one player also allows us to describe the equilibria as having the all-pay auction form.

5.1 Cluster

In this section, we use the same assumptions as in Section 3. However, each group has a separate Defender, indexed as $1, \ldots, m$, so $K_i = N_i$. We also assume that each Defender is identical, so $|N_i| = \frac{n}{m}$ is the same for all i. We focus on the case of a single cluster. The Attacker's second stage decision is unchanged, so f^* is as before. The individual Defender's value is then $\frac{1}{m}\frac{\mathcal{L}^*}{c_D} = \frac{1}{m}v_D^*$; given the Attacker's uniform randomization over which f^* systems to attack, the Defenders split the loss equally in expectation. In the $a_i = d_i = 0$ case, the Attacker wins the tie if and only if $v_A^* > \frac{1}{m^2}v_D^*$.

Proposition 4

There exists a correlated equilibrium of the cluster graph game with private investments where the strategies have the all-pay auction form, and the following hold:

1. $w_A^* = v_A^*$ *and* $w_D^* = \frac{1}{m^2}v_D^*$
2. *The equilibrium is otherwise equivalent to that in Proposition 1.*

Proof. See Appendix A.4. □

Discussion of Proposition 4 We can collect the players' payoffs for the cluster network under the private defence assumption as follows:

Attacker	High Value	$(v_A^* - \frac{1}{m^2} v_D^*) c_A$
	Low Value	0
Defenders	High Value	$-m \frac{m+1}{2} v_A^* c_D$
	Low Value	$-v_D^* c_D + \frac{v_D^*}{m^2 v_A^*} \frac{m-1}{2} \frac{v_D^*}{m} c_D$

We see that the externality is made more severe by increasing m, as each Defender has a smaller share of the overall value of the cluster. However, this also increases the number of attack vectors, spreading the Defender's investments across more groups while not shrinking aggregate investment. This effect already appears in the public cluster example, which had Attacker high value payoff of $(v_A^* - \frac{1}{m} v_D^*) c_A$. The Attacker's benefit from the externality in particular comes from a reduction in the aggregate investment, $m \frac{1}{m^2} v_D^* = \frac{1}{m} v_D^*$. This benefit increases as m increases. Since $m > 1$, the Defender in their high value case is worse off due to the externality (the public cluster payoff is $-m v_A^* c_D$). This difference does get worse as m rises (more severe externality).

Surprisingly, the Defenders are actually better off due to the externality in their low value case. The corresponding public cluster payoff of $-v_D^* c_D$ equals only the first term of the Defender's low value payoff, the second term being positive. Although the externality causes the Defenders to invest less than their best response if they acted as one, they actually benefit from this in the form of lower costs. The Attacker responds to the less vigorous defence by reducing their own offensive investments, so the Defenders do not necessarily suffer much from increased probability of loss. This beneficial effect is limited to the Defenders' low value case, because the low value determines equilibrium investment sizes. This beneficial effect also mostly shrinks as m rises, because the probability of investment falls ($\frac{v_D^*}{m^2 v_A^*} \downarrow$), meaning investment sizes matter less.

From a policy perspective, these results imply that measures that encourage private defenders to internalize network externalities (e.g. regulations, common standards) may not always be beneficial. They would be beneficial when the Defenders are relatively high value, e.g. perhaps when the systems in question are critical infrastructure or otherwise of deep importance. However, when dealing with systems that are relatively low stakes from the Defender's perspective or having many attack vectors, such policies may increase costs without meaningful improvement in security. Hence, these results suggest that a more targeted attempt to deal with the externality problem could be preferable to an excessively broad policy measure.

5.2 Star

Here, we maintain almost the same assumptions as in Section 4. However, there are now multiple Defenders. The central system Defender (Defender 1) controls the first $k \geq 1$ systems (the central system and $k-1$ of the peripheral systems). There are $n-k$ other Defenders, each controlling just one peripheral system. We index these other Defenders as $k+1$ through n. This allows the central Defender to have some value (as the $k=1$ case is trivial) while maintaining tractability. Let $\rho \equiv \frac{k-1}{n-1}$ be Defender 1's share of the peripheral systems. Also, let $\phi \equiv \frac{\rho}{1+\rho}$ (which is a strictly increasing function of ρ). Note that $\phi \in \left[0, \frac{1}{2}\right]$. Both ρ and ϕ are measures of the central Defender's internalization of the benefits of defence.

Proposition 5
There exists a correlated equilibrium of the star graph game with private investments that has the all-pay auction form, where the following hold:

1. $w_A^* = (n-1)v_A, w_D^* = \phi(n-1)v_D$
2. *Conditional on investing at all, the Attacker invests in only the central system (probability $1-\phi$) or only the peripheral systems (probability ϕ).*
3. *The equilibrium is otherwise equivalent to that in Proposition 3.*

Proof. Consider the case where $v_A < \phi v_D$. The Attacker's expected payoff is the following:

$$\frac{a_1}{u_1}(n-1)v_A + \frac{u_1 - a_1}{u_1}\sum_{i=2}^{n}\frac{a_i}{u_i}v_A - a_1 - \sum_{i=2}^{n} a_i$$

The partial derivative with respect to a_1 is

$$\frac{(n-1)v_A}{u_1} - \frac{1}{u_1}\sum_{i=2}^{n}\frac{a_i}{u_i}v_A - 1$$

Given $a_2 = \ldots = a_n = 0$, indifference for a_1 is achieved when $u_1 = (n-1)v_A = w_A^*$. Then, $a_k > 0$ for any $k \geq 2$ implies that the Attacker strictly prefers $a_1 = 0$. The partial derivative with respect to any a_i for $i \geq 2$ is $\frac{u_1 - a_1}{u_1}\frac{v_A}{u_i} - 1$. Conditional on $a_1 = 0$, indifference is achieved when $u_i = v_A = \frac{w_A^*}{n-1}$. Moreover, if $a_1 > 0$, the Attacker's payoff is maximized when $a_i = 0$.

Let p_A^C denote the Attacker's probability of investing in the central system (conditional on investing at all). Defender 1's expected payoff is the following:

$$-(1-p_A)\sum_{j=2}^{k}\left[(1-p_A^C)\left(1-\frac{d_j}{u_j}\right) + p_A^C\left(1-\frac{d_1}{u_1}\right)\right]c_D v_D - c_D\sum_{j=1}^{k} d_j$$

Their indifference conditions are

$$(1-p_A)p_A^C(k-1)\frac{v_D}{u_1} = 1, (1-p_A)(1-p_A^C)\frac{v_D}{u_j} = 1, \forall j \in \{2,\ldots,k\}$$

Combining these two indifference equations and substituting for u_j yields

$$p_A^C(k-1)\frac{v_D}{u_1} = (1-p_A^C)\frac{v_D}{u_j} \Leftrightarrow p_A^C = \frac{1}{1+\rho} = 1-\phi$$

Then, plugging into the d_1 indifference equation and solving for p_A:

$$(1-p_A)\frac{1}{1+\rho}(k-1)\frac{v_D}{u_1} = 1 \Leftrightarrow p_A = \frac{\phi v_D - v_A}{\phi v_D}$$

Note that Defender i's expected payoff when $i \geq k+1$ has the same form as each term $j = 2, \ldots, k$ of Defender 1's payoff, so these Defenders are also indifferent.

Now, consider the case where $v_A > \phi v_D$. We use p_D to denote the probability that the Defenders do not invest at all. In this case, $p_A = 0$ but either $p_D > 0$. The Attacker's payoff is the following:

$$\left[p_D + (1-p_D)\frac{a_1}{u_1}\right](n-1)v_A + (1-p_D)\left[1-\frac{a_1}{u_1}\right]\sum_{i=2}^{n}\frac{a_i}{u_i}v_A - \sum_{j=1}^{n}a_j$$

The partial derivative with respect to a_1 is the following:

$$(1-p_D)\frac{(n-1)v_A}{u_1} - (1-p_D)\sum_{i=2}^{n}\frac{a_i}{u_i}\frac{v_A}{u_1} - 1$$

When $a_i = 0$ for all $i \geq 2$, indifference is equivalent to $(1-p_D)\frac{(n-1)v_A}{u_1} = 1$. The partial derivative with respect to a_i for $i \geq 2$ is $(1-p_D)\left[1-\frac{a_1}{u_1}\right]\frac{v_A}{u_i} - 1$. When $a_1 = 0$, the indifference condition is $(1-p_D)\frac{v_A}{u_i} = 1$.

The central Defender's payoff is the following:

$$-\sum_{j=2}^{k}\left[(1-p_A^C)\left(1-\frac{d_j}{u_j}\right) + p_A^C\left(1-\frac{d_1}{u_1}\right)\right]v_D - \sum_{j=1}^{k}d_j$$

The indifference conditions are

$$p_A^C\frac{(k-1)v_D}{u_1} = 1, (1-p_A^C)\frac{v_D}{u_i} = 1, \forall i \in \{2, \ldots, k\}$$

The other Defenders j for $j \geq k+1$ have payoffs analogous to any single term $i = 2, \ldots k$ of the central Defender's payoff, so they are also made indifferent when $u_j = u_i$.

This gives us four indifference conditions and four unknowns: p_D, p_A^C, u_1, and u_i. The unique solution is $p_D = \frac{v_A - \phi v_D}{v_A}$, $p_A^C = \frac{1}{1+\rho} = 1-\phi$, $u_1 = (n-1)\phi v_D$, $u_i = \phi v_D$, which is exactly as described in the Proposition. Also, note that Defender 1 has no payoff discontinuity at zero investment, because if they

deviate to very small, positive investments, the Attacker will still almost surely succeed in attacking all $k-1$ of their valuable systems. $\square$

Discussion of Proposition 5 Comparing Proposition 5 to the public star result Proposition 3, there is only one major difference: $\frac{1}{2}v_D$ has been replaced with ϕv_D, which is a lower value. As promised, the public star is a special case of the private star, where $\rho = 1$ (so $k = n$ and $\phi = \frac{1}{2}$). The private investment game favors the Attacker, compared to the public investment game. First of all, the condition where the Attacker is "high value" ($v_A > \phi v_D$) is more likely to hold. Second, in this case the Defenders invest less, which also benefits the Attacker. ϕ is a measure of the extent to which Defender 1 internalizes the benefits of their investment in the central system. When ϕ is nearly $\frac{1}{2}$, we have close to the public investment star. When ϕ is near 0, there is a large externality. Defender 1 invests less in the central system, making all Defenders' peripheral investments less valuable, and so these decrease as well. All Defenders effectively have a lower value when the externality is more extreme. Moreover, the Attacker wants to invest in the easy to attack central system, so they do this with probability $1-\phi > \frac{1}{2}$, which is more than in the public case.

The Attacker's and aggregate Defenders' payoffs are given in the table below:

Attacker	High Value	$(n-1)(v_A - \phi v_D)c_A$
	Low Value	0
Defenders	High Value	$-(n-1)2v_A c_D - (n-k)\frac{1}{2\rho}v_A c_D$
	Low Value	$-(n-1)2\phi v_D c_D - (n-k)(1-\phi)\left[1-\frac{1}{2}\phi\frac{v_D}{v_A}\right]v_D c_D$

As expected, the Attacker's payoff rises in the private case relative to the public case, as $\phi < \frac{1}{2}$. The externality causes the central Defender to be less inclined to invest, so all players' equilibrium investments fall. In the Defender's high value case, we see that they are worse off relative to the public model. The second term $-(n-k)\frac{1}{2\rho}v_A c_D$ reflects the lower payoffs of Defenders $k+1$ through n as they may be attacked via the central system even when they

invest the maximum. As the externality gets worse (ρ gets smaller), this extra loss is even larger.

Surprisingly, the low value Defenders are not worse off due to the externality. The second term $-(n-k)(1-\phi)\left[1-\frac{1}{2}\phi\frac{v_D}{v_A}\right]v_D c_D$ is again the loss to Defenders $k+1$ through n since they may be attacked even when investing the maximum. However, this maximum investment (the first term $-(n-1)2\phi v_D c_D$) is now lower than in the public model. In all models considered in this paper, the equilibrium investment magnitudes (i.e., the severity of the cyber war) are determined by the value of the lower value player(s). The Defenders' externality effectively reduces their value, making the conflict less competitive and reducing the costs of cyber defence in equilibrium.

In fact, rewriting the Defenders' payoff in a form more similar to the public case yields $-(n-1)v_D c_D + (n-1)\frac{1-\rho}{\rho}\frac{1}{2}\phi^2\frac{v_D}{v_A}v_D c_D$. This exceeds the Defender's payoff in the public case (which is just the first term). The second term is a non-monotonic function of ρ, initially equal to 0, then rising, then falling, and finally hitting 0 again when $\rho = 1$ (i.e. no externality). Intuitively, when $\rho = 0$, all Defenders simply give up, since they know the central Defender will invest nothing. Therefore, there are no investment costs to save. Then, as ρ rises, the Defenders are becoming able to deter the Attacker (even sometimes if they invest nothing), so they must benefit. However, the Attacker is responding to this by becoming more likely to target the periphery, thereby becoming less predictable. This reduces the marginal benefit of the increase in ρ. Meanwhile, the competition is becoming more intense and investment costs rise.

A notable distinction between the public star and the private star is that the public star functioned as though there were two attack vectors to the entire star. Although in principle the central system counts for more than a typical attack vector, when the Defender internalizes all of the benefits of defending the center, they defend it that much more aggressively, so it still only effectively counts as one attack vector in equilibrium. However, this is no longer the case with the externality. The central Defender is not investing enough, and this exposes more of the latent vulnerability of having a central system that connects to all other systems. In some respects, this functions as if the

network had more attack vectors. Specifically, since ϕ takes the place of $\frac{1}{2}$ in the equilibrium cases and investment sizes, there are effectively $\frac{1}{\phi} = 1 + \frac{1}{\rho}$ attack vectors. As the externality becomes more significant (ρ becomes smaller), the effective number of attack vectors becomes larger. However, this analogy to attack vectors is imperfect. Adding more attack vectors also adds to the number of types of defensive investments required, so even though investment sizes shrink, the Defender does not benefit. However, making the externality larger shrinks the investment sizes and does not increase the number of investment types, which may benefit the Defender.

5.3 Discussion

There are several observations common to both the cluster and star networks with private investments. The externality resulting from private investments makes the Attacker high value, Defender low value case more likely. In this case, the externality reduces investments of both players on both the intensive (how much to invest) and extensive (invest vs. not invest) margins. This benefits the Attacker, because they may reduce their investment costs without reducing their success probabilities. Surprisingly, it benefits the Defender as well. Being unable to collectively best respond reduces the intensity of the conflict in a way that benefits both parties. In the Attacker low value, Defender high value case, the Attacker is no better off (they are still indifferent about even attempting a cyber attack), and the Defender is worse off. In this case, the externality does not reduce the intensity of the conflict, because the externality reduces the Defenders' individual values, but only the low value player's (the Attacker's) value determines the investment sizes.

There is one primary qualitative difference between the results on the private cluster and private star. As shown in Proposition 1 and Proposition 3, the star is better for the Defender than the cluster (in all but one case). This is because the star effectively has fewer attack vectors. The Attacker is especially inclined to target the central system, but the Defender responds to this in equilibrium and defends it to the point that the center and periphery are

targeted equally often in equilibrium. However, the Defender will only do this if they fully internalize the benefits of defending the center. In the private investments case, the central Defender is not investing enough, and the Attacker is targeting the center more than half the time. The common central system becomes more of a liability as a result of the externality. As the externality becomes more extreme, the advantages of the star vanish.

6 Conclusion

Cyberwarfare between nation states has become increasingly common in recent years. We address several important questions that this phenomena raises. What determines the severity of cyberwarfare outcomes? What is the role of network structures in either worsening or improving outcomes? What is the consequence of private provision of defensive capabilities in networks that link multiple agents? In this paper, we addressed these questions with an Attacker-Defender game theory model.

We identify several factors that can help explain the determinants of cyberwarfare and rationalize trends that have been observed in recent decades. We model cyberwarfare as being similar to an all-pay auction, where the Attacker is able to infiltrate a system if it invests more than the Defender. As in other such auction contexts, the intensity of investment on both sides is greatest when the effective valuations of the two players are close to each other. If the Defender is the higher value player – which may be true in many applications – faster improvements in countries' offensive vs. defensive abilities would drive an increase in conflict intensity. We also find that network structures that tend to minimize the number of attack vectors available to an Attacker tends to be beneficial for the Defender because it allows defensive resources to be used more efficiently. A star network with a single central system that links to multiple peripheral systems tends to be beneficial for the Defender because it limits the number of attack vectors.

We also considered a problem that is common especially in liberal democracies: externalities that arise when private entities are in charge of defending

their own systems. When systems are interlinked, private entities may not internalize the benefit of their investment on downstream systems and so there could be underinvestment in this environment. We find that this is indeed a possibility, consistent with considerations of positive network externalities. However, we find that the opposite outcome is also possible: if the Defender is the low effective value player, public provision may actually lead to overinvestment and reduced welfare for the Defenders as compared to private provision. This is because public provision can lead the Attacker to invest more intensely in order to secure infiltration. Taken together, these results imply that policies that impose more centralized standards or regulations on the private sector may be especially beneficial for systems with large associated losses for the Defender but could be unnecessarily costly if applied to less critical 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.

Daron Acemoglu, Azarakhsh Malekian, and Asu Ozdaglar. Network security and contagion. *Journal of Economic Theory*, 166:536–585, 2016.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.

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

Diego A Cerdeiro, Marcin Dziubiński, and Sanjeev Goyal. Individual security, contagion, and network design. *Journal of Economic Theory*, 170:182–226, 2017.

Alessandro Fedele and Cristian Roner. Dangerous games: A literature review on cybersecurity investments. *Journal of Economic Surveys*, 36(1):157–187, 2022.

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

Sanjeev Goyal and Adrien Vigier. Attack, defence, and contagion in networks. *The Review of Economic Studies*, 81(4):1518–1542, 2014.

Trygve Haavelmo. *A Study in the Theory of Economic Evolution*. North-Holland Publishing Company, 1954.

Howard Kunreuther and Geoffrey Heal. Interdependent security. *Journal of risk and uncertainty*, 26:231–249, 2003.

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.

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

Jakub Przetacznik. Russia's war on Ukraine: Timeline of cyber-attacks. Technical report, European Parliament, June 2022.

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.* Harvard University Press, 1960.

Hal Varian. System reliability and free riding. In *Economics of information security*, pages 1–15. Springer, 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/`.

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

A Omitted Proofs

A.1 Proof of Proposition 2

Proof. For any final set of attacked systems F, the Attacker's expected payoff (excluding the sunk costs of the first stage) is additively separable across clusters, where each Cluster j's term is $f_j b + (1-q)^{f_j} f_j B$, where f_j is the number of systems attacked in Cluster j. The first derivative is $b + (1-q)^{f_j} B + \ln(1-q)(1-q)^{f_j} f_j B$. This is always > 0 when $f_j = 0$, so there is incentive to attack systems. The local maximum, when it exists, solves the first-order condition $b + (1-q)^{f_j} B + \ln(1-q)(1-q)^{f_j} f_j B = 0$. The local maximum is $f_j = \frac{W(-\sigma_A e)-1}{\ln(1-q)}$, where $W(\cdot)$ is the principal branch of the Lambert W (product logarithm) function. This exists if and only if the argument to the W function satisfies $-\sigma_A e \geq -\frac{1}{e} \Leftrightarrow \sigma_A \leq \frac{1}{e^2}$. This corresponds to a local maximum when the inequality is strict. If it holds with equality, it is an inflection point. If it is violated, there is no local maximum, and the solution is the corner solution $f_j = |C_j^V|$. This solution for f_j (if it exists) is guaranteed to be positive. Also, note that this solution for f_j may exceed $|C_j^V|$, in which case the Attacker should attack the entire cluster.

Plugging local maximum f_j into the expected payoff yields

$$\begin{aligned}
&(\frac{W(-\sigma_A e)-1}{\ln(1-q)})b + (1-q)^{\frac{W(-\sigma_A e)-1}{\ln(1-q)}} \frac{W(-\sigma_A e)-1}{\ln(1-q)} B \\
=&(\frac{W(-\sigma_A e)-2}{\ln(1-q)})b - \frac{e^{W(-\sigma_A e)-1}}{\ln(1-q)} B
\end{aligned}$$

This is larger than the expected payoff of attacking the entire cluster if and only if

$$\begin{aligned}
&(\frac{W(-\sigma_A e)-2}{\ln(1-q)})b - \frac{e^{W(-\sigma_A e)-1}}{\ln(1-q)} B - |C_j^V| b - (1-q)^{|C_j^V|} |C_j^V| B > 0 \\
\Leftrightarrow &\left[(\frac{W(-\sigma_A e)-2}{\ln(1-q)}) - |C_j^V|\right] \sigma_A - \left[\frac{e^{W(-\sigma_A e)-1}}{\ln(1-q)} + (1-q)^{|C_j^V|} |C_j^V|\right] > 0,
\end{aligned}$$

which is precisely the condition used in the statement of the Proposition. □

A.2 Proof of Corollary 1

Proof. First, note that $W(-\sigma_A e)$ is negative and strictly decreasing over the parameter range where there is an interior solution, so $f_j^* = \frac{W(-\sigma_A e)-1}{\ln(1-q)}$. Since $\ln(1-q) < 0$, f_j^* strictly increases as σ_A rises.

Now, consider the marginal effect on increasing σ_A on the LHS of the major attack condition (in 2.a of the Proposition):

$$\begin{aligned}
&\left[\frac{W(-\sigma_A e)-2}{\ln(1-q)} - |C_j^V|\right] + \frac{-eW'(-\sigma_A e)}{\ln(1-q)}\sigma_A - \frac{-W'(-\sigma_A e)e^{W(-\sigma_A e)}}{\ln(1-q)} \\
&= \left[\frac{W(-\sigma_A e)-2}{\ln(1-q)} - |C_j^V|\right] + \frac{-e\frac{W(-\sigma_A e)}{-\sigma_A e[1+W(-\sigma_A e)]}}{\ln(1-q)}\sigma_A - \frac{-\frac{W(-\sigma_A e)}{-\sigma_A e[1+W(-\sigma_A e)]}e^{W(-\sigma_A e)}}{\ln(1-q)} \\
&= \frac{W(-\sigma_A e)-1}{\ln(1-q)} - |C_j^V|
\end{aligned}$$

Therefore, whenever the cluster size is not a binding constraint on the local maximum, increasing σ_A decreases the LHS (more conducive to full infiltration). When cluster size is a binding constraint, it will still be a binding constraint after increasing σ_A (the local maximum increases). □

A.3 Proof of Corollary 2

Proof. Let $f_j^{**} > \frac{W(-\sigma_A e)-1}{\ln(1-q)}$ be the local minimum of the Attacker's payoff, if it exists. The marginal effect of increasing cluster size $|C_j^V|$ on the LHS on the major attack condition is $-\sigma_A - (1-q)^{|C_j^V|} - \ln(1-q)(1-q)^{|C_j^V|}|C_j^V|$. This is negative whenever $|C_j^V| > f_j^{**}$. Also, note that the LHS of the major attack condition diverges to $-\infty$ as $|C_j^V| \to \infty$. As a result, there is a cutoff $\overline{N}$, where full infiltration is guaranteed if $|C_j^V| \leq f_j^*$ or if $|C_j^V| \geq \overline{N}$ and that there is guaranteed to be only partial infiltration of f_j^* systems otherwise. □

A.4 Proof of Proposition 4

Proof. First, suppose that $\frac{1}{m}v_D^* \geq mv_A^*$. Assuming that the Defender uses independent uniform distributions, the Attacker's payoff is

$$\left[1 - \prod_{i \in M}\left(1 - \frac{a_i}{u_i}\right)\right]\mathcal{B}^* - c_A \sum_{i \in M} a_i$$

The derivative of this with respect to any a_i is

$$\prod_{k \in M \setminus \{i\}}\left(1 - \frac{a_k}{u_k}\right)\frac{\mathcal{B}^*}{u_i} - c_A$$

The Attacker is indifferent conditional on $a_k = 0, \forall k \in M \setminus \{i\}$ if and only if $u_i = \frac{\mathcal{B}^*}{c_A} = v_A^*$. Then, the Attacker strictly prefers $a_i = 0$ whenever $a_k > 0$ for any $k \neq i$.

Let p_A be the probability that the Attacker invests in none of the groups. Now, assuming that the Attacker only ever invests in one group in the cluster, that this group is chosen uniformly, and that conditional on investing the Attacker uses the same uniform distribution as the Defender, Defender i's expected payoff is

$$-(1 - p_A)\left(\frac{1}{m}\sum_{j \in M}\left(1 - \frac{d_j}{u_j}\right)\right)c_D\frac{1}{m}v_D^* - c_D d_i$$

Indifference with respect to any d_i is

$$(1 - p_A)\frac{c_D v_D^*}{m^2 u_i} - c_D = 0 \Leftrightarrow p_A = \frac{\frac{1}{m}v_D^* - mv_A^*}{\frac{1}{m}v_D^*}$$

Now, suppose that $mv_A^* > \frac{1}{m}v_D^*$. Let p_D be the probability of X (none of the Defenders invest). After signal Y, Defender i's expected payoff is

$$-\frac{1}{m}\sum_{j \in M}\left(1 - \frac{d_j}{u_j}\right)c_D\frac{1}{m}v_D^* - c_D d_i$$

Defender i is indifferent only if $\frac{\frac{1}{m}v_D^*}{mu_i} = 1$. Therefore, $u_i = \frac{v_D^*}{m^2}$. After signal X, the Attacker is guaranteed to infiltrate some system based on the tie breaking rule, so Defender i's expected payoff is $-c_D\frac{1}{m}v_D^* - c_D d_i$, and choosing $d_i = 0$ is the best response. The Attacker's expected payoff is

$$\left[1 - (1 - p_D)\prod_{i \in M}\left(1 - \frac{a_i}{u_i}\right)\right]\mathcal{B}^* - c_A\sum_{i \in M} a_i$$

The derivative with respect to any a_i is

$$(1 - p_D)\prod_{k \in M \setminus \{i\}}\left(1 - \frac{a_k}{u_k}\right)\frac{\mathcal{B}^*}{u_i} - c_A$$

Conditional on $a_k = 0, \forall k \neq i$, the Attacker is indifferent if $(1 - p_D)\frac{\mathcal{B}^*}{u_i} = c_A$. Substituting in u_i yields

$$(1 - p_D)\frac{m\mathcal{B}^*}{\frac{1}{m}v_D^*} = c_A \Leftrightarrow p_D = \frac{mv_A^* - \frac{1}{m}v_D^*}{mv_A^*}$$

Whenever $a_k > 0$ for some $k \neq i$, then the Attacker strictly prefers $a_i = 0$. Therefore, everyone is best responding. □