Deterrence in the Cyber Realm:

Public versus private cyber capability

Nadiya Kostyuk*

September 15, 2019

DRAFT

Abstract

Can cyber deterrence work? Empirical research on cyber deterrence has told us relatively little about the deterrence ability of cyber institutions — defined as publicly observable proactive efforts aimed at signaling a country's level of cyber offensive and defensive capability. Using an incomplete-information model, I demonstrate that cyber institutions sometimes deter adversarial strategic cyber operations (e.g., election interference, attacks against critical infrastructure objects) by prevention and threat of punishment. However, states tend to sub-optimally over-invest resources in these institutions. In particular, weak cyber states tend to over-invest to convince strong cyber adversaries that they are strong, whereas strong cyber states over-invest so that adversaries do not believe that they are weak states pretending to be strong. By doing so, these states reduce their overall cyber capability. Through my interviews with cybersecurity experts, intelligence reports, and examples of attempted election interference campaigns, I establish the empirical plausibility of my results. My focus on the strategic use of cyber institutions as a deterrent represents a departure from existing literature — which has focused only on cyber operations — and has important policy implications.

Word count: 10,253 (total)

*Doctoral Candidate, University of Michigan, Ann Arbor; Pre-doctoral Fellow in Cyber Security and Policy at the Fletcher School of Law and Diplomacy and the School of Engineering, Department of Computer Science, Tufts University; nadiya@umich.edu; Web: http://www-personal.umich.edu/ nadiya/index.html

Cyber deterrence is a highly debated topic in scholarly discussions.[1] These discussions have focused mostly on how major powers can use cyber operations[2] to signal their cyber capabilities to deter their adversaries in cyberspace. However, researchers have concluded that this type of deterrence is ineffective because of the "cyber attribution challenge" – the difficulty of identifying the perpetrators of cyber operations (Borghard and Lonergan, 2017; Brantly, 2016; Gartzke, 2013; Libicki, 2009; Lindsay and Gartzke, 2015; Council et al., 2009; Nye Jr, 2017; Valeriano, Jensen and Maness, 2018).[3] Since signals can be misinterpreted, cyber operations intended to signal the country's cyber capability instead might increase uncertainty about the distribution of power, thereby making escalation or war more likely (Fearon, 1995; Powell, 2002; Reed, 2003; Smith and Stam, 2004).[4]

Perhaps for the reasons listed above, over the last decade, nations have been working on a new approach to deter adversarial cyber attacks. Specifically, they have begun creating *cyber institutions* (CBI). I define *cyber institutions* as *publicly observable proactive efforts aimed at signaling a country's level of cyber-offensive and/or defensive capability* ("cyber capability" hereafter). Unlike cyber operations, which often have no "return address" due to the attribution challenge or are too secretive to be revealed because they aim to collect intelligence, cyber institutions send an immediate, rough estimate of the nation's cyber capability. For example, Sweden created an agency to protect its citizens from misinformation and designated its election systems as critical infrastructure objects that will be under the government's protection to signal an increase in its cyber defenses for the purpose of deterring foreign powers from interfering in its 2018 national elections (Cederberg, 2018, 2). Given that Sweden is a nation with weak cyber capabilities, *can*

[1] Some scholarly works include Borghard and Lonergan 2017; Brantly 2016; Gartzke 2013; Libicki 2009; Lindsay and Gartzke 2015; Council et al. 2009; Nye Jr 2017; Valeriano, Jensen and Maness 2018.

[2] I use *Joint Publication 3 13 Information Operations* (2014, II-9)'s definition of *cyber operations*: "the employment of cyberspace capabilities where the primary purpose is to achieve objectives in or through cyberspace."

[3] Scholars also distinguish deterrence by normative taboos and entanglement (Brantly, 2018; Nye Jr, 2017). Because I focus on in-kind operations — cyber-to-cyber deterrence, these other types of deterrence are not the main focus of this paper.

[4] Ball (1993) argues that such uncertainty can serve as a deterrent.

this strategy work for weak cyber nations? Is deterrence possible in cyberspace?

To address these questions, this paper develops the first theory in the international-relations literature to explain how states use cyber institutions as a way to deter adversaries that might be contemplating strategic cyber attacks that threaten the country's prosperity and security (Fernandino, 2018). I focus on strategic cyber attacks because (1) it is impossible to replicate nuclear weapons' absolute deterrence in cyberspace, and (2) states tend to prioritize deterrence of these attacks over low-level cyber operations because they have uncertain and dramatic consequences and cannot be completely defended against.

The central claim of my theory is that the state can deter her adversary contemplating a cyber attack by signaling her cyber capability via cyber institutions. I argue that this approach is more efficient than using non-cyber foreign policy tools, such as a diplomatic, economic, or military response, because they are too costly for the state to use and/or are ineffective. These non-cyber tools are ineffective because the adversary contemplating a cyber attack has a good estimate of how likely the state is to use her non-cyber tools given her past history. The adversary is willing to risk costs imposed by these tools because the capacity of these non-cyber tools is unlikely to rapidly change (Fearon, 2002, 6). This approach of signaling cyber capability via cyber institutions is more effective than signaling via cyber operations because it (1) preserves the state's cyber operations whose value diminishes after their use and (2) provides the adversary with an immediate, often rough proxy for the state's cyber capability. Because this estimate is not always accurate, the adversary might overestimate the state's capability, and as a result, given the state's already high resolve, be deterred. I assume that the state has a high resolve to use cyber capability against her adversary, considering the high-stakes situation – the adversary is contemplating an attack (Press, 2005).

To explain when such an approach can work, I develop an incomplete-information model in which the adversary's decision to attack is endogenous to the state's type. I argue that there

is an optimal allocation of limited resources between cyber institutions and covert cyber activity (e.g., ongoing cyber operations) in terms of actual developed capability, but my model equilibria show that states tend to make sub-optimal choices. Cyber institutions influence an adversary's decision to attack only in limited cases, but states tend to sub-optimally over-invest their resources in publicly observable cyber institutions instead of distributing these resources between cyber institutions and covert cyber activity to maximize their overall cyber capability. Weaker states over-invest in cyber institutions to signal higher cyber capability than they possess and strong cyber states over-invest so that their adversaries do not believe that they are weak states pretending to be strong.

Scarce data on cyber incidents and cyber institutions,[5] compounded by the secrecy surrounding national security issues, make it difficult to undertake a rigorous empirical test of the findings of my model. Instead, I use a series of interviews with cybersecurity experts, intelligence reports, and examples of attempted election interference, to establish the empirical plausibility of my theory and demonstrate that cyber institutions can deter strategic cyber attacks in limited cases (Interview 2018: #48, 49). For example, through the creation of cyber institutions, Sweden was able to deter Russia from proceeding with its influence campaign during Sweden's 2018 elections. However, cyber institutions had no effect on Russia's decisions on whether to attack the 2017 German elections and 2016 U.S. elections. In the German case, non-cyber factors shifted Russia's cost-benefit calculus in favor of not attacking even before cyber institutions were put in place. In the U.S. case, Moscow was willing to pay any cost and to risk any potential U.S. (cyber or non-) retaliation for its influence campaign; no cyber institutions could have stopped this campaign.

This paper makes a number of theoretical contributions to the existing international-relations

[5]Currently, there are only two published datasets on cyber incidents: Valeriano, Jensen and Maness 2018's Dyadic Cyber Incident Dataset and Kostyuk and Zhukov 2019's data on conflict in Ukraine. I am currently developing the datasets on cyber institutions.

literature. First, it helps us better understand the nature of deterrence in the information age. My unique strategic logic of cyber institutions as a proxy for a country's cyber capability represents a departure from existing literature. While most scholarly works on cyber-to-cyber deterrence focus on deterrence by punishment using cyber operations (Baliga et al., 2018; Gartzke and Lindsay, 2015; Lindsay, 2013; Rid, 2013; Valeriano and Maness, 2014; Valeriano, Jensen and Maness, 2018), this project presents a new theory that explains how cyber institutions can deter by both prevention and threat of punishment, by signaling an increase in both offensive and defensive cyber capabilities.

Second, much of the literature on cyber coercion either focuses on intelligence and policy dilemmas confronting major powers or seeks to adapt existing theories of coercive diplomacy and deterrence to explain the strategic behavior of major powers in cyberspace (Borghard and Lonergan, 2017; Brantly, 2016; Gartzke, 2013; Libicki, 2009; Lindsay and Gartzke, 2015; Nye Jr, 2017; Valeriano, Jensen and Maness, 2018). I, instead, seek to explain the strategic behavior of weaker cyber states or middle powers when these states suspect a cyber attack. In these situations, middle powers often must rely on their own cyber capabilities because their allies are unlikely to help. Cyber defenses are unique to each country and not easily transferable; close allies are often reluctant to share their cyber offenses (as compared with their willingness to offer military assistance during territorial invasions).

Third, this paper uses new sources to provide the first-ever theoretically informed explanation of how nations can use its cyber capability to deter state-sponsored interference campaigns. Most works on this relatively new phenomenon are limited to descriptive policy reports or formal accusations that trace the evidence to potential perpetrators (Brattberg and Maurer, 2018*b*; Cederberg, 2018; Galante and EE, 2018; Mueller, 2019; USDepartmentOfJustice, 2018*c*,*a*).

Before explaining the logic of my deterrence-by-cyber-institutions model (Section 2), I define

what constitutes election interference campaigns (Section 1). Then, I evaluate the expectations derived from the model using the case studies of the 2017 German (Section 3.1), 2018 Swedish (Section 3.2), and 2016 U.S. elections (Section 3.2). Online Appendix 1 contains formal statements of all propositions and their proofs.

1 Defining Election Interference

Before moving to my model, I define what types of actions can constitute *election interference* and the role a state can play in this process.[6] Table 1 lists the six types of actions that the state can take by using cyber and/or information operations (e.g., propaganda, fake news) as well as the three different levels of state-involvement to influence an election outcome (Galante and EE, 2018).

Table 1: TYPES OF ELECTION INTERFERENCE

		Level of State Involvement		
		State-directed	*State-encouraged*	*State-aligned*
Cyber Operations	*Infrastructure Exploitation*	✓		✓
	Vote Manipulation	✓		
Information Operations	*False Front Engagement*		✓	
	Sentiment Amplification		✓	✓
	Fabricated Content	✓	✓	
Cyber and Information Operations	*Strategic Publication*			✓
		2014 U.S. elections	*2016 U.K. referendum*	*2017 French elections*
		Election examples		

First, the state can distort data or system functionality by infrastructure exploitation. For example, the Russian state used cyber operations to compromise voter registration databases and campaign finance databases in thirty-nine states during the 2016 U.S. elections (Riley and Robertson, 2017). Second, the state can use cyber operations to manipulate vote by changing vote tallies, input, or transmission. For example, the Russian state destroyed key files of

[6] I use "election interference" and "influence campaign" interchangeably in this paper.

Ukraine's Central Election Commission's programs that monitor the tallying of votes a few days before the 2014 Ukrainian elections. Third, the state can use cyber operations to illicitly obtain sensitive data, such as internal communications, and then strategically release them to tarnish the reputation of electoral candidates. For instance, after having obtained sensitive information through the 2015 U.S. Democratic National Committee (DNC) hack, the Russian state published this information on DCLeaks.com and WikiLeaks to damage Hillary Clinton's candidacy. The state can also use information operations to execute either independently or together the following three actions — false-front engagement, sentiment amplification, and fabricated content. The Internet Research Agency (IRA) — a Russian company whose owner Yvgeniy Prigozhin has ties to Putin — often uses these three techniques at the same time when it creates social media accounts impersonating Americans who often make incorrect claims to exploit divisive political sentiments (USDepartmentOfJustice, 2018*c*).

In addition to these six types of actions used to influence elections, I distinguish three possible levels of state involvement in influence campaigns. First, the state can direct a campaign. Cyber operations in the 2014 Ukrainian and 2016 U.S. elections attributed to Advanced Persistent Threat (APT) 28 — a part of the Russian military's main intelligence directorate, the GRU (Alperovitch, 2016) — are examples of state-directed interference. Second, the state can encourage interference by ensuring that the third party with knowledge of the "state's objectives can partake with reasonable assurance that these efforts will be viewed favorably" (Galante and EE, 2018, 6). For instance, prior to the 2016 Brexit referendum, IRA operated an extensive social media pro-Brexit campaign but no evidence confirmed that the Kremlin directed this campaign. Last, the state has no involvement in an election interference campaign, although the interference is aligned with state objectives. For example, the interference into the 2017 French elections seemed to align with the objectives of the Russian state. But the French National Agency for the Security of Information Systems (ANSSI) confirmed that the Kremlin was not

behind the interference and presented the simplicity of the attacks as evidence pointing to an actor with lower cyber capabilities than the Russian state (*France says no trace of Russian hacking Macron*, 2017). This paper focuses on examples of *state-directed* attempts to interfere into foreign elections by using cyber and information operations to execute any of the mentioned-above six types of actions.

2 A Theory of Cyber Deterrence

The theory presented here examines a specific context in which a challenger contemplates a strategic cyber attack against a defender.[7] The challenger decides to carry out a cyber attack and accepts the risks and costs of non-cyber foreign policy tools that the defender might use because he is able to predict the inefficiency of these tools given her past history (Fearon, 2002, 6). In such cases given the ineffectiveness of her non-cyber policy tools, she is left with cyber tools to deter an immediate threat. Even though the challenger may have a rough estimate of the defender's cyber capability, the dynamic nature of cyber tools in comparison to non-cyber policy tools creates a higher degree of uncertainty about the defender's cyber capability. The defender can choose to signal her cyber capability via cyber operations but they do not constitute a good signaling mechanism because their value diminishes after the first use. Additionally, these signals might not be received, given the difficulty of attributing cyber operations. I argue that the defender's choice to signal via public cyber institutions for example, is more effective because it allows the defender to preserve her cyber operations and provide the challenger with an immediate, albeit uncertain, estimate of the defenders cyber capability. This uncertain estimate of the defendant's cyber capability, coupled with the defender's already high resolve to preserve her in this high-stake scenario, might lead the challenger to overestimate the defender's

[7] I employ the language from the traditional deterrence literature and refer to an adversary as a *challenger* and a defending state as a *defender*. Because the model focuses only on a strong challenger, I have omitted "strong" and refer only to a "challenger" for the remainder of the paper.

capability and, as a result, be deterred (Press, 2005).

I will now develop the model more explicitly. I start by explaining how the defender can signal her cyber capability (Section 2.1) and then examine the defender's (Section 2.2) and the challenger's (Section 2.3) optimal actions, which are depicted in an extensive form game (Section 2.4) and presented in model equilibria (Section 2.5).

2.1 Signaling Cyber Capability

For simplicity, I distinguish two ways a defender can signal her cyber capability — invest in public cyber institutions (CBI) and invest in covert cyber activities (CCA).

I define public *cyber institutions* as *publicly observable proactive efforts aimed at signaling a defender's level of cyber-offensive and/or defensive capability*. Out of the plethora of ways the defender can publicly signal her capability, I explore the following three: (1) the creation of a new agency, program, or initiative, and/or the adoption of a new doctrine, strategy, or policy to address some aspect of cybersecurity (e.g., U.S. Cyber Command, the 2011 Department of Defense (DoD) Strategy for Operating in Cyberspace); (2) the addition of new cyber roles to existing agencies or cyber provisions to existing policies (e.g., making the Ministry of Education responsible for the development of a nationwide curriculum to improve computer science skills); and (3) the attribution of cyber operations (e.g., U.S. Department of Justice indictments (USDepartmentOfJustice 2014-USDepartmentOfJustice 2018*b*)).

Public cyber institutions can deter by prevention and/or by threat of punishment. For example, building a firewall, updating computer software, or establishing programs to educate various groups about cyber and information threats aims to deter by prevention. France aimed to deter through the threat of punishment when it published its first doctrine for offensive cyber operations (*Public Elements of the Doctrine on Military Cyber Offensive*, 2019), stating that the country was "not afraid" of using cyber "weapons" in response to cyber threats (Laudrain,

2019). The *U.S. Department of Defense Cyber Strategy* (2015, 3) uses deterrence by prevention when it explains the Department of Defense's (DoD's) role in defending its information networks and uses the threat of punishment when it deploys the Cyber Mission Force to defend the United States against cyber attacks of "significant con-sequence." I investigate situations when both deterrence mechanisms are at play because one institution can signal both cyber-offensive and defensive capability and a few institutions can signal either defensive or offensive capability. I view a cyber institution as a continuous variable and measure it by the amount of resources the defender devotes to its creation.

I define *covert cyber activities* (CCA) as *non-public development of state cyber capabilities and ongoing cyber operations.* Besides secret cyber operations, CCA include any department, program, or initiative that secretly creates cyber capability and does not send any signal about this existing capability. For instance, the U.S. National Security Agency (NSA) has been developing cyber capabilities for some before it became known that the agency had such capabilities. Secrecy that distinguishes CCA from CBI creates a fine line between the two. For example, the *Stuxnet* worm used against an Iranian nuclear enrichment facility was an example of a CCA until its discovery and attribution to the U.S. and Israeli governments in 2010. Once attributed, the *Stuxnet* worm became a CBI that sent a clear signal about the level of offensive cyber capability that existed within the U.S. and Israeli governments. Using primary evidence from my interviews, I assume that both CCA and CBI are jointly optimal for the development of a state's cyber capability, and I opt to focus on the effect of cyber institutions on deterrence.[8]

The relationship between a defender's cyber capability and her chances at deterrence is not straightforward — an increase in one does not necessarily mean an increase in the other. For deterrence to be successful the defender does not have to have strong cyber capabilities — she needs only to appear to have these capabilities. Because perceived and not actual cyber capa-

[8]Section 2 gives an overview of my interviews.

bilities are all that matter (Jervis, 1976), a defender with weak cyber capabilities might strategically invest more in publicly observable CBI if such an investment maximizes her chances at deterrence, even if it decreases her overall cyber capability. Notably, while CBI can help bolster the effectiveness of cyber capabilities, over-investment in CBI takes valuable resources away from developing cyber capabilities. For instance, a polished strategy document or newly-constructed headquarters is a poor substitute for a cyber covert activity like getting into a potential challenger's electricity grid and preparing for a future attack that could be used for retaliation or demonstrating the defender's cyber capability.

2.2 Defenders

The model assumes that defender D already has some existing cyber capacity[9] when she decides which level of CBI to implement. This capacity is a prerequisite for the level of CBI the defender can establish and determines the defender's type θ. The defender can be either *strong* or *weak* ($\theta \in \{S, W\}$).[10] Nature selects that a defender is strong with probability q and that she is weak with $1 - q$.[11] If $I_\theta(r)$ is the CBI implemented by D_θ that used resources $r \in [0, 1]$, then $I_S(r) > I_W(r)$ and $I_S^{'}(r) > I_W^{'}(r) \geq 0$ for all r.[12] This means that if both types have the same amount of resources and invest the same amount into CBI, the strong type will end up with more sophisticated CBI that signal greater cyber capability, and, as a result, has greater deterrence chances (Figure 1).

In addition to investing in cyber-institutions, the defender spends her remaining resources on covert cyber activity (CCA), denoted by $N_\theta(r)$. Then $N_\theta(r) = (1 - r)n_\theta$, where n_θ is the

[9] Capacity refers to the defender's total cyber capability at the beginning of the game.

[10] I do not distinguish the defender's type using its capability/vulnerability ratio because with an increase in her Internet usage and, as a result, her vulnerability, there is an increase in a defender's ability to defend her networks and address some of these vulnerabilities.

[11] For simplicity, the model assumes that q is 50% — a defender is strong half of the time.

[12] $I_\theta(r)$ is a twice-differentiable and concave function with $I_\theta^{'}(r) > 0$ and $I_\theta^{''}(r) < 0$. The model assumes functions $I_\theta(r)$ to be strictly increasing in r and for the marginal returns to be strictly decreasing.

rate at which CCA change as resources change, determined by the defender's type θ. Similarly, because a strong type has a higher level of existing cyber capacity within her government than a weak type, it makes it easier for a strong type to implement CCA, and thus $n_S > n_W$ (Figure 1).

Together, the sum of resources spent on CBI and CCA equals to resources spent on creating a new level of cyber capability: $c_\theta(r) = I_\theta(r) + N_\theta(r)$,[13] where $c_\theta(r)$ is cyber capability that a defender of type θ creates when she implements a publicly observable CBI. This cyber capability captures the defender's overall ability to deter the challenger. Let

$$\hat{c}_\theta = max_{r\in[0,1]} c_\theta(r) = max_{r\in[0,1]} I_\theta(r) + N_\theta(r) \tag{1}$$

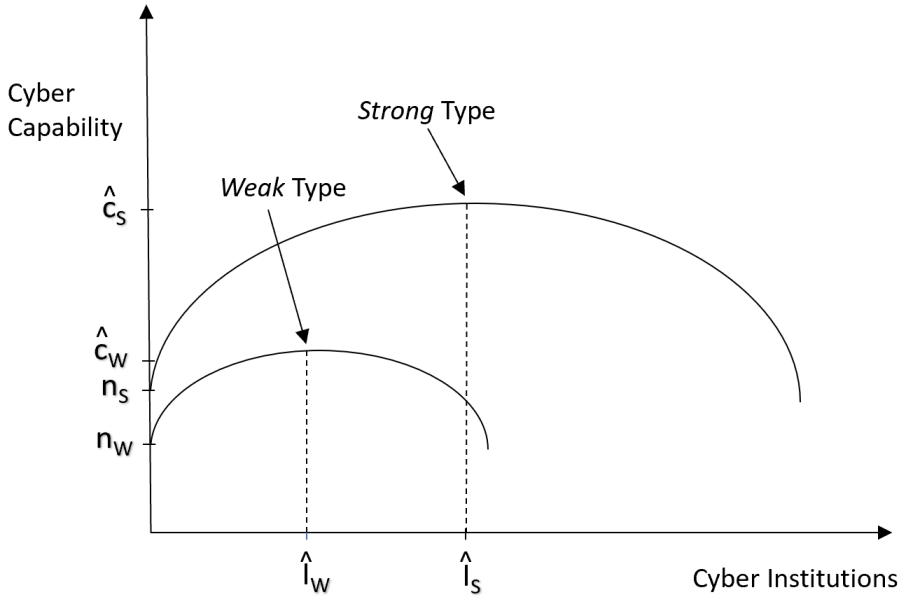
be the largest cyber capability a defender of type θ can attain. Let $\hat{r}_\theta$ be the value of r that maximizes a defender's cyber capability in Equation 1 and $\hat{I}_\theta = I_\theta(\hat{r}_\theta)$ be the corresponding CBI that achieves this.

Equation 1 displays two important points. First, a defender's cyber capability does not necessarily increase simply by investing more in her CBI because such a strategy takes valuable resources away from CCA. Second, weak and strong types will budget their resources differently, if both types maximize their cyber capability. Specifically, strong types will invest more in cyber institutions ($\hat{r}_S > \hat{r}_W$), will have a higher level of cyber capability ($\hat{c}_S > \hat{c}_W$), and, as a result, will be more effective in deterring their adversaries, than will weak types. This is because, to maximize her cyber capability, the defender will budget her resources so that her marginal returns, in terms of cyber capability from CCA, equals that from CBI ($I'_\theta(\hat{r}_\theta) = n_\theta$).[14] After weak types have reached this point, strong types can still achieve higher cyber capability by investing more resources in CBI, as Figure 1 shows. As a result, the deterrence chances are higher at the point where strong types reach their optimal level of CBI.

[13] This model assumes that c is decreasing in r beyond some point and $I'_\theta(1) < n < I'_\theta(0)$ for $\theta \in S, W$.

[14] As a concave function, $\hat{c}_\theta$ achieves its maximum when $c'(\hat{r}_\theta) = 0$. Since $c'(\hat{r}_\theta) = I'_\theta(\hat{r}_\theta) + N'_\theta(\hat{r}_\theta) = I'_\theta(\hat{r}_\theta) - n_\theta$, $I'_\theta(\hat{r}_\theta) = n_\theta$.

Figure 1: THE RELATIONSHIP BETWEEN RESOURCES INVESTED IN CYBER INSTITUTIONS AND A DEFENDER'S CYBER CAPABILITY



If the defender successfully deters the challenger and the challenger decides not to attack, the defender receives value from deterring her challenger, $V_D \in [0,1]$. If the defender does not deter the challenger using her CBI, she pays the cost of being attacked, $C_D \in [0,1]$. The challenger's choice of action is endogenous to the defender's CBI, which signals to the challenger the defender's type, her cyber capability, and how damaging retaliation will be if the challenger attacks this type and/or how good the defender's cyber defenses are, allowing the challenger to estimate how costly it will be for him to break these defenses. The challenger wants to avoid attacking the strong defender because her retaliation will do more damage to the challenger than will retaliation by a weak type and/or it will be much costlier to break the defenses of the strong type than to break those of the weak type.

If the challenger attacks, the defender's cyber defenses may or may not be sufficient to stop this attack from getting through. The challenger's probability of successful attack (γ) depends on the defender's cyber capability. I write $\gamma \equiv \gamma(c)$,[15] $\gamma_S = \gamma(\hat{c}_S)$ and $\gamma_W = \gamma(\hat{c}_W)$ are the

[15]The model assumes γ is decreasing in c and $\gamma \in [0,1]$.

lowest probabilities that the challenger successfully attacks strong and weak defenders.

If the challenger's attack is successful, the defender decides whether to retaliate. If she retaliates, she pays the cost of attempted retaliation, $C_R \in [0,1]$, regardless of whether retaliation is successful. If retaliation is successful, the defender additionally receives value from successful retaliation, $V_R \in [0,1]$.[16] The probability that this retaliation is successful (P) depends on the defender's cyber capability. I write $P \equiv p(c)$,[17] and $P_S = p(\hat{c}_S)$ and $P_W = p(\hat{c}_W)$ are the greatest probabilities that the defender's strong and weak type successfully retaliate against the challenger.

2.3 Challengers

The model assumes the challenger cannot observe the defender's type directly, but he has a prior probability of her type, derived from, for example, past cyber operations attributed to the defender and the defender's technological and scientific abilities.[18] I assume that the inability to confirm some of these claims and indicators, among other variables, make the challenger uncertain about the defender's type. Having observed public CBI, the challenger (possibly) updates his beliefs about whether he faces the defender's strong type (μ), which factors into his decision whether to attack her. The challenger attacks the defender whenever his net gains from attacking outweigh his net gains from not attacking:

$$P_B V_C - P_B \sigma_{R_\theta} P_A C_P - C_C > R. \tag{2}$$

In Equation 2, R is the challenger's reservation utility, which represents the challenger's net gains from not attacking the defender ($R \geq 0$). $V_C \in [0,1]$ is the value that the challenger

[16] Because a defender's main goal is to deter a challenger, the model assumes that $V_D > V_R$.

[17] The model assumes that P is increasing in c and $P \in [0,1]$.

[18] The model assumes that both players have a common prior and that q is a true probability that a defender is a strong type.

receives from attacking the defender. $C_P \in [0,1]$ is the challenger's expected cost from the defender's retaliation. $C_C \in [0,1]$ is the challenger's expected cost from attacking the defender. σ_{R_θ} is the probability distribution that D_θ retaliates against C as a function of whether C attacks D_θ, having observed her CBI. P_A is the challenger's expectation that the defender will retaliate successfully having observed the defender's CBI. I estimate this expectation as $P_{A(I)} = \mu(I)p(c_S(I)) + (1-\mu(I))p(c_W(I))$, where $P_S = p(\hat{c}_S)$ and $P_W = p(\hat{c}_W)$ are the greatest probabilities that the defender's strong and weak type successfully retaliate against the challenger. In Equation 2, the defender's cyber capability serves as a proxy for the challenger's expectation of the defender's retaliation — the larger the challenger's expectation of the defender's cyber capability, the more likely he is to believe that the defender will retaliate against him after being attacked, and the more damaging he believes this retaliation will be. As a result, the larger the expectation of the defender's cyber capability, the more deterred the challenger will be from attacking her.

P_B is the challenger's expectation that the defender's cyber shields will successfully hold against his attack, having observed the defender's CBI. I estimate this expectation as $P_{B(I)} = \mu(I)\gamma(c_S(I)) + (1-\mu(I))\gamma(c_W(I))$, where $\gamma_S = \gamma(\hat{c}_S)$ and $\gamma_W = \gamma(\hat{c}_W)$ are the lowest probabilities that the challenger successfully attacks strong and weak defenders or the highest probabilities that the defender's cyber shields successfully hold against the challenger's attack. Similarly, in Equation 2, the defender's cyber capability serves as a proxy for the challenger's expectation of how unbreakable the defender's defenses are — the larger the challenger's expectation of the defender's cyber capability, the more likely the challenger is to believe that the defender's cyber defenses will hold against his attack and, as a result, the more deterred the challenger will be from attacking the defender. Lastly, Equation 2 demonstrates the difference between my model of cyber deterrence and the model of deterrence by non-cyber means — in the former, even if the challenger's attack is not successful, the defender's retaliation is

assumed.

Because the challenger observes the same cyber institutions to estimate the probability of the defender's retaliation and the probability that her defenses withstand his attack, I assume that $P_A = P_B$.[19] Since the defender's programs devoted to CCA are not completely observable, the challenger makes this decision while still facing some degree of uncertainty about the defender's true type, $\mu(I) \in (0, 1)$.

2.4 Game & Solution Concept

Game. Figure 2 displays my two-player game with incomplete information concerning the defender's type. It proceeds as follows:

0. Nature N has chosen to start the game at the open dot and chooses the defender's type D_θ, where $\theta \in \{S, W\}$. With probability q, N selects that D_θ is strong and with probability $1 - q$, N selects that D is weak.

1. D_θ learns her type and is faced with choosing the level of CBI to implement to deter a challenger C. The choices are denoted by the sector of the circle starting at D_θ. The arc represents resources, distributed between 0 and 1 for simplicity, that D can choose to invest in CBI. Even though the arc is a continuum of choices, D_θ can choose either a weak or strong cyber institution to signal her capability. Both types prefer to deter C.

2. After D_θ implements CBI, C observes it, (possibly) updates his beliefs about D_θ's type and her level of cyber capability, and decides whether he wants to attack D_θ, considering the possibilities that his attack might not succeed and that D_θ might retaliate. These scenarios for how C can react to D_θ's CBI are represented by the two filled-in dots next to Challenger. C chooses whether to attack or not D_θ. C's choices are represented by the lines

[19] In the future iteration of this paper, I plan to derive model equilibria where $P_A \neq P_B$.

emanating from the two filled-in dots. The important point here is that when C chooses whether to attack, he does so knowing what CBI D_θ implements, but not knowing D_θ's type with any certainty. This is depicted in the picture by dashed red lines around the solid dots, representing C's information set. In this information set, C can see one of the following combinations of CBI: (1) D_S and D_W implement the same level of CBI; (2) D_S and D_W implement different levels of CBI that are respectively typical for their types ($D_S \rightarrow \hat{I}_S$, and $D_W \rightarrow \hat{I}_W$). If C decides not to attack, D_θ successfully deters C and receives V_D. The game ends.

3. If C decides to attack D_θ, his attack can either succeed or not. If he succeeds, he receives the value from attacking V_C, while paying the cost C_C of attacking. If he does not, he pays C_C and receives no value. The probability of a successful attack is determined by D_θ's overall cyber capability and how she distributes her resources at Stage 1. With probability γ_θ, N selects that C's attack is successful, and with probability $1 - \gamma_\theta$, N selects that C's attack is not successful.

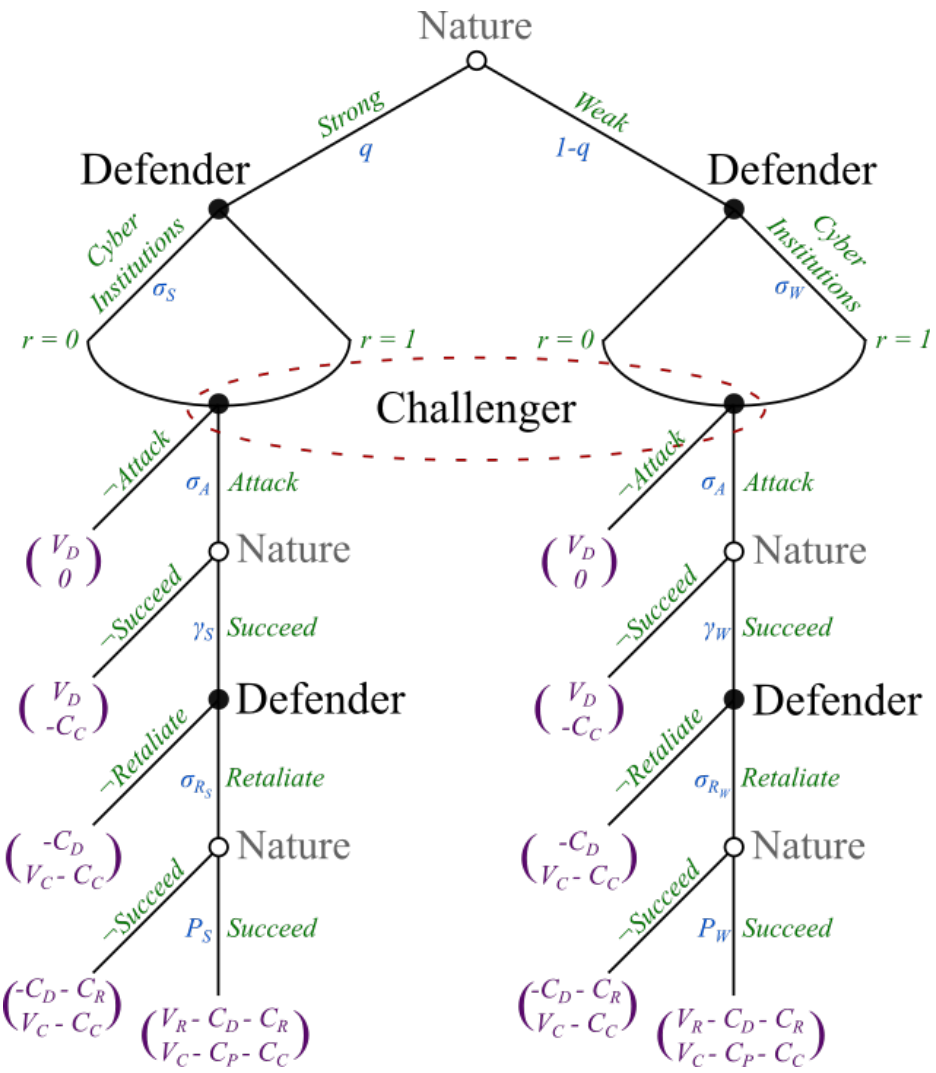
4. If C's attack succeeds, D_θ must decide whether to retaliate against C.[20] If she does not, D_θ pays the cost of being attacked C_D, and the game ends.

5. If D_θ decides to retaliate against C, D_θ pays the cost of retaliating C_R. The probability of a successful retaliation is determined by D_θ's overall cyber capability and how she distributes her resources at Stage 1. With probability P_θ, N selects that D_θ's retaliation is successful and with probability $1 - P_\theta$, N selects that D_θ's retaliation is not successful. If D_θ's retaliation is successful, D_θ receives the value V_R, and C pays the cost of suffering retaliation C_P. Regardless of whether or not the retaliation is successful, the game ends.

[20]The defender knows who the challenger is because the model operates in a closed system with two players and cyber attribution is no longer a challenge. I do not model the cyber attribution challenge for the following two reasons: (1) a state's decision to attribute cyber operations is no longer a technical challenge but is instead a political decision (Interview, 2018: #15, #30; Rid and Buchanan, 2015; Soldatov and Borogan, 2017); and (2)Baliga et al. (2018) models the feasibility of deterrence when the cyber attribution challenge is present.

6. Payoffs are received.

Figure 2: EXTENSIVE FORM GAME TREE OF DETERRENCE BY CYBER INSTITUTIONS



Solution Concept. An equilibrium to the model is defined by the set of strategies and beliefs:

$$(\sigma_S, \sigma_W, \sigma_{R_S}, \sigma_{R_W}, \sigma_A, \mu),$$

which are probability distributions that (1) strong and (2) weak types of D implement each possible level of CBI; the probability distribution that (3) strong and (4) weak types of D retaliate

against C as a function of whether C attacks this type, having observed her CBI; (5) the probability that C attacks D as a function of D's CBI that C observes; and (6) C's posterior probability that D is of a strong type, given CBI he observes.

The solution concept is Perfect Bayesian equilibrium $(\sigma_S, \sigma_W, \sigma_{R_S}, \sigma_{R_W}, \sigma_A, \mu)$, which has four components. First, σ_{R_θ} is the probability distribution over $[0, I_\theta(1)]$ for each type of D, θ. $\sigma_{R_\theta} > 0$ only if this probability maximizes D_θ's expected payoff, given C's decision to attack her having observed D's CBI ($\sigma_A : [0, I_S(1)] \rightarrow [0,1]$). D retaliates against C when net gains from retaliation are at least as good as net gains from non-retaliation. D's expected payoff is shown in Equation 3.

$$\sigma_{R_\theta}(I) > 0 \Leftrightarrow I \in \arg\max_I \Big[(1-\sigma_{R_\theta})(-C_D) + \sigma_{R_\theta}\Big[p(c_\theta)[V_R - C_D - C_R] + (1-p(c_\theta))(-C_R)\Big]\Big] \quad (3)$$

Second, there is a probability distribution σ_θ over $[0, I_\theta(1)]$ for each type of D, θ. $\sigma_\theta > 0$ only if this probability maximizes D_θ's expected payoff, given C's decision to attack her having observed D's CBI ($\sigma_A : [0, I_S(1)] \rightarrow [0,1]$). D's expected payoff is shown in Equation 4.

$$\sigma_\theta(I) > 0 \Leftrightarrow I \in \arg\max_I \Big[\sigma_A\Big[\gamma_\theta\Big[\sigma_{R_\theta}[p(c_\theta)[V_R - C_D - C_R] + (1-p(c_\theta))(-C_R)] + (1-\sigma_{R_\theta})(-C_D)\Big] + (1-\gamma_\theta)V_D\Big] + (1-\sigma_A)V_D\Big] \quad (4)$$

Third, C attacks D with positive probability ($\sigma_A : [0, I_S(1)] \rightarrow [0,1]$) when net gains from attacking are at least as good as net gains from not attacking, given his expectations of D's type:

$$\sigma_A(I) \in \arg\max_{\sigma_A \in [0,1]} \Big[(1-\sigma_A)R + \sigma_A\Big[P_B\Big[\sigma_{R_\theta}[P_A(V_C - C_P - C_C) + (1-P_A)(V_C - C_C)] + (1-\sigma_{R_\theta})(V_C - C_C)\Big] + (1-P_B)(-C_C)\Big]\Big] \quad (5)$$

In Condition 5, $P_{A(I)} = \mu(I)p(c_S(I)) + (1-\mu(I))p(c_W(I))$ and $P_{B(I)} = \mu(I)\gamma(c_S(I)) + (1-$

$\mu(I))\gamma(c_W(I))$. Since Condition 5 is a linear function of σ_A, $\sigma_A^* \in \{0,1\}\ \forall P_{A(I)} \neq \bar{P}$.

Fourth, C updates his posterior beliefs (μ: $[0, I_S(1)] \rightarrow [0,1]$) about D's type using Bayes' Rule

$$\mu(I) = \frac{q\sigma_S}{q\sigma_S + (1-q)\sigma_W}, \tag{6}$$

$\forall I$ such that $\sigma_S(I) + \sigma_W(I) > 0$.

2.5 Equilibrium Results

In this section, Lemmas 1- 3 describes the pure strategy Perfect Nash Equilibrium (SPNE) results that generate testable hypotheses presented in Propositions 1- 4.[21]

Defender's choice to retaliate. I start with the bottom of the game tree and examine the defender's choice to retaliate.

Lemma 1. *A defender retaliates if her value from this retaliation, given the probability that this retaliation is successful, is greater than her cost ($P_\theta V_R > C_R$).*

Lemma 1 describes three results. First, and surprisingly, the defender does not consider her cost of being attacked when contemplating retaliation. Second, D_θ's probability of successful retaliation P_θ increases with D_θ's retaliation cost C_R. C_R can increase because P_θ increases and V_R decreases. This scenario where D's retaliation value is low but D invests significant resources to retaliate against C, leading to D's high probability of successful retaliation, is unlikely because D's value has to be greater than her cost for her to retaliate. C_R can also increase because P_θ decreases and V_R increases. In this scenario, D's retaliation value is high but it is costly to retaliate, leading to a low probability of successful retaliation. C_R can also increase because both P_θ and V_R increase. In this scenario, D's retaliation value is high and she invests

[21] The Online Appendix provides all formal statements and proofs (Section 1).

significant resources to retaliate against C, leading to D's high probability of successful retaliation. Any of the last two scenarios can explain D's cost-benefit calculation to retaliate in this model.

Third, the defender retaliates if her gains are greater than her costs. I assume that different types of defenders gain different values and pay different costs from retaliation. The defender's weak type does not have significant cyber capabilities, making her gains from retaliation significantly lower than her cost ($V_R < C_R$). As a result, the defender's weak type does not retaliate when attacked. On the contrary, the defender's strong type gains significant value from retaliation. Not only does she punish the challenger, she also teaches potential perpetrator not to cyber-attack her. As a result, the strong type always retaliates when attacked. Lemma 2 summarizes this logic.

Lemma 2. *Only the defender's strong type retaliates when attacked because her value from this retaliation, given the probability that this retaliation is successful, is greater than her cost ($P_\theta V_R > C_R$).*

Challenger's choice to attack. I move up the game tree and examine the challenger's choice to cyber-attack. Lemma 3 defines the critical value when the challenger is indifferent between attacking and not attacking.

Lemma 3. *A critical value of P_A for which the challenger is just indifferent between attacking or not is:*

(a) $\bar{P} = \frac{C_C+R}{V_C}$, *if* $\sigma_{R_\theta} C_P = 0$, *and*

(b) $\bar{P} = \frac{V_C+\sqrt{V_C^2-4\sigma_{R_\theta} C_P(C_C+R)}}{2\sigma_{R_\theta} C_P}$, *if* $\sigma_{R_\theta} C_P \neq 0$.

Lemma 3 (a) demonstrates that if the defender does not retaliate and/or C's cost of retaliation is zero, the challenger's willingness to attack the defender ($\bar{P}$) rises with the costs from attacking, given the defender's cyber shields (C_C), and the value from his reservation utility (R), and

decreases with the attack value (V_C). Lemma 3 (b)[22] demonstrates that if the defender retaliates and C's cost of retaliation is not zero, the challenger's willingness to attack the defender ($\bar{P}$) rises with the attack value (V_C) and with a decrease in the probability distribution that D_θ retaliates against C as a function of whether C attacks D_θ, having observed her CBI (σ_{R_θ}), the costs from potential retaliation (C_P), from attack itself, given the defender's cyber shields (C_C), and the value from his reservation utility (R). The challenger will attack if his expected probability that the defender's retaliation is successful and that the defender's defenses withstand his attack, having observed the defender's cyber institutions, is lower than his critical value of $\bar{P}$ ($P_A < \bar{P}$).

Let us consider different regions with different relationships between $P_{S/W}$ and $\bar{P}$. When $\bar{P} < P_W < P_S$, a challenger never attacks because the challenger's gains from not attacking are much higher than from attacking. As a result, a defender is left with nothing but maximizing her cyber capability. When $\bar{P} > P_S > P_W$, the challenger always attacks because the challenger's gains from attacking are much higher than from not attacking. As a result, the defender's action cannot significantly influence the challenger's calculus and any attempt to do that will be futile. Having no influence over the challenger's behavior, the defender is left with nothing but maximizing her cyber capability. As a result, the challenger's decision to attack D is *independent* of the defender's type in these two regions. Proposition 1 summarizes these results.

Proposition 1. *In the following two equilibria, the challenger's decision to attack is independent of the defender's cyber institutions.*

(a) When a challenger never attacks because his gains from not attacking are much higher than from attacking ($\bar{P} < P_W < P_S$), a defender is left with nothing but maximizing her cyber capability ($\hat{I}_\theta$). Because the challenger does not attack, the defender does not need to consider to retaliate ($\sigma_{R_\theta} = 0$).

[22] Because $\bar{P}$ is a probability, I assume that $V_C^2 \geq 4\sigma_{R_\theta} C_P(C_C + R)$.

(b) *When the challenger always attacks because his gains from attacking are much higher than from not attacking ($\bar{P} > P_S > P_W$), the defender is left with nothing but maximizing her cyber capability ($\hat{I}_\theta$). Only the defender's strong type retaliates because retaliation is too costly for the weak type ($\sigma_{R_S} = 1$).*

The strategic use of cyber institutions to deter the challenger, if it occurs at all, must occur in the *signaling region*, where $P_W \leq \bar{P} \leq P_S$. Only in this region there is both the *need* for and the *possibility* of deterrence. Here, the defender's cyber institutions can influence the challenger's behavior — the challenger will take different actions depending on the defender's true type. Specifically, the challenger will attack the defender if he knew the defender was weak and will avoid attacking if he knew the defender was strong. Because of this, the weak type has an incentive to imitate the strong type. Proposition 2 summarizes these results.

Proposition 2. *If a challenger prefers to attack a defender's weak type and avoid attacking the defender's strong type ($P_W < \bar{P} < P_S$), the weak type imitates the strong type's cyber institutions and, as a result, deters the challenger. Because the challenger does not attack, the defender does not need to consider retaliation ($\sigma_{R_\theta} = 0$).*

Proposition 2 captures the possibility that the defender invests more resources in cyber institutions and, as a result, deters the challenger. The logic of deterrence in this case, however, is not straightforward – *instead of distributing her resources between public cyber institutions and cyber covert activity to maximize her cyber capability, the defender invests most resources into cyber institutions to hide her cyber weakness.*

A defender's strong type is aware that the defender's weak type is trying to imitate her. The strong type does not want to be confused with the weak type that is more likely to motivate a challenger to attack her. As a result, the strong type alters her behavior to clearly distinguish herself from the weak type to ensure that the challenger is deterred. In particular, she

spends enough resources to reach a level of CBI that the weak type cannot attain. This situation demonstrates that there develops a competition between different types of the defender. Proposition 3 summarizes the results.

Proposition 3. *If a challenger prefers to attack a defender's weak type and avoid attacking the defender's strong type ($P_W < \bar{P} < P_S$), the strong type might decide to spend enough resources to reach a level of CBI that the weak type cannot attain. If she does that, the weak type implements a cyber institution typical for her type ($\hat{I}_W$) and gets attacked by the challenger. The weak type does not retaliate because retaliation is too costly for her ($\sigma_{R_\theta} = 0$).*

The challenger is aware that the defender's weak type might be imitating her strong type. This makes him uncertain about the type of the defender he faces. As a result, the challenger decides to mix between attacking and not, even if he sees CBI typical for the defender's strong type. Proposition 4 summarizes this result.

Proposition 4. *If a challenger mixes between attacking and not attacking a defender's strong type ($P_W < \bar{P} = P_S$), the defender is left with nothing but maximizing her cyber capability ($\hat{I}_\theta$). The defender's strong type retaliates when attacked ($\sigma_{R_S} = 1$).*

Table 2 lists all model equilibria and assumptions. The left region ($\bar{P} < P_W < P_S$) depicts a situation where cyber institutions create a false impression that they are effective in deterring challengers. Even though the challenger does not cyber-attack a defender in this region, the defender's cyber institutions has no influence over this challenger's decision. The right region ($P_W < P_S < \bar{P}$) depicts the opposite situation — cyber institutions create a false impression that they cannot deter challengers. But, in fact, they have no influence over the challenger's choice to attack. The left middle region ($P_W < \bar{P} < P_S$) depicts two scenarios where deterrence by cyber institutions may work. In the top *pooling* scenario, the defender's weak and strong types implement the same CBI, making the challenger believe that he is facing the defender's

strong type. As a result, the challenger does not attack either of the defender's type. In the bottom *strategic separation* scenario, the defender's strong type increases her CBI to some level that the defender's weak type is no longer able to imitate her strong type to clearly distinguish her strong type from her weak type. Here, the challenger attacks the defender's weak type and avoids her strong type. Lastly, the right middle region ($P_W < \bar{P} = P_S$) depicts the situation where the challenger mixes between attacking and not attacking the defender's strong type because he is uncertain about which type of the defender he is facing.

Table 2: MODEL ASSUMPTIONS & EQUILIBRIA

Assumptions	$\bar{P} < P_W < P_S$	$P_W < \bar{P} < P_S$	$P_W < \bar{P} = P_S$	$P_W < P_S < \bar{P}$
Equilibria	$D_S \rightarrow \hat{I}_S; R$ $D_W \rightarrow \hat{I}_W; \neg R$ $C \rightarrow \neg A$	$D_S \rightarrow I_1; \neg R$ $D_W \rightarrow I_1; \neg R$ $C \rightarrow \neg A$ when observes I_1	$D_S \rightarrow \hat{I}_S; R$ $D_W \rightarrow \hat{I}_W; \neg R$ $C \rightarrow$ mixes	$D_S \rightarrow \hat{I}_S$ $D_W \rightarrow \hat{I}_W$ $C \rightarrow A$
		$D_S \rightarrow I_0; \neg R$ $D_W \rightarrow \hat{I}_W; \neg R$ $C \rightarrow \neg A$ when observes I_0		
Results	***CBI has not effect***	***CBI may deter***	***CBI may deter***	***CBI has not effect***

P_S / P_W – D_S / D_W's greatest probabilities of successfully retaliating/defending against C;
$\bar{P}$ — C's willingness to attack; $I_1 > I_W$, $I_0 > I_S$.

3 Evidence

The novelty, secrecy, and sensitivity of the topic of cyber deterrence prevents me from conducting a rigorous empirical test of my findings.[23] Instead, this paper aims to demonstrate the empirical plausibility of my theory and provide support for my model equilibria, using intelligence reports, sixty-five interviews with cybersecurity experts from twenty-five countries (Section 3.1), and examples of the 2016 U.S., 2017 German, and 2018 Swedish elections.

My comparative case study method focuses on Kremlin-directed attempts to influence electoral campaigns in Western democracies. I choose the most similar cases for my comparison;

[23] I plan to conduct such a test in my future research by using either a large-N empirical analysis or historical or archival research to lay out careful evidence for a case study.

they share the same cyber-capable attacker (Russia), have similar targets (Western democracies), use the same methods (cyber and information operations executed by the same set of actors), have the same purpose (election interference), and have similar timeframes (2016-2018). They differ, however, in the level of cyber institutions that the targets implement.

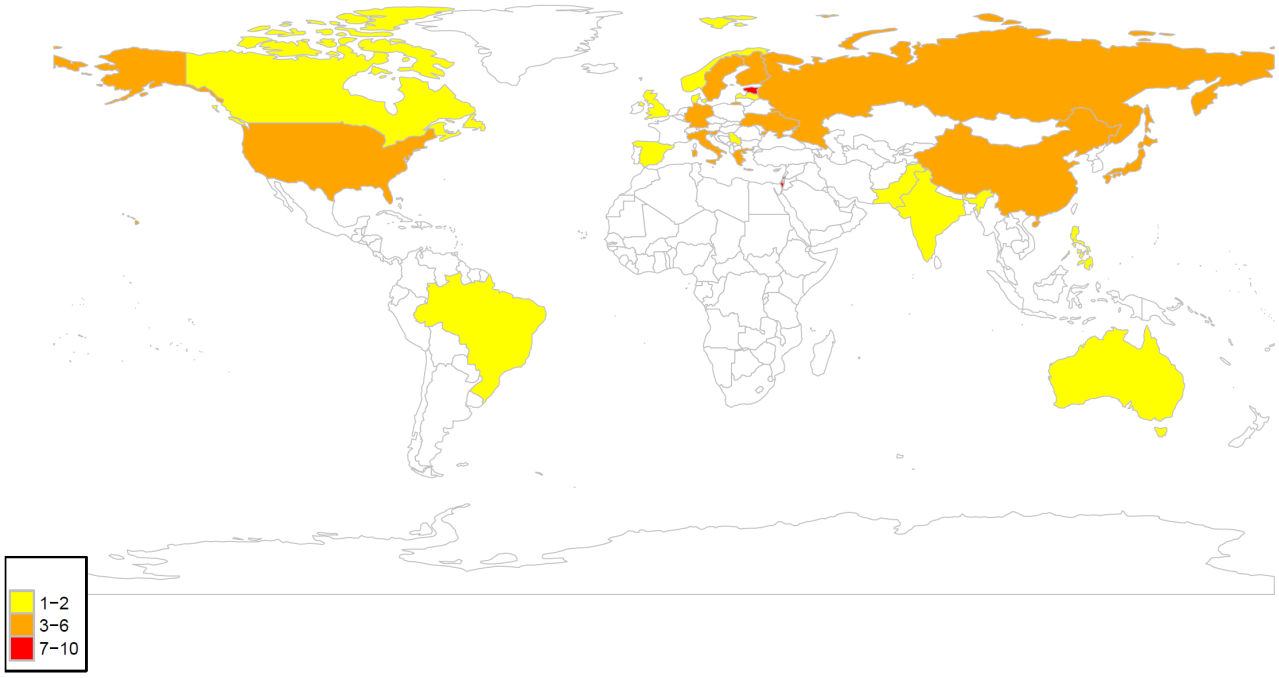
My model predicts that Moscow is deterred when it overestimates its target's cyber capability as a result of observing the target's cyber institutions. This can be seen in the 2017 Swedish election (Section 3.2). In the other two scenarios, cyber institutions had no effect on the Kremlin's decision whether to attack. In the German 2017 elections, a combination of non-cyber factors shifted Russia's cost-benefit calculus in favor of not attacking even before cyber institutions were put in place (Online Appendix, Section 3.1). In the U.S. 2016 elections, Moscow was willing to pay any cost and risk any potential U.S. (cyber and/or non-cyber) retaliation for an influence campaign that could help create a global authoritarian fraternity (Online Appendix, Section 3.2).

As with most deterrence studies, my evidence suffers from two types of critiques (Lupovici, 2018). First is the stated versus actual intent to deter. Even though official documents or public claims state that cyber institutions are created to deter adversaries, it is impossible for one to discern whether actual cyber deterrence was even attempted. While this is a valid concern, the distinction between actual and stated intents is not as important as adversarial perception of this intent. This is connected to the second critique — the deterrent effect. Specifically, one might find it difficult to discern whether cyber institutions have an effect on an adversary's strategic calculations. The fact that adversaries tend to perceive such general assertions of capacity and intention as if they were directed specifically at them partially addresses these two critiques (Segal, 2014).

3.1 Interviews

Overview. I conducted sixty-five interviews with cybersecurity experts from twenty-five countries who were either current or former government employees. These interviews were either conducted in-person or via video calls or emails. Figure 3 displays the number of interviews that I conducted between February and December of 2018. In Israel and Estonia, I was able to conduct nine interviews during this time.

Figure 3: NUMBER OF INTERVIEWS PER COUNTRY (FEBRUARY-DECEMBER 2018)



While in total, I contacted 231 experts in 47 countries, many were reluctant to speak to me even off the record because of the sensitivity of the topic of offensive cyber capabilities. In addition to the common issue that all researchers face of people not responding to emails, there are also three additional barriers. First, many governments have been developing so-called "cyber weapons" for a while but have not publicly announced such efforts. Second, governments do not want to admit that their deterrence efforts have been failing, thereby revealing

the insecurity of their system. Third, even governments that have announced the development of their cyber institutions may not want others to find out that they have not been able to fulfill their proposed plans. Strong governmental control in autocratic regimes might explain hesitation even among non-governmental experts to engage in an interview with me. It took me about forty-four hours to conduct all interviews, with each interview's duration ranging between fifteen minutes and three hours, and a median of an hour and mean of 1.48 hours.

General Themes. My interviewees pointed to two trends in the defender's behavior. First is pooling behavior among weak cyber nations where instead of developing cyber capabilities solely for intelligence collection, for instance, these weaker cyber countries start developing cyber offensive capabilities within their militaries. The main purpose of this *loud* signal, contrary to quiet cyber intelligence operations, is to signal the country's readiness to go beyond its national borders to punish cyber aggressors (Interview, 2018: #11). However, careful examination of these public signals shows that the stated capability is not always present. For instance, when asked for concrete details about the recruitment and training of cyber soldiers, silence, vague responses, or the excuse that many countries were finding it difficult to recruit cyber warriors into their forces followed (Interview 2018: #11, #20, # 35, #49). This discrepancy between stated and actual capabilities hint that countries use easy to observe but difficult to verify cyber institutions to make their adversaries overestimate their existing cyber capabilities in the hopes of deterring them from attacking.

Second is a strategic separation of strong cyber nations where despite their significant cyber capabilities, cyber powers continue to publicly invest in their cyber institutions to differentiate themselves from weaker nations. For instance, Russia has established information warfare units (Reuters, 2017) and is committed to invest between $200 million and $250 million USD per year to significantly strengthen its cyber-offensive capabilities, and to create a cyber-deterrent

that "will equate to the role played by nuclear weapons" (Gerden, 2016). Such additional investments and multiple cyber institutions signal "mostly failure" of the already implemented efforts (Interview, 2018: #20). If deterrence had worked and the country was ready and confident in its ability to defend itself in cyberspace, it would not need "to make [any additional] noise" (Interview, 2018: #20).

My interviewees profess the belief that deterrence is working in the case of strategic cyber-attack scenarios but remain skeptical of cyber institutions as an effective deterrent mechanism, citing the difficulty of demonstrating this effect empirically as a major challenge (Interview 2018: #48, 49). They point to two signs of successful deterrence. First, the decision to design cyber weapons with more care and precision shows that states have started practicing some restraint as more nations become cyber-dependent (e.g., built-in restrictions in the WannaCry and NotPetya attacks). Second, cyber powers have changed their cyber strategies. With an increase in China's reliance on the Internet, resulting in an increase in China's own cyber vulnerability, the country has changed its cyber force posture from brinkmanship to calibrated escalation, signaling to its adversaries that it wants to avoid full-scale retaliation (Cunningham, 2018).

3.2 2018 Swedish National Elections: Deterrence works

The 2018 Swedish national elections fall into the top middle left region of Table 2 ($P_W < \bar{P} < P_S$) Sweden, as a middle power, was able to create an impression that it possessed significant cyber defenses and offenses and as a result was able to deter the Russian government through its cyber institutions. I argue that before these institutions were put into place, Russia was considering to interfere in Sweden's elections; however, the new cyber institutions made Russia overestimate Sweden's cyber capability, effectively deterring Russia.

Why did Russia consider interfering in the Swedish electoral process? Sweden's non-alignment policy has always served as a guarantee of Russia's security. Recently, however, Sweden shifted

its military non-alignment position by strengthening its international defense cooperation with NATO (Kunz, 2015). In response, Russia's defense minister Sergei Shoigu described Sweden's involvement in NATO activities as "worrying" and added that "such steps...[were] forcing us to take response measures" (*Russia Concerned by Efforts to Draw Finland, Sweden Into NATO - Defense Minister*, 2018). Military actions and/or economic sanctions are possible response measures, although they have not been pursued. The Ukrainian and Syrian conflicts, combined with NATO-Swedish defense cooperation (even if it falls short of collective defense) have most likely prevented Russia from pursuing the military option.

Russia is, on the other hand, a mastermind of influence campaigns and has been preparing a strong foundation for such a campaign on the Swedish population for some time. Starting in 2014, the Swedish information landscape witnessed an increase in disinformation campaigns, led by trolls, bots, and Kremlin-sponsored media outlets, such as *Sputnik International* (Kragh and Åsberg, 2017, 774). Even after the Sweden-targeted version of *Sputnik International* was terminated in the spring of 2016, other cyber and information campaigns aimed at influencing the Swedish public opinion continued. Russian actors were behind a series of distributed denial-of-service (DDoS) attacks[24] against Swedish news sites and a disinformation campaign about NATO in the Swedish media.[25] Moreover, in 2016, the Swedish authorities reported an increase in information campaigns aimed at "polarizing Swedish society, undermining stability, and spreading falsehoods" leading up to the 2018 Swedish elections (Cederberg, 2018, 5). A 2018 U.S. Senate report confirms this by providing evidence that Sweden remained one of the few "favorite target[s] of the Kremlin's propaganda machine" (*Putin's Assymetric Assault on Democracy in Russia and Europe*, 2018, 109).

These ongoing cyber and information operations were bolstered by a political climate of anti-immigration sentiment in both the Swedish parliament and populace. Immigration has

[24]These attacks flood a website with multiple requests, making it crash.

[25]See https://www.documentcloud.org/documents/4627057-16-2517-CKK-2017-09-15-State-Production-3.html

always been a contentious issue in Swedish politics, recently exacerbated by the Syrian refugee crisis. Over the last five years, Sweden, a country of ten million, has welcomed 165,000 asylum-seekers from the Middle East (Cerrotti, 2017). To demonstrate how polarizing the topic of immigration is in the Swedish politics and among its population, let us look at the Sweden Democrats. Using immigration as one of its agenda items, the far-right Sweden Democrats became the third largest party in the Swedish parliament in 2018 after having occupied only two seats in 2010 (Johnson and Evans, 2018). The party's anti-immigration stance and the divide in the general population over the immigration issue presented the perfect canvas for the Kremlin's influence campaigns.

However, by 2018, in the immediate lead-up to the election, Kremlin-linked disinformation campaigns seemed to fade, and there was no outright election interference (Cederberg, 2018). Apparently, Moscow decided not to aid anti-NATO Swedish political parties, such as the Sweden Democrats in the 2018 election. The question is *why*? What stopped Russia from interfering in the Swedish elections?

Could NATO's military capabilities have deterred this military and cyber power? It is unlikely because the NATO-Sweden cooperation defense pact prioritizes security in the Baltic Sea region and the "develop[ment of] interoperable capabilities and maintain[ance of] the ability of the Swedish Armed Forces to work with those of NATO and other partner countries in multinational peace-support operations" (*Relations with Sweden*, 2018). In other words, this pact is meant to protect Sweden from a physical invasion by Russia (akin to those in Ukraine and Georgia), but not from election interference. Thus, Sweden had to rely on its own military as its government most likely lacked the political will to use a military response to Russia's influence campaign, even if such a campaign meant a violation of sovereignty in the form of election interference.

In addition to NATO's and the Swedish military's capabilities, economic sanctions by the

West were also unlikely to succeed in deterrence. This is because the conflicts in Ukraine and Syria have taught the West a valuable lesson in the failure of economic sanctions to change Moscow's behavior (Gricius, 2018). If neither military nor economic options deterred Russia, could Swedish cyber institutions have affected Russia's calculus of attack?

Even prior to the 2016 U.S. election interference, the Swedish government took significant publicly observable steps to improve its cyber defense and offense in order to deter future information and cyber operations. Having witnessed Russia's interference in the U.S. elections, Sweden assumed that its 2018 election would be the Kremlin's next target and substantially increased their already ongoing efforts.

Swedish defense started with clearly defined priorities. As early as 2015, Sweden's Defense Policy named protection of democracy as one of Sweden's security objectives (*Sweden's Defense Policy: 2016 to 2020*, 2015) — an objective reiterated in two national cybersecurity strategies released in the first half of 2017 (*National Cybersecurity Policy*, 2017). The documents laid out a series of measures aimed at creating better cyber defenses to "raise the threshold [of] attacking Sweden" (*National Cybersecurity Policy*, 2017, 19). Importantly, Sweden has designated its election systems as critical infrastructure.

The government also increased the budgets of existing agencies to include election protection into their scope, and it established new agencies, forums, and programs to protect against election interference and disinformation campaigns. The Swedish government's crisis preparation and response agency became the main authority for election coordination. Through a special Cabinet decision, the Swedish Civil Contingencies Agency (MSB) together with the Swedish Security Services and the Election Authority was tasked with coordinating election protection. The Swedish Agency for Public Management became responsible for "coordinating Swedish public agencies which could carry out 'psychological defense"' (SverigesRadio, 2017) during peacetime to "improve the ability of Swedish society to withstand pressure from a potential

opponent" (*Sweden's Defense Policy: 2016 to 2020*, 2015, 5).

Swedish security services (SÄPO), the Swedish Police Authority, the Election Authority, and MSB established a high-level national forum, responsible for briefing election administrators on potential threats (Cederberg, 2018). Having created a confidential report using past cases of election interference, this forum traveled throughout the country to educate local election administrators about how to better protect against these cyber threats (Cederberg, 2018, 15). Similar education campaigns were undertaken by media outlets and political parties (Sveriges-Radio, 2017) and were also carried out in schools (Roden, 2017). Prepared for total defense awareness, including a cyber emergency, MSB produced a pamphlet titled "If Crisis or War Comes" and "sent it to all 4.7 million Swedish households" (Brattberg and Maurer, 2018*a*, 29). During all this preparation, SÄPO and other governmental agencies were relatively open and transparent about the initiatives they undertook to address potential interference. Such clear communication might have increased public awareness and the likelihood that Swedes would practice better cyber hygiene (Brattberg and Maurer, 2018*a*).

In addition to building better defenses, Sweden invested significant resources in improving the cyber capabilities of its intelligence agencies to detect external threats and of its military forces to respond to them (Rettman and Kirk, 2018). The National Defence Radio Establishment (FRA) and the Military Intelligence and Security Service (MUST) — both responsible for signals intelligence — installed special detection and warning systems to guard against foreign powers hacking into sensitive agencies. In preparation for elections, the government increased expenditure on signals intelligence and added to it new projects that included "the development of advanced offensive smart technologies and tools that have the capacity to weaponise counter-strike actions against...perpetrators" (O'Dwyer, 2018).

Sweden also strengthened its military posture. For the first time in more than two decades, the Swedish government decided to substantially increase its defense budget, some of which

was to be spent on active cyber capabilities (*Sweden's Defense Policy: 2016 to 2020*, 2015, 4-5). The country re-introduced military conscription, with some of these new recruits contributing to the cyber work force. Most importantly, during the country's preparatory efforts to deal with any potential election interference by foreign powers, Sweden's Prime Minister Stefan Löfven emphasized the country's military cyber offensive capability and the government's willingness to use it. For instance, when discussing a three-point plan to stop foreign powers from influencing the 2018 Swedish elections, Löfven publicly claimed that the Swedish Armed Forces were capable of carrying "active operations in the cyber environment" (SverigesRadio, 2017). At a security conference in January 2018, Löfven clearly communicated the country's willingness to act in case of election interference: "To those thinking about trying to influence the outcome of the elections in our country: Stay away!" (as quoted in Cederberg (2018, 11)).

The Swedish government's persistence, drive, and transparency in establishing its cyber institutions aimed at protecting its elections most likely convinced Russia that an influence campaign would have been too costly. But this interference would not have been as costly as Russia likely expected. First, there is a discrepancy between the Swedish government's commitments and the implementation of these commitments (Cederberg, 2018, 29). For instance, despite the Swedish government's announced intention, not only was there no psychological defense agency created prior to the elections, the government did not even appoint an investigator responsible for determining this agency's scope. Second, government initiatives did not always translate into more resilient cyber defense capability. For instance, even though Sweden's security agencies have been publicly working with political parties and media outlets to increase their awareness of how to deal with cyber and information threats, the information these agencies presented did not necessarily address the needs of campaign officials and journalists. The same concern applies to the Swedish public. Even factoring in Sweden's relatively small population, it has never become clear if and how public awareness campaigns would translate into

behavioral change. Kostyuk and Wayne (2019) demonstrate that respondents fail to engage in safer online behavior even though they intended to do so when they were exposed to a cyber attack. For these reasons, it appears that Sweden was able to deter Russia from interfering in its 2018 elections by making the Kremlin overestimate Sweden's existing cyber capability.

4 Discussion and Implications

This study has revealed several important patterns of the strategic use of cyber institutions to deter challengers. To begin with, a challenger's choice to attack a defender via cyberspace is independent of the defender's cyber institutions in two equilibria. In the first, the challenger has no interest in attacking the defender, giving the false impression of deterrence success (the left region of Table 2), while in the second, the challenger has decided to attack the defender even before observing her cyber institutions, giving the false impression of deterrence failure (the right region of Table 2). Inadequate signaling or factors such as domestic politics, budgetary and legal constraints, or organizational and strategic culture, might explain these two scenarios. Researchers and policy-makers should consider these situations in their analyses, before concluding the effectiveness or ineffectiveness of cyber institutions for deterrence. They should also carefully consider the link between the defender's cyber institutions and the challenger's willingness to attack — even if a cyber institution gives the challenger information about the defender that he did not previously possess, not all challengers will base their decision of whether to cyber-attack on this information.

This model also demonstrates that cyber institutions can indeed play a strategic role. In the left middle region of Table 2, the challenger only attacks the defender if he perceives her as cyber weak and avoids attacking if he perceives her as cyber strong. As a result, weak cyber nations choose to over-invest their limited resources into public cyber institutions and under-invest in covert cyber activity to appear strong. Strong cyber nations, in their turn,

over-invest their limited resources into cyber institutions to distinguish themselves from weak cyber nations pretending to be cyber strong. This sub-optimal distribution of limited resources, which makes the defender weaker in her overall cyber capability, can be worth-while because it may deter the challenger. These results have important theoretical and policy implications.

First, they shed light on a theoretical debate surrounding cyber deterrence. Similar to Tor (2017), this article stresses the need to re-think our reference to absolute nuclear deterrence as a matrix of deterrence success for cyber operations. Countries do not create cyber capability to deter low-level cyber operations; instead, their goal is to stop adversaries from executing strategic cyber attacks, which can cause detrimental damage to the country's economy, prosperity, and security. As the cyber threat landscape grows and changes, states tend to update their definitions of strategic cyber attacks to reflect this change. Less than a decade ago, for instance, countries focused only on the protection of their critical infrastructure (*Presidential Policy Directive 21*, 2013). Following the 2016 U.S. elections, many nations added election protection to their top national security priorities (*National Cybersecurity Policy*, 2017).

As countries constantly re-define what constitutes strategic cyber attacks, adversaries become more creative in the execution of cyber operations aimed at achieving their strategic goals. For example, in response to Russia's meddling in the 2016 U.S. elections, the U.S. government took significant steps to protect its 2018 elections. In response to this stern measure, Russian bots and trolls adjusted their behavior and started operating during the election off-season. For instance, there was a spike in Russian bot and troll tweets in the summer after the 2016 U.S. election (Roeder, 2018). These influence campaigns, even if conducted during election off-seasons, shape public opinion and might affect public voting behavior. Emerging digital technologies, such as artificial intelligence, bring a new set of challenges that governments should be prepared to address. "Deep learning" technology, a method in which computers learn how to solve certain tasks based on the analysis of large information sets, allows

to automatically create fake images and videos that are indistinguishable from real content.

Second, over-investment in public cyber capability demonstrates that states are changing their deterrent tactic. Over the last two decades, states have mainly invested in their secret cyber capability to deter adversaries. But this tactic is inefficient because: it is costly—the value of cyber operations diminishes after their first use; and ineffective due to the difficulty of cyber attribution. Over-investment in public cyber capability, on the other hand, allows even weak cyber nations to deter their strongest adversaries.

While signaling via public cyber capability might deter in limited cases, it is not clear how long this tactic will be effective. By creating cyber institutions, nations sometimes purposefully make their overall cyber capability weaker to appear stronger in public. With time, weak cyber nations attempting to imitate strong cyber nations are more likely to be exposed as weak nations and their cyber institutions will become less effective in deterring adversaries. For example, in 2012, Norway, as a middle power, announced the creation of a cyber command, but by 2018, this command was not close to becoming operational. Thus, any future cyber institutions announced by Norway may not deter its adversaries.

This example also demonstrates that states should take the signaling of cyber capability via cyber institutions with a grain of salt. While cyber institutions serve as a cyber threat assessment barometer because they allow a challenger to estimate a state's ability to conduct cyber operations, the state's willingness to use these operations, and the scope of a potential retaliation by the state, this assessment is not precise. In addition to public cyber institutions, challengers should examine other indicators, such as economic and technological achievements and the government's reliance on the private sector for cyber capability, to better estimate the state's existing cyber capability. This cumulative approach to estimate the state's cyber capability will help governments better evaluate options that minimize the risk of escalation.

My approach of studying deterrence by cyber institutions is not without limitations. My

model oversimplifies real-world scenarios by making assumptions to identify the causal effect of cyber institutions on deterrence chances. Specifically, my model views the challenger's decision to attack and the defender's choice to establish cyber institutions as a one-time decision. In practice, cyber institutions present a state's cumulative effort to boost its cyber capability. Similarly, influence campaigns, for example, are composed of many small campaigns that span an extended period. As a result, it is not easy to distinguish between situations in which deterrence by cyber institutions fails to deter election interference and those in which cyber institutions were not even considered by the challenger. This is because it is hard to determine how much updating-of-beliefs took place during different stages of the influence campaign.

Moreover, my model studies the immediate deterrence scenarios in which the challenger is contemplating a cyber-attack against the defender. In this high-stakes scenario, I assume that the defender's resolve is high. Future reiterations of this model should relax this assumption, and, perhaps, consider general deterrence scenarios to separate resolve from capability and study the individual, potentially diverging effects of these two factors on deterrence success. For example, one could study the scenario in which cyber institutions signal a state's lack of resolve because it chooses to invest in cyber institutions — a more ambiguous option of signaling cyber capabilities — instead of other, more precise signaling options.

Lastly, my model equates the defender's probability of successfully retaliating against the challenger with the probability that her cyber defenses hold. When the government creates a cyber institution, this institution often signals an increase in both offensive and defensive cyber capabilities (Schneider, 2019). But it might be worth investigating scenarios in which this is not necessary the case. For instance, if a cyber institution only signals an increase in cyber offense, the threat of cyber retaliation might not deter a country that does not have many cyber targets, like North Korea, but might deter a country with many cyber targets, like the United States. A cyber institution that only signals an increase in cyber defenses, on the other hand, might deter

countries that view attacking well-protected targets as too costly. As a result, by practicing both the deterrence by prevention and by the threat of punishment, the country maximizes its chances at deterrence because it increases the cost of attacking for all attackers.

Even though states usually attempt to deter by both prevention and the threat of punishment, nations should simultaneously work on their cyber strategies in addition to building the capability to maximize the desired deterrent effect. For instance, by building offensive cyber capabilities, government officials seem to assume that "cyber capabilities alone have a deterrent effect without taking into consideration the strategic requirements that come with deterrence by the threat of punishment, namely credibly holding assets at risk and signaling desired behavior while being willing to face consequences in case of an escalation" (Schulze and Herpig, 2018). But it is often unclear whether political will exists to launch a retaliatory cyber attack against, for example, Russia or China, and face the potential consequences of entering an escalation cycle with these adversaries. "If deterrence fails it is usually because someone thought he saw an 'option' that the [...] government had failed to dispose of, a loophole that it hadn't closed against itself" (Schelling, 2008, 44). As countries try to flex their cyber muscles, they must spell out their cyber strategies in order to close the loopholes that adversaries can exploit. Until this is done, the pessimistic view of cyber deterrence will persist.

References

Alperovitch, Dmitri. 2016. "Bears in the midst: Intrusion into the Democratic National Committee." *CrowdStrike Blog* 15.

Baliga, Sandeep, SOM Kellogg, Ethan Bueno de Mesquita and Alexander Wolitzky. 2018. Deterrence with imperfect attribution. Technical report mimeo, 2018. URL http://home. uchicago. edu/˜ bdm/PDF/deterrence. pdf.

Ball, Desmond. 1993. "Arms and affluence: military acquisitions in the Asia-Pacific region." *International Security* 18(3):78–112.

Borghard, Erica D and Shawn W Lonergan. 2017. "The Logic of Coercion in Cyberspace." *Security Studies* 26(3):452–481.

Brantly, Aaron. 2018. "Conceptualizing Cyber Deterrence by Entanglement.".

Brantly, Aaron Franklin. 2016. *The Decision to Attack: Military and Intelligence Cyber Decision-making*. University of Georgia Press.

Brattberg, Erik and Tim Maurer. 2018*a*. "How Sweden is preparing for Russia to hack its election." *BBC News* .

Brattberg, Erik and Tim Maurer. 2018*b*. *Russian Election Interference: Europe's Counter to Fake News and Cyber Attacks*. Vol. 23 Carnegie Endowment for International Peace.

Cederberg, Gabriel. 2018. "Catching Swedish Phish: How Sweden is Protecting its 2018 Elections." *Defending Digital Democracy Project, Belfer Center for Science and International Security* .

Cerrotti, Rachel. 2017. "Sweden was among the best countries for immigrants. That's changing." *PRI* .

Council, National Research et al. 2009. *Technology, policy, law, and ethics regarding US acquisition and use of cyberattack capabilities*. National Academies Press.

Cunningham, Fiona. 2018. "Maximizing Leverage: Explaining Chinas Cyber Force Posture.".

Fearon, James. 2002. "Selection effects and deterrence." *International Interactions* 28(1):5–29.

Fearon, James D. 1995. "Rationalist explanations for war." *International organization* 49(3):379–414.

Fernandino, Lisa. 2018. "Cybercom to Elevate to Combatant Command." *U.S. Department of Defense* .

France says no trace of Russian hacking Macron. 2017.

Galante, Laura and Shaun EE. 2018. *Defining Russian Election Interference: An analysis of select 2014 to 2018 cyber enabled incidents*. Atlantic Council.

Gartzke, Erik. 2013. "The myth of cyberwar: bringing war in cyberspace back down to Earth." *International Security* 38(2):41–73.

Gartzke, Erik and Jon R Lindsay. 2015. "Weaving tangled webs: offense, defense, and deception in cyberspace." *Security Studies* 24(2):316–348.

Gerden, Eugene. 2016. "Russia to spend 250 million US dollars strengthening cyber-offensive capabilities.".

Gricius, Gabriella. 2018. "Why Russian Sanctions are Ineffective." *Global Security Review* .

Interview. 2018. "Interviews on Cyber Institutions as a Deterrent.".

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton, Princeton University Press.

Johnson, Simon and Catherine Evans. 2018. "Anti-immigration Sweden Democrats poll record high ahead of September election." *Reuters* .

Joint Publication 3 13 Information Operations. 2014.

Kostyuk, Nadiya and Carly Wayne. 2019. "Communicating Cybersecurity: Citizen Risk Perception of Cyber

Threats.".

Kostyuk, Nadiya and Yuri M. Zhukov. 2019. "Invisible Digital Front: Can cyber attacks shape battlefield events?" *Journal of Conflict Resolution* 63:317–347.

Kragh, Martin and Sebastian Åsberg. 2017. "Russias strategy for influence through public diplomacy and active measures: the Swedish case." *Journal of Strategic Studies* 40(6):773–816.

Kunz, Barbara. 2015. "Swedens NATO workaround: Swedish security and defense policy against the backdrop of Russian revisionism." *Enote. Focus Strateguque* .

Laudrain, Arthur P.B. 2019. "Frances New Offensive Cyber Doctrine." *Lawfare* .

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

Lindsay, Jon R. 2013. "Stuxnet and the limits of cyber warfare." *Security Studies* 22(3):365–404.

Lindsay, Jon R and Erik Gartzke. 2015. "Coercion through Cyberspace: The Stability-Instability Paradox Revisited." *Typescript, University of California, San Diego* .

Lupovici, Amir. 2018. "Toward a Securitization Theory of Deterrence." *International Studies Quarterly* .

Mueller, RS. 2019. "Report on the investigation into Russian interference in the 2016 presidential election." *US Department of Justice* .

National Cybersecurity Policy. 2017.

Nye Jr, Joseph S. 2017. "Deterrence and Dissuasion in Cyberspace." *International Security* 41(3):44–71.

O'Dwyer, Gerard. 2018. "Sweden steps up cyber defence measures.".

Powell, Robert. 2002. "Bargaining theory and international conflict." *Annual Review of Political Science* 5(1):1–30.

Presidential Policy Directive 21. 2013.

Press, Daryl Grayson. 2005. *Calculating credibility: How leaders assess military threats*. Cornell University Press.

Public Elements of the Doctrine on Military Cyber Offensive. 2019.

Putin's Assymetric Assault on Democracy in Russia and Europe. 2018.

Reed, William. 2003. "Information, power, and war." *American Political Science Review* 97(4):633–641.

Relations with Sweden. 2018.

Rettman, Andrew and Lisbeth Kirk. 2018. "Sweden raises alarm on election meddling." *EU Observer* .

Reuters. 2017. "Russia sets up information warfare units - defence minister." *Reuters* .

Rid, Thomas. 2013. *Cyber war will not take place*. Oxford University Press.

Rid, Thomas and Ben Buchanan. 2015. "Attributing cyber attacks." *Journal of Strategic Studies* 38(1-2):4–37.

Riley, Michael and Jordan Robertson. 2017. "Russian cyber hacks on US electoral system far wider than previously known." *Bloomberg* 13.

Roden, Lee. 2017. "Sweden's government wants newspapers to pay less tax in an effort to combat fake news." *The Local* .

Roeder, Ollie. 2018. "Why Were Sharing 3 Million Russian Troll Tweets.".

Russia Concerned by Efforts to Draw Finland, Sweden Into NATO - Defense Minister. 2018.

Schelling, Thomas C. 2008. *Arms and influence: With a new preface and afterword*.

Schneider, Jacquelyn. 2019. Cyber and Cross Domain Deterrence: Deterring Within and From Cyberspace. In *Cross-Domain Deterrence: Strategy in an Era of Complexity*, ed. Jon R. Lindsay and Erik Gartzke. Oxford University Press chapter 5.

Schulze, Matthias and Sven Herpig. 2018. "Germany Develops Offensive Cyber Capabilities Without A Coherent Strategy of What to Do With Them." *Council on Foreign Relations* .

Segal, Adam. 2014. "What Briefing Chinese Officials On Cyber Really Accomplishes." *Forbes* .

Smith, Alastair and Allan C Stam. 2004. "Bargaining and the Nature of War." *Journal of Conflict Resolution* 48(6):783–813.

Soldatov, Andrei and Irina Borogan. 2017. *The Red Web: The Struggle Between Russia's Digital Dictators and the New Online Revolutionaries*. Public Affairs.

SverigesRadio. 2017. "Swedish PM warns of foreign influence ahead of 2018 poll." *Radio Sweden* .

Sweden's Defense Policy: 2016 to 2020. 2015.

Tor, Uri. 2017. "Cumulative Deterrenceas a New Paradigm for Cyber Deterrence." *Journal of Strategic Studies* 40(1-2):92–117.

U.S. Department of Defense Cyber Strategy. 2015.

USDepartmentOfJustice. 2014. "U.S. Charges Five Chinese Military Hackers for Cyber Espionage Against U.S. Corporations and a Labor Organization for Commercial Advantage.".

USDepartmentOfJustice. 2018*a*. "Grand Jury Indicts 12 Russian Intelligence Officers for Hacking Offenses Related to the 2016 Election.".

USDepartmentOfJustice. 2018*b*. "North Korean Regime-Backed Programmer Charged With Conspiracy to Conduct Multiple Cyber Attacks and Intrusions.".

USDepartmentOfJustice. 2018*c*. "United States of America v. Internet Research Agency LLC et al.".

Valeriano, Brandon, Benjamin Jensen and Ryan C. Maness. 2018. *Cyber Strategy: The Evolving Character of Power and Coercion*. Oxford University Press.

Valeriano, Brandon and Ryan C Maness. 2014. "The dynamics of cyber conflict between rival antagonists, 2001–11." *Journal of Peace Research* 51(3):347–360.

Online Appendix:
Mathematical Proofs, Interviews, and Case Studies

"Deterrence in the Cyber Realm: Public versus private cyber capability"

Nadiya Kostyuk
University of Michigan

September 15, 2019

Contents

1 Proofs

I am using the following definitions in the proofs:

- Let's define $I_\theta(r) = I$, then $r = I_\theta^{-1}(I)$.
- $c_\theta(I) \equiv I + (1 - I_\theta^{-1}(I))n$ is type θ's cyber capability if she chooses CBI level, I;
- $I_{\sigma_\theta} \equiv \{I : \sigma_\theta > 0\}$ is set of CBI for type θ of D that will be chosen with positive probability under strategy σ_θ;
- $I_\theta(\bar{P}) \equiv \{I : p(c_\theta(I)) > \bar{P}\}$ is the set of CBI for type θ of D that leads to successful retaliation probability that is higher than $\bar{P}$.

Due to the characteristics of this game, there might exist a lot of equilibria. Therefore, I introduce the notion of *outcome equivalence*. Two equilibria — $(\sigma_S, \sigma_W, \sigma_{R_S}, \sigma_{R_W}, \sigma_A, \mu)$ and $(\sigma'_S, \sigma'_W, \sigma'_{R_S}, \sigma'_{R_W}, \sigma'_A, \mu')$ — are outcome-equivalent if the expected payoffs of each player are the same under these two equilibria (Fudenberg and Tirole, 1991). The game has a unique equilibrium outcome if any two equilibria are outcome-equivalent. Therefore, I can pick one representative equilibrium if the game has a unique equilibrium outcome.

Defender's choice to retaliate. Examining D's choice to retaliate, I have the following utilities:

$$EU_D(\neg Retaliate) = -C_D, \text{and}$$

$$EU_D(Retaliate) = P_\theta(V_R - C_D - C_R) + (1 - P_\theta)(-C_D - C_R)$$

D is indifferent between retaliating and not retaliating when $EU_D(Retaliate) = EU_C(\neg Retaliate)$, i.e.,

$$P_\theta(V_R - C_D - C_R) + (1 - P_\theta)(-C_D - C_R) = -C_D$$

$$P_\theta V_R - P_\theta C_D - P_\theta C_R - C_D - C_R + P_\theta C_D + P_\theta C_R = -C_D$$

$$P_\theta V_R = C_R, \tag{1}$$

where

- P_θ is the probability that D_θ successfully retaliates, which is determined by D_θ's overall cyber capability and how she distributes her resources at Stage 1,
- V_R is the value that D_θ receives from successful retaliation,
- C_D is D_θ's cost of being attacked, and
- C_R is the cost that D_θ pays for retaliating.

Equation 1 depicts two interesting results. First, it demonstrates that C_D—D_θ's cost of being attacked—does not influence her decision to retaliate. Second, it shows that C_R increases when $P_\theta V_R$ increases. C_R can increase because (1) P_θ increases and V_R decreases, (2) P_θ decreases and V_R increases, or (3) both P_θ and V_R increase.

The first case depicts the unlikely scenario where D's retaliation value is low. But D invests significant resources to retaliate against C, leading to D's high probability of successful retaliation.

The second case depicts the scenario where D's retaliation value is high. But it is costly to retaliate. As a result, the probability of successful retaliation is low. The last case depicts the scenario where D's retaliation value is high. D invests significant resources to retaliate against C, leading to D's high probability of successful retaliation.

Challenger's choice to attack. Examining C's choice of action, I have the following utilities:

$$EU_C(\neg Attack) = R, \text{and}$$

$$\begin{aligned} EU_C(Attack) &= P_B\Big[\sigma_{R_\theta}\big[P_A(V_C - C_P - C_C) + (1-P_A)(V_C - C_C)\big] + (1-\sigma_{R_\theta})(V_C - C_C)\Big] + (1-P_B)(-C_C) = \\ &= P_B\Big[\sigma_{R_\theta}\big[P_A V_C - P_A C_P - P_A C_C + V_C - C_C - P_A V_C + P_A C_C\big] + \\ &\quad + (1-\sigma_{R_\theta})(V_C - C_C)\Big] + (1-P_B)(-C_C) = \\ &= P_B\Big[\sigma_{R_\theta}\big[-P_A C_P + V_C - C_C\big] + (1-\sigma_{R_\theta})(V_C - C_C)\Big] + (1-P_B)(-C_C) = \\ &= P_B\Big[-\sigma_{R_\theta} P_A C_P + \sigma_{R_\theta} V_C - \sigma_{R_\theta} C_C + V_C - C_C - \sigma_{R_\theta} V_C + \sigma_{R_\theta} C_C\Big] + (1-P_B)(-C_C) = \\ &= -P_B \sigma_{R_\theta} P_A C_P + P_B V_C - P_B C_C - C_C + P_B C_C = -P_B \sigma_{R_\theta} P_A C_P + P_B V_C - C_C \end{aligned}$$

C is indifferent between attacking and not attacking when $EU_C(Attack) = EU_C(\neg Attack)$, i.e.,

$$-P_B \sigma_{R_\theta} P_A C_P + P_B V_C - C_C = R \tag{2}$$

Because the model assumes that $P_A = P_B$,

$$-P_A^2 \sigma_{R_\theta} C_P + P_A V_C - C_C = R, \text{or}$$

$$P_A^2 \sigma_{R_\theta} C_P - P_A V_C + C_C + R = 0$$

If $\sigma_{R_\theta} C_P = 0$, then solving for P_A, we obtain

$$\begin{aligned} -P_A V_C + C_C + R &= 0 \\ P_A &= \frac{C_C + R}{V_C} \end{aligned} \tag{3}$$

If $\sigma_{R_\theta} C_P \neq 0$, solving this for P_A, we obtain

$$P_A = \frac{V_C \pm \sqrt{V_C^2 - 4\sigma_{R_\theta} C_P (C_C + R)}}{2\sigma_{R_\theta} C_P}$$

Without loss of generality, I assume C's reservation utility, R, is equal to 0. As a result,

$$P_A = \frac{V_C \pm \sqrt{V_C^2 - 4\sigma_{R_\theta} C_P C_C}}{2\sigma_{R_\theta} C_P}. \tag{4}$$

Because P_A is a probability, two clarifications should be made about Equation 4. First, I assume that $V_C^2 \geq 4\sigma_{R_\theta} C_P C_C$, so that both solutions are real numbers. Second, because all values in this equation are distributed between 0 and 1, I only focus on the positive solution of Equation 4 and I

denoted it by $\bar{P}$. $\bar{P}$ can be interpreted as the cut-off point for D's successful retaliation probability when C is indifferent between attacking and not attacking.

C's choice of action only depends on the relationship between his critical value of being indifferent between attacking or not attacking $\bar{P}$ and his perceived probability of D's successful retaliation against him, P_A, given that he observes CBI. Here, $P_A = p_S(I)\mu(I) + p_W(I)(1-\mu(I))$, where $\mu(I)$ is C's belief that D is strong type and $p_\theta(I) = p(c_\theta(I))$ for $\theta \in \{S, W\}$ is D_θ's successful retaliation probability, given the cyber capability that this type obtains, having implemented CBI. Because this cyber capability is not always optimal, $p_\theta \leq \hat{P}_\theta$, where $\hat{P}_\theta$ is D_θ's maximum probability of successful retaliation. Specifically, we have

1. $P_A > \bar{P}$, C does not attack
2. $P_A < \bar{P}$, C attacks
3. $P_A = \bar{P}$, C mixes between attacking and not attacking.

I consider the following three cases.

Case 1: $\bar{P} < \hat{P}_W < \hat{P}_S$. I am going to show that in this region the unique equilibrium outcome is: C does not attack, D_θ get her first-best outcomes — D_θ deters C and D_θ does not retaliate.

First, I consider potential separating equilibria in which $\mathcal{I}_{\sigma_S} \cap \mathcal{I}_{\sigma_W} = \emptyset$. Then I know C's belief $\mu(I)$ is

$$\mu(I) = \begin{cases} 1, & \text{for } I \in \mathcal{I}_{\sigma_S} \\ 0, & \text{for } I \in \mathcal{I}_{\sigma_W} \\ [0,1], & \text{o.w. (off the equilibirum path)} \end{cases}.$$

The corresponding on-path P_A is

$$P_A = \begin{cases} p(c_S(I)), & \text{for } I \in \mathcal{I}_{\sigma_S} \\ p(c_W(I)), & \text{for } I \in \mathcal{I}_{\sigma_W}. \end{cases}.$$

Since $\bar{P} < \hat{P}_W < \hat{P}_S$, $\mathcal{I}_\theta(\bar{P}) \neq \emptyset$, for both types of D. Each type's best response is choosing any level of CBI within the set $\mathcal{I}_\theta(\bar{P})$ which leads to $P_A > \bar{P}$. Hence, C does not attack on the equilibrium path. Both types of D get their first-best outcomes and have no incentives to deviate. Therefore, it is a equilibrium.

Second, I consider equilibria in which $\mathcal{I}_{\sigma_S} \cap \mathcal{I}_{\sigma_W} \neq \emptyset$. Then I know C's belief $\mu(I)$ is

$$\mu(I) = \begin{cases} \frac{q\sigma_S(I)}{q\sigma_S(I)+(1-q)\sigma_W(I)}, & \text{for } I \in \mathcal{I}_{\sigma_S} \cup \mathcal{I}_{\sigma_W} \\ [0,1], & \text{otherwise} \end{cases}.$$

The corresponding on-equilibrium path P_A is $P_A = p_S(I)\mu(I) + p_W(I)(1-\mu(I))$. I know that if both types of D choose CBI from their own $\mathcal{I}_{\sigma_\theta}$ set, then for any belief $\mu(I)$, $P_A = p_S(I)\mu(I) + p_W(I)(1-\mu(I)) \Rightarrow P_A > \bar{P}\mu(I) + (1-\bar{P})\mu(I) \Rightarrow P_A > \bar{P}$. Hence C does not attack. Both types of D get their first-best outcomes and have no incentives to deviate. A representative equilibrium

of the above unique equilibrium outcome is: $\sigma_S(\hat{I}_S) = \sigma_W(\hat{I}_W) = 1$, $\sigma_{R_\theta}(I) = 0$, and $\sigma_A(I) = 0$ for $I \in \{\hat{I}_S, \hat{I}_W\}$, $\mu(\hat{I}_S) = 1, \mu(\hat{I}_W) = 0$, $\sigma_A(I) \in [0,1]$ for $I \notin \{\hat{I}_S, \hat{I}_W\}$, $\mu(I) \in [0,1]$ for $I \notin \{\hat{I}_S, \hat{I}_W\}$.

Case 2: $\hat{P}_W < \bar{P} < \hat{P}_S$. In this case, D_W has an incentive to imitate D_S. I consider two different assumptions in this case. I define $H \equiv \mathcal{I}_S(\bar{P}) \cap (I_W(1), I_S(1)]$. If $H \neq \emptyset$, D_W cannot imitate any CBI in H because it is beyond her capability for any $I \in (I_W(1), I_S(1)]$, while choosing $I \in H$ is both doable and profitable for D_S. If $H = \emptyset$, D_W can imitate CBI in H.

Case 2.1: $H = \emptyset$. Under this assumption, there is no separating equilibria. Suppose there exists a separating equilibrium in which $\mathcal{I}_{\sigma_S} \cap \mathcal{I}_{\sigma_W} = \emptyset$. Then we know C's belief $\mu(I)$ is

$$\mu(I) = \begin{cases} 1, & \text{for } I \in \mathcal{I}_{\sigma_S} \\ 0, & \text{for } I \in \mathcal{I}_{\sigma_W} \\ [0,1], & \text{o.w. (off equilibirum path)} \end{cases}.$$

The corresponding on-path P_A is

$$P_A = \begin{cases} p(c_S(I)), & \text{for } I \in \mathcal{I}_{\sigma_S} \\ p(c_W(I)), & \text{for } I \in \mathcal{I}_{\sigma_W}. \end{cases}.$$

Let's consider D_W's strategy. Because $p(c_W(I)) < \bar{P}$ for any $I \in \mathcal{I}_{\sigma_W}$, C attacks D_W. In order for this to be an equilibrium, $p(c_S(I)) < \bar{P}$ for all $I \in \mathcal{I}_{\sigma_S}$, otherwise D_W will deviate to $\mathcal{I}_{\sigma_S}$. That is, $\mathcal{I}_{\sigma_S} \cap \mathcal{I}_\theta(\bar{P}) = \emptyset$. In particular, $\hat{I}_S \notin \mathcal{I}_{\sigma_S}$. Then, $\sigma'_S(\hat{I}_S) = 1$ is a profitable deviation for D_S. Specifically, D_S's expected utility when she deviates to $\mathcal{I}_{\sigma_S}$ is:

$$EU_{D_S}(\sigma'_S) = \sigma'_A\Big[\gamma(c_S(\hat{I}_S))\Big[\sigma'_{R_S}\big[p(c_S(\hat{I}_S))[V_R - C_D - C_R] + (1 - p(c_S(\hat{I}_S))(-C_D - C_R)\big] + (1 - \sigma'_{R_S})(-C_D)\Big] + (1 - \gamma(c_S(\hat{I}_S)))V_D\Big] + (1 - \sigma'_A)V_D$$

$$EU_{D_S}(\sigma'_S) = \sigma'_A\Big[\gamma(c_S(\hat{I}_S))\Big[\sigma'_{R_S}\big[p(c_S(\hat{I}_S))V_R - p(c_S(\hat{I}_S))C_D - p(c_S(\hat{I}_S))C_R - C_D - C_R + p(c_S(\hat{I}_S))C_D + p(c_S(\hat{I}_S))C_R\big] - C_D + \sigma'_{R_S}C_D\Big] + (1 - \gamma(c_S(\hat{I}_S)))V_D\Big] + (1 - \sigma'_A)V_D$$

$$EU_{D_S}(\sigma'_S) = \sigma'_A\Big[\gamma(c_S(\hat{I}_S))\Big[\sigma'_{R_S}p(c_S(\hat{I}_S))V_R - \sigma'_{R_S}C_D - \sigma'_{R_S}C_R - C_D + \sigma'_{R_S}C_D\Big] + (1 - \gamma(c_S(\hat{I}_S)))V_D\Big] + (1 - \sigma'_A)V_D$$

$$EU_{D_S}(\sigma'_S) = \sigma'_A\Big[\gamma(c_S(\hat{I}_S))\Big[\sigma'_{R_S}p(c_S(\hat{I}_S))V_R - \sigma'_{R_S}C_R - C_D\Big] + V_D - \gamma(c_S(\hat{I}_S))V_D\Big] + (1 - \sigma_A)'V_D$$

$$EU_{D_S}(\sigma'_S) = \sigma'_A\Big[\gamma(c_S(\hat{I}_S))\sigma'_{R_S}p(c_S(\hat{I}_S))V_R - \gamma(c_S(\hat{I}_S))\sigma'_{R_S}C_R - \gamma(c_S(\hat{I}_S))C_D + V_D - \gamma(c_S(\hat{I}_S))V_D\Big] + (1-\sigma_A)'V_D$$

$$EU_{D_S}(\sigma'_S) = \sigma'_A\gamma(c_S(\hat{I}_S))\sigma'_{R_S}p(c_S(\hat{I}_S))V_R - \sigma'_A\gamma(c_S(\hat{I}_S))\sigma'_{R_S}C_R - \sigma'_A\gamma(c_S(\hat{I}_S))C_D - \sigma'_A\gamma(c_S(\hat{I}_S))V_D + V_D$$

If D_S does not deviate and C attacks, then:

$$EU_{D_S}(\sigma_S) = \gamma(c_S(\hat{I}_S))\Big[\sigma_{R_S}\big[p(c_S(\hat{I}_S))[V_R - C_D - C_R] + (1-p(c_S(\hat{I}_S))(-C_D-C_R)\big] + (1-\sigma_{R_S})(-C_D)\Big] + (1-\gamma(c_S(\hat{I}_S)))V_D$$

$$EU_{D_S}(\sigma_S) = \gamma(c_S(\hat{I}_S))\sigma_{R_S}p(c_S(\hat{I}_S))V_R - \gamma(c_S(\hat{I}_S))\sigma_{R_S}p(c_S(\hat{I}_S))C_R - \gamma(c_S(\hat{I}_S))C_D + V_D - \gamma(c_S(\hat{I}_S))V_D$$

Because $\sigma'_A \in \{0,1\}$, $EU_{D_S}(\sigma'_S) \geq EU_{D_S}(\sigma_S)$. Because $c_S(\hat{I}_S) > c_S(I)$ for any $I \in \mathcal{I}_{\sigma_S}$,

$$EU_S(\sigma'_S) > \int_{I\in\mathcal{I}_{\sigma_S}} \gamma(c_S(\hat{I}_S))\sigma'_{R_S}p(c_S(\hat{I}_S))V_R\sigma_S(I)dI - \int_{I\in\mathcal{I}_{\sigma_S}} \gamma(c_S(\hat{I}_S))\sigma'_{R_S}C_R\sigma_S(I)dI - \int_{I\in\mathcal{I}_{\sigma_S}} \gamma(c_S(\hat{I}_S))C_D\sigma_S(I)dI - \int_{I\in\mathcal{I}_{\sigma_S}} \gamma(c_S(\hat{I}_S))V_D\sigma_S(I)dI + V_D$$

Now, let's consider potential pooling equilibria in which $\mathcal{I}_{\sigma_S} \cap \mathcal{I}_{\sigma_W} \neq \emptyset$. Let's pick any $\bar{I} \in \mathcal{I}_{\sigma_S} \cap \mathcal{I}_{\sigma_W}$, $\mu(\bar{I}) = \frac{q\sigma_S(\bar{I})}{q\sigma_S(\bar{I})+(1-q)\sigma_W(\bar{I})}$.

1. Suppose $P_{A(\bar{I})} < \bar{P}$, i.e., $\sigma_A(\bar{I}) = 1$. Then at least one type of D has an incentive to deviate. For example, if $\bar{I} \neq \hat{I}_S$, then $\hat{I}_S$ gives D_S a higher expected payoff than $\bar{I}$. A profitable deviation would be shifting the probability assigned to $\bar{I}$ to $\hat{I}_S$,

$$\sigma'_S(I) = \begin{cases} \sigma_S(I), & \text{for } I \notin \{\bar{I}, \hat{I}_S\} \\ 0, & \text{for } I = \bar{I} \\ \sigma_S(\bar{I}) + \sigma_S(\hat{I}_S), & \text{for } I = \hat{I}_S \end{cases}.$$

   It is easy to check that $\sigma'_S(I)$ indeed is a strategy (a probability distribution, $\int_I \sigma'_S(I)dI = 1$). Using the definition of D_S's expected payoff, we have:

$$EU_{D_S}(\sigma'_S) - EU_{D_S}(\sigma_S) = \Bigg[\sigma'_A\Big[\gamma(c_S(\hat{I}_S))\Big[\sigma'_{R_S}\big[p(c_S(\hat{I}_S))(V_R - C_C - C_R) + p(c_S(\hat{I}_S))(-C_D - C_R)\big] + (1-\sigma'_{R_S})(-C_D)\Big] + (1-\gamma(c_S(\hat{I}_S)))V_D\Big] + (1-\sigma'_A)V_D\Bigg] -$$

$$\Bigg[\gamma(c_S(\bar{I}))\Big[\sigma_{R_S}\big[p(c_S(\bar{I}))(V_R - C_C - C_R)\big] + (1-p(c_S(\bar{I}))(-C_D - C_R)\Big] + (1-\sigma_{R_S})(-C_D)\Big] + (1-\gamma(c_S(\bar{I})))V_D\Bigg]\sigma_S(\bar{I})$$

This is because:

$$EU_{D_S}(\sigma_S) = \int \sigma_S(I)\Bigg[\sigma_A\Bigg[\gamma(c_S(I))\Big[\sigma_{R_\theta}\big[p(c_S(I))(V_R - C_C - C_R) + (1 - p(c_S(I)))(-C_D - C_R)\big] + (1 - \sigma_{R_\theta})(-C_D)\Big] + (1 - \gamma(c_S(I)))V_D\Bigg] + (1 - \sigma_A)V_D\Bigg] dI$$

$$EU_{D_S}(\sigma_S^{'}) = \int \sigma_S(I)^{'}\Bigg[\sigma_A\Bigg[\gamma(c_S(I))\Big[\sigma_{R_\theta}\big[p(c_S(I))(V_R - C_C - C_R) + (1 - p(c_S(I)))(-C_D - C_R)\big] + (1 - \sigma_{R_\theta})(-C_D)\Big] + (1 - \gamma(c_S(I)))V_D\Bigg] + (1 - \sigma_A)V_D\Bigg] dI$$

Let's assume

$$\Bigg[\sigma_A\Bigg[\gamma(c_S(I))\Big[\sigma_{R_\theta}\big[p(c_S(I))(V_R - C_C - C_R) + (1 - p(c_S(I)))(-C_D - C_R)\big] + (1 - \sigma_{R_\theta})(-C_D)\Big] + (1 - \gamma(c_S(I)))V_D\Bigg] + (1 - \sigma_A)V_D\Bigg] = m(I).$$

$\sigma_S = \sigma_S^{'}$, for all $I \neq \bar{I}, \hat{I}_S$.

$$EU_{D_S}(\sigma_S^{'}) - EU_{D_S}(\sigma_S) = \int \sigma_S(I)^{'} m(I) dI - \int \sigma_S(I) m(I) dI =$$

$$\int_{I \neq \bar{I}, \hat{I}_S} \sigma_S(I)^{'} m(I) dI + \int_{I \in \bar{I}, \hat{I}_S} \sigma_S(I)^{'} m(I) dI - \int_{I \neq \bar{I}, \hat{I}_S} \sigma_S(I) m(I) dI - \int_{I \in \bar{I}, \hat{I}_S} \sigma_S(I) m(I) dI =$$

$$\int_{I \in \bar{I}, \hat{I}_S} \sigma_S(I)^{'} m(I) dI - \int_{I \in \bar{I}, \hat{I}_S} \sigma_S(I) m(I) dI =$$

$$\sigma_S^{'}(\bar{I}) m(\bar{I}) + \sigma_S^{'}(\hat{I}_S) m(\hat{I}_S) - \sigma_S(\bar{I}) m(\bar{I}) - \sigma_S(\hat{I}_S) m(\hat{I}_S) =$$

$$0 + \big[\sigma_S(\bar{I}) + \sigma_S(\hat{I}_S)\big] m(\hat{I}_S) - \sigma_S(\bar{I}) m(\bar{I}) - \sigma_S(\hat{I}_S) m(\hat{I}_S) =$$

$$\sigma_S(\bar{I}) m(\hat{I}_S) + \sigma_S(\hat{I}_S) m(\hat{I}_S) - \sigma_S(\bar{I}) m(\bar{I}) - \sigma_S(\hat{I}_S) m(\hat{I}_S) = \sigma_S(\bar{I}) \big[m(\hat{I}_S - m(\bar{I}))\big].$$

Because $\sigma_A \in \{0, 1\}$,

$$EU_{D_S}(\sigma_S^{'}) - EU_{D_S}(\sigma_S) \geq \Bigg[\gamma(c_S(\hat{I}_S))\Big[\sigma_{R_S}^{'}\big[p(c_S(\hat{I}_S))(V_R - C_C - C_R) + p(c_S(\hat{I}_S))(-C_D - C_R)\big] + (1 - \sigma_{R_S}^{'})(-C_D)\Big] + (1 - \gamma(c_S(\hat{I}_S)))V_D\Bigg] - \Bigg[\gamma(c_S(\bar{I}))\Big[\sigma_{R_S}\big[p(c_S(\bar{I}))(V_R - C_C - C_R)\big] + (1 - p(c_S(\bar{I}))(-C_D - C_R)\Big] + (1 - \sigma_{R_S})(-C_D)\Big] + (1 - \gamma(c_S(\bar{I})))V_D\Bigg]\sigma_S(\bar{I})$$

Because $c_S(\hat{I}_S) > c_S(\bar{I})$, $EU_{D_S}(\sigma'_S) - EU_{D_S}(\sigma_S) > 0$. Hence, $\sigma'_S(I)$ is a profitable deviation.

The above logic applies to the case $\bar{I} \neq \hat{I}_W$ as well. But $\bar{I}$ cannot be $\hat{I}_S$ and $\hat{I}_W$ at the same time. Therefore, at least one of D_S and D_W has an incentive to deviate.

2. Suppose $P_{A(\bar{I})} > \bar{P}$, i.e., $\sigma_A(\bar{I}) = 0$. D_S and D_W get the first-best outcomes and have no incentive to deviate. As long as there exists some I such that $P_{A(\bar{I})} > \bar{P}$, I could have pooling equilibria. Now let's find a sufficient condition for this. Define $g(I) = p_S(I)q + p_W(I)(1-q)$, then $g(I) = p_S(I)q + p_W(I)(1-q) = p(c_S(I))q + p(c_W(I))(1-q) = p(I + (1 - I_S^{-1}(I))n)q + p(I + (1 - I_W^{-1}(I))n)(1-q)$. One sufficient condition is

$$\max_{I \in [0, I_W(1)]} g(I) > \bar{P} \tag{5}$$

If Inequality 5 is satisfied, then there exists a set $\mathcal{I}_g \equiv \{I \in [0, I_W(1)] : g(I) > \bar{P}\}$. Hence a representative pooling equilibrium is: $\sigma_S(I_1) = \sigma_W(I_1) = 1$, $\sigma_{R_\theta}(I_1) = 0$, $\sigma_A(I_1) = 0$, $\mu(I_1) = q$, $\mu(I) = 0$ for $I = I_1$, and $\sigma_A(I) \in [0,1]$, $\mu(I) \in [0,1]$ for $I \neq I_1$, where I_1 is a number in $\mathcal{I}_g$.

Case 2.2: $H \neq \emptyset$. Under this assumption, $\mu(I) = 1$ for any $I \in H$. Therefore, $P_{A(I)} = p_S(I)\mu(I) + p_W(I)(1-\mu(I))$ and $p_S(I) > \bar{P}$, for any $I \in H$. As long as D_S chooses CBI in H, she get her first best outcome and D_W cannot imitate D_S. Since C does not attack D_S, there is no need for D_S to consider retaliation against C ($\sigma_{R_S} = 0$). But what will D_W do in this situation? D_W will retaliate if her expected utility from retaliating is higher than her expected utility from not retaliting. Specifically, $EU_{D_W}(Retaliate) > EU_{D_W}(\neg Retaliate) \Rightarrow p(c_W(\hat{I}_W))(V_R - C_D - C_R) + (1 - p(c_W(\hat{I}_W))(-C_D - C_R) > -C_D$, meaning that D_W retaliates if $p(c_W(\hat{I}_W))V_R > C_R$, as demonstrated by Equation 1.

Hence there exists a separating equilibrium outcome. A representative equilibrium is: $\sigma_S(I_0) = 1, \sigma_W(\hat{I}_W) = 1, \sigma_{R_S}(I_0) = 0, \sigma_{R_W}(\hat{I}_W) = 1, \sigma_A(I_0) = 0, \sigma_A(I) = 1, \mu(I_0) = 1, \mu(\hat{I}_W) = 0 \,\text{for}\, I \in \{I_0, \hat{I}_W\}, \text{and}\, \sigma_A(I) \in [0,1], \mu(I) \in [0,1]$ for $I \notin \{I_0, \hat{I}_W\}$, where I_0 is any number in H. The existence of pooling equilibria shall follow Case 2.1 (part 2).

Case 3: $\hat{P}_W < \hat{P}_S < \bar{P}$**.** In this case, $\mathcal{I}_\theta(\bar{P}) = \emptyset$ for D_θ. Hence, for any CBI and for $\mu(I)$, $P_{A(I)} = p_S(I)\mu(I) + p_W(I)(1-\mu(I))$. If $\hat{P}_W < \hat{P}_S < \bar{P}$, $p_S(I)\mu(I) + p_W(I)(1-\mu(I)) < \bar{P}\mu(I) + (1-\bar{P})\mu(I) \Rightarrow P_{A(I)} < \bar{P}$, meaning that C attacks when observing any I. Then, D_θ's expected payoff is

$$\gamma_\theta \Big[\sigma_{R_\theta}\big[p(c_\theta)(V_R - C_P - C_R) + (1 - p(c_\theta))(-C_D - C_R)\big] + (1 - \sigma_{R_\theta})(-C_D)\Big] + (1 - \gamma_\theta)V_D$$

D_θ maximizes $p(c_\theta(I))$ to maximize her expected payoff. I have defined that $\hat{I}_\theta = \text{argmax}\, c_\theta(I)$, since $p(\cdot)$ is a increasing function, $\hat{I}_\theta = \text{argmax}\, p(c_\theta(I))$. As mentioned earlier, D_θ retaliates when $p(c_\theta(\hat{I}_\theta))V_R > C_R$. Therefore, for any equilibrium $\sigma_S(\hat{I}_S) = \sigma_W(\hat{I}_W) = 1$, $\sigma_{R_S}(\hat{I}_S) = \sigma_{R_W}(\hat{I}_W) = 1$, $\sigma_A(I) = 1$ for any I, $\mu(\hat{I}_S) = 1, \mu(\hat{I}_W) = 0$, and $\mu(I) \in [0,1]$ for any $I \notin \{\hat{I}_S, \hat{I}_W\}$.

Case 4: $\bar{P} = \hat{P}_W < \hat{P}_S$. If $\bar{P} = \hat{P}_W < \hat{P}_S$,

$$C : \begin{cases} \text{mixes} & I = \hat{I}_W, \\ \text{does not attack} & I = \hat{I}_S. \end{cases}$$

In this situation D_W has an incentive to deviate to some $I'_W = \hat{I}_S$ and not to be attacked by C.

Case 5: $\hat{P}_W < \bar{P} = \hat{P}_S$. If $\hat{P}_W < \bar{P} = \hat{P}_S$,

$$C : \begin{cases} \text{attacks} & I = \hat{I}_W, \\ \text{mixes} & I = \hat{I}_S. \end{cases}$$

Equation 4 shows that C is indifferent between attacking and not attacking when $\bar{P} = \frac{\sqrt{V_C^2 - 4\sigma_{R_\theta} C_P (C_C + R)}}{2\sigma_{R_\theta} C_P}$. Let's assume that C attacks when he observes $\hat{I}_S$ with probability α. In this case, D_S will not imitate D_W because C attacks D_W. Let's check if there is any profitable deviation for D_W. If D_W does not imitate D_S, she receives

$$EU_{D_W}(\hat{I}_W, \hat{I}_S) = -V_D + \hat{P}_W(V_R - C_D - C_R) < 0.$$

But, if D_W imitates D_S, then

$$c_W(\hat{I}_S) = \hat{I}_W(\hat{I}_W^{-1}(\hat{I}_S)) + (1 - \hat{I}_W^{-1}(\hat{I}_S))n = \hat{I}_S + (1 - \hat{I}_W^{-1}(\hat{I}_S))n.$$

This is because solving for the level of r, such that a weak type mimics a strong type, I get: $I_W(r) = \hat{I}_S \rightarrow r = \hat{I}_W^{-1}(\hat{I}_S)$. As a result, D_W imitates D_S, when

$$\bar{P}_2 = \mathcal{P}(c_W(\hat{I}_S)). \tag{6}$$

Then,

$$EU_{D_W}(I'_W = \hat{I}_S, \hat{I}_S) = (1-\alpha)(0) + \alpha[-V_D + \bar{P}_2(V_R - C_D - C_R)] = \alpha[-V_D + \bar{P}_2(V_R - C_D - C_R)].$$

Let's consider the following three cases that consider different levels of α.

Case 5.1.: $\alpha < \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)}$.

D_W's payoff from mimicking D_S is

$$\alpha[-V_D + \bar{P}_2(V_R - C_D - C_R)].$$

Her payoff from not mimicking D_S is

$$-V_D + \hat{P}_W(V_R - C_D - C_R).$$

Because

$$\alpha[-V_D + \bar{P}_2(V_R - C_D - C_R)] > -V_D + \hat{P}_W(V_R - C_D - C_R),$$

$$\alpha < \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)}.$$

There is no equilibrium here, because D_W has an incentive to deviate and imitate D_S.

Case 5.2.: $\alpha > \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)}$.

If $\alpha > \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)}$, then

$$C : \begin{cases} \text{attacks} & I = \hat{I}_W, \\ \text{mixes} & I = \hat{I}_S. \end{cases}$$

As a result,

$$D : \begin{cases} D_W, & I = \hat{I}_W, \\ D_S, & I = \hat{I}_S. \end{cases}$$

Here we have a *mixed strategy separating equilibrium*, where $\sigma_\theta = I_\theta, \sigma_{R_S}(\hat{I}_S) = 0, \sigma_{R_W}(\hat{I}_W) = 1, \sigma_A(I) = \alpha, \mu(I_S) = 1, \mu(I_W) = 0$ for $I \in \{\hat{I}_W, \hat{I}_S\}$, and $\sigma_A(I) \in [0,1], \mu(I) \in [0,1]$ for $I \notin \{\hat{I}_W, \hat{I}_S\}$. Let's solve for the optimal level of α when this equilibrium holds (C mixes when sees $\hat{I}_S$):

$$\alpha^* = \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)} =$$

$$\frac{V_D - \hat{P}_W(V_R - C_D - C_R) + \bar{P}_2(V_R - C_D - C_R) - \bar{P}_2(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)} =$$

$$1 + \frac{(\bar{P}_2 - P_W)(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)} \tag{7}$$

Equation 7 shows α^* decreases as P_W and V_D increase and α^* increases as $\bar{P}_2$ increases. As a result, for all α's that are above α^*, this equilibrium holds.

Case 5.3.: $\alpha = \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)}$.

If $\alpha = \frac{V_D - \hat{P}_W(V_R - C_D - C_R)}{V_D - \bar{P}_2(V_R - C_D - C_R)}$, there are two possible scenarios:

1. If D_W does not imitate D_S, then we have the same equilibrium as defined in *Case 5.2.*
2. If D_W imitates D_S, we do not have an equilibrium, as explained in *Case 5.1.*

■

2 Anecdotal Evidence from the Interviews on the Topic of Cyber Institutions as a Deterrent

I first examine the defender's behavior and then take a look at the challenger's choice of actions.

The fact that more than one hundred of the world's militaries are said to have some sort of organization or unit to address "cyber warfare" might be suggestive of two trends. First is these countries' deterrent intent. When a country starts developing its cyber offensive capability, the most natural fit is to place its cyber offensive operations within its signals intelligence (SIGINT) agencies because these agencies are the most equipped to deal with "cyber" (Interview, 2018: #3). Such a move signals that the country is in the process of utilizing the advantages of cyberspace

to primarily collect intelligence. If a country proceeds to the next step and starts creating cyber offensive capabilities within its military, it signals its capability and intent to use its cyber offensive tools to go beyond its national borders to punish cyber aggressors (Interview, 2018: #11). Contrary to cyber intelligence, which includes quiet penetration operations that aim to stay undetectable in adversarial networks as long as possible, this *loud* signal of readiness to attack is meant to deter potential perpetrators. A recent U.S. Department of Defense's (DoD) strategy of "active defense," defined as "the employment of limited offensive action and counterattacks to deny a contested area or position to the enemy" (*US DoD Active Defense*, 2019), is an example of such a cyber institution meant to deter adversaries.

Second is pooling behavior of these nations (Singer and Friedman, 2014). Without providing specific details regarding the capabilities of the created units, nations hope to hide their low cyber capabilities behind these cyber institutions and convince their adversaries that their cyber capabilities are real and growing. While there are many examples of this pooling behavior, let us take a quick look at Norway, which created its cyber command back in 2012 (Interview, 2018: #4, #15). Despite an eagerness to invest in its cyber offensive capabilities, Norway has not made much progress in the development of these capabilities. A desire to be "forward leaning," which one interviewee describes as a typical Norwegian feature, likely explains why the country was so quick to document its plan to create a cyber military unit (Interview 2018: #11), but it does not explain why the country has been slow in implementing this plan. Instead, the country's desire to pool with strong cyber nations is a more plausible explanation for this behavior.

Norway is not an exception in this regard. During the last few years, many countries have become eager to announce the creation of offensive cyber capabilities within their governments' militaries and the adjustment of their cyber doctrines to reflect this change. For instance, the pacifistic nature of the German and Japanese constitutions does not prevent these nations from developing cyber military units and from slowly shifting their cyber defensive postures to more offensive ones (Matsubara, 2018; Schulze and Herpig, 2018). Recently, French armed forces minister Florence Parly unveiled the country's first doctrine for offensive cyber operations (*Public Elements of the Doctrine on Military Cyber Offensive*, 2019), stating that France is "not afraid" of using cyber "weapons" in response to cyber threats (Laudrain, 2019).

Careful examination of these public actions and declarations may lead to the conclusion that the stated capability is not always present. For example, with plans to establish cyber military units, countries tend to report how many cyber soldiers these units will have in a three- or five-year period. While militaries often rely on contractors for the development of a computer code, these contractors are forbidden to execute actual operations on behalf of military. As a result, these newly recruited cyber soldiers should possess at least some basic computer skills to be able to execute cyber operations against enemies. When asked for concrete details about the recruitment and training of cyber soldiers, silence, vague responses, or the excuse that many countries were finding it difficult to recruit cyber warriors into their forces followed (Interview 2018: #11, #20, # 35, #49).

There are two possible explanations of this discrepancy between stated and actual capabilities. First, it can hint that countries use public cyber institutions, which are easy to observe but difficult to verify, as their main strategy of signaling their cyber capability and resolve in hopes that their adversaries overestimate their true cyber arsenals and intentions. Second, it can be the common maturation trend within militaries in which doctrine and organizational outpaces operations capabilities. China, for example, has an aspirational doctrine since 2001 but took a few years to build up operational capabilities to be able to implement that doctrine in 2005. This is when it make its

military cyber organizations/doctrine public.

While many nations create cyber institutions to pool with so-called *cyber powers*,[1] these powers, in their turn, implement a more significant level of cyber institutions to differentiate themselves from weaker cyber nations. For instance, Russia established information warfare units (Reuters, 2017) and has committed between $200 million and $250 million USD per year to significantly strengthen its cyber-offensive capabilities and to create a cyber-deterrent that "will equate to the role played by nuclear weapons" (Gerden, 2016). The U.S. DoD's 2017 cyber budget of $6.7 billion USD was devoted to "strengthening cyber defenses and increasing options available in case of a cyber-attack" (U.S.DepartmentOfDefense, 2016). Even though its exact number remains unknown, some of this budget was spent on the implementation the U.S. cyber strategy of active defense or on the establishment of other cyber institutions. If deterrence had worked and the country was ready and confident in its ability to defend itself in cyberspace, it would not need "to make [any additional] noise" (Interview, 2018: #20). Multiple efforts — more investment in cyber institutions, in this case — can simply signal "mostly failure" of the already implemented efforts (Interview, 2018: #20).

Now, let us examine *how such actions affect the challenger's decisions.* My interviewees were rather skeptical — in line with the model's results — of cyber institutions as an effective deterrent mechanism, citing the difficulty of demonstrating this effect empirically as a major challenge. Because cybersecurity is a relatively new area of national security, and most information about decision-making processes regarding this area remains classified, there is no publicly-available evidence suggesting that policy-makers, for example, have decided not to attack a country in cyberspace because they were afraid of that country's potential cyber response. We can assume that U.S. adversaries became more concerned about their networks after Snowden, Stuxnet, and Shadow Brokers, for instance, revealed the skill and organizational capacity of the National Security Agency (NSA). But, an official confirmation of this assumption is lacking.

My interviewees professed the belief that deterrence is working in the case of strategic cyber-attack scenarios – blackouts attacks, for example (Interview 2018: #48, 49). They point to two signs of successful deterrence. First, the decision to design cyber weapons with more care and precision shows that states have started practicing some restraint as more nations become cyber-dependent. Attackers attempt to limit unintended and unpredictable consequences of their cyber operations, hoping not to cross vaguely defined "red lines" of acceptable cyber behavior and trigger a response in the digital domain. Supporting this explanation, one of my interviewees pointed to the built-in restrictions in the WannaCry and NotPetya attacks as "evidence that government agencies might be restraining themselves" (Interview, 2018: #48). Without these restrictions, the consequences of these operations could have been more devastating (Vanderburg, 2017).

Second, cyber powers have changed their cyber strategies. With an increase in China's reliance on the Internet, resulting in an increase in China's own cyber vulnerability, the country has changed its cyber force posture from brinkmanship to calibrated escalation, signaling to its adversaries that it wants to avoid full-scale retaliation (Cunningham, 2018). Moreover, evidence from Estonia's intelligence reports demonstrate a change in Russia's cyber strategies. With an increase in the country's cyber capability, Estonia's newly established cyber institutions have fallen victim to low-level attacks, most likely coming from Russia in the form of phishing and spear-phishing emails that target the private emails of diplomats, politicians, and people involved in the military and national security (*Estonian Annual Review* 2009–*Estonian Annual Review* 2017).[2]

[1] China, Iran, Israel, North Korea, Russia, the United Kingdom, and the United States (Vavra, 2017).

[2] This change of strategy is most likely the result of NATO's extended deterrence (Gannon and Lindsay, 2017). I

As with most deterrence studies, my evidence suffers from two types of critiques (Lupovici, 2018). First is the stated versus actual intent to deter. Even though official documents state that cyber institutions are created to deter adversaries, it is impossible for one to discern whether actual cyber deterrence was even attempted. While this is a valid concern, the distinction between actual and stated intents is not as important as adversarial perception of this intent. This is connected to the second critique — the deterrent effect. Specifically, one might find it difficult to discern whether cyber institutions have an effect on an adversary's strategic calculations. The fact that adversaries tend to perceive such general assertions of capacity and intention as if they were directed specifically at them partially addresses these two critiques (Segal, 2014).

3 Election Examples

In the main manuscript, I present the case of the 2018 Swedish elections where cyber institutions deterred the Kremlin from interfering into this election (Section 3.2). Here, I focus on the two cases where cyber institutions had not effect on the Kremlin's decision of whether to attack. In the German 2017 elections, a combination of non-cyber factors shifted Russia's cost-benefit calculus in favor of not attacking even before cyber institutions were put in place (Section 3.1). In the U.S. 2016 elections, Moscow was willing to pay any cost and to risk any potential U.S. (cyber or non-) retaliation for its influence campaign that could help create a global authoritarian fraternity (Section 3.2).

3.1 2017 German Federal Elections: Why stop half-way through?

The 2017 German federal elections fall into the left region of Table 2 ($\bar{P} < P_W < P_S$) because *German cyber institutions had no effect on Russia's decision to stay away from the 2017 German federal elections.* Instead, a combination of non-cyber factors shifted Russia's cost-benefit calculus in favor of not attacking even before cyber institutions were put in place.

The history of cyber and information operations attributed to the Kremlin a few years prior to the elections demonstrate that the Russian state had interest in German elections interference. Information operations started in 2013, when three key German-language Kremlin-linked propaganda outlets — *Sputnik Deutsch*, *RT Deutsch*, and *NewsFront Deutsch* — entered the German market (Nimmo, 2017) and were later joined by trolls[3] and bots.[4] The 2015 and 2016 cyber operations against the parliament and political parties demonstrate that Moscow was also eager to obtain sensitive information for potential future use (Herpig, 2017). But value from these efforts for an effective influence campaign were overwhelmed by potential costs resulting from the following factors.

First, Germany's balanced media systems, the lack of polarization among the German public, Germany's multiparty and proportional system, and a "gentleman's agreement" between major political parties not to use any information leaked as a result of cyber attacks made it difficult for Moscow to sow confusion in the public (Schwirtz, 2017). Second, the only clear beneficiary of these campaigns was the Alliance for Germany (AfD) — the rest of parties supported the sanction regime against Moscow. But when AfD entered the race, they only had between eight and ten percent of

discuss this possibility and its implications in the main manuscript (Section 4).

[3] A troll is someone who argues for extreme views without credible sources.

[4] An automated program that runs over the Internet.

vote, which was not enough to gain the majority. Third, the use of old-fashioned paper ballots on the local level afforded Germany with an accurate recount in the event that digitized votes used on the federal level were compromised (Herpig, 2017). Lastly, high-level deterrent rhetoric by German politicians referencing a deterioration of the relationship between the two countries if interference occurred likely played a part.[5] Although Putin and Merkel's relationship has eroded over time, alienating Germany — an important bridge between the West and Russia — was not in Russia's interest.

If not these factors, what other factors could have stopped Russia from election interference? Brattberg and Maurer (2018) and Schwirtz (2017) suggest that a failure to influence the 2017 French presidential elections made the Kremlin re-think its approach, since it lost the element of surprise. While quite plausible, the evidence later revealed interference into the French elections was not directed by the Kremlin, although it was aligned with its objectives (Galante and EE, 2018). In preparation to the elections, Germany was also heavily investing in its cyber defense and offense, hoping to deter the aggressor. But, these institutions played no role in the Kremlin's decision to stay away from the German elections because Berlin's threat of punishment lacked credibility and its deterrence by prevention was failing due to poorly-constructed defenses.

To deter by punishment, the German government identified hybrid threats and cyberwarfare as key security concerns and re-iterated the central role of the German army in defending against such threats (*White Paper*, 2016). But, in practice, the German army could only protect its own systems (Brady, 2017).[6] Moreover, how to protect the 2017 election was a single agenda item, discussed at the Federal Security Council — a body that only meets when the country faces the most serious threats — in March 2017. The *hack-back* strategy was an option that the council considered. If implemented, this strategy would allow the German government to launch offensive cyber attacks against attackers before they could do any real damage (Schwirtz, 2017). For instance, in response to the attack against the parliament, the German government could remove servers where stolen parliament data was located. The implementation of the hack-back strategy, however, required two major changes in the German constitution that were unlikely to be adopted in time prior to the elections: (1) to move the responsibility for hack-backs from a state to federal level, and (2) to allow the military to respond to cyber operations below the threshold of armed conflict (Schulze, 2019; Schwirtz, 2017).

To deter by prevention, German's Federal Office for Information Security ran penetration tests to detect any vulnerability in its systems and networks and the parliament strengthened its computer security (Schwirtz, 2017). The *Cyber Security Strategy for Germany* (2016, 8) set up a National Cyber Response Center that reports to the Federal Office of Information Security (BSI), a domestic intelligence agency, and deals with response measures for various information technology (IT) incidents. In its turn, BSI created a mobile Quick Reaction Force and, for the first time ever, BSI briefed party campaigns about intrusion vectors and cyber-hygiene (Schulze, 2019). Other government-organized campaigns aimed at increasing the cost of a potential influence campaign included: monitoring the Internet for misinformation campaigns meant to sway the election, provisions

[5] During its spring 2017 visit to Moscow, the "Chancellery emissary delivered a stern warning"; in May, Merkel herself issues a warning to Putin, by saying "she assumes 'German parties will be able to decide their election campaign among themselves"'(Beuth et al., 2017); and in June, German President Frank-Walter Steinmeier warned Moscow that "Were [it] to interfere in the election of the Bundestag, then the share of commonalities will necessarily decrease further. That would be damaging for both sides" (as quoted in Brattberg and Maurer (2018, 18)).

[6] If asked by BSI, Bundeswehr can also help protect critical infrastructure in a crisis. But, cooperation between BSI and Bundeswehr on an operational level is much easier said than done (Schulze, 2019).

of government-subsidies to media and local publishers for the creation or expansion of fact-checking programs, and education programs of German politicians on basic cyber hygiene and digital threats by private companies, such as Facebook and Google (Auchard and Sterling, 2017; Scott, 2017).

But, deterrence by prevention was unlikely at play here for the following two reasons. First, while public outreach efforts are important, it is not clear how quickly educational campaigns lead to a change in behavior. Kostyuk and Wayne (2019) demonstrate that even though respondents who are exposed to a cyber attack to which they can personally relate report a desire to engage in safer online behavior, they fail to do so. Second, the means of these initiatives did not always translate into their ends. For instance, early in 2017, the Office of the Federal Returning Officer established a verified Twitter account to share any news about clarifications of potential fake news that could disrupt the electoral process (Brattberg and Maurer, 2018). But considering that there are only nineteen percent of Germans on Twitter (as of 2018),[7] the impact of this initiative remained unclear (Scott, 2017). This low number might explain why there was no major coordinated bot activity on Twitter around the election period (Rosenberger and Berger, 2017). These few examples demonstrate that the efforts of the German government to increase the cost of potential interference campaigns did not necessarily translate into cyber capability. As a result, they had no effect on Russia's decision to stay away from the 2017 German elections.

3.2 2016 U.S. Election: Full speed ahead

The 2016 U.S. presidential elections fall into the right region of Table 2 ($P_W < P_S < \bar{P}$) because *U.S. cyber institutions had no effect on the Kremlin's decision to interfere into the 2016 U.S. presidential elections*. Having witnessed the worldwide impact of the mass disclosures of the U.S. government's treatment of private data, Moscow was willing to pay any cost and to risk any potential U.S. (cyber or non-) retaliation for its influence campaign that could help create a global authoritarian fraternity. Having made up its mind before or in 2014, the Russian government was an unstoppable tank moving towards its target.

Moscow's online campaigns and the Russian intelligence-gathering mission that began in 2014 demonstrate the seriousness of Russia's intentions in implementing its democracy containment doctrine (USDepartmentOfJustice, 2018). Specifically, the Russian influence campaign took root back in 2014, when the Internet Research Agency (IRA) began operating a social media campaign, "designed to provoke and amplify political and social discord in the United States" (Mueller, 2019, 4). With time, this campaign evolved into "a targeted operation that...favored candidate Trump and disparaged candidate Clinton" (Mueller, 2019, 4). In addition to these overt online operations, the Russian government also used covert cyber attacks to achieve its goal. For example, in July 2015, Russia's General Staff Main Intelligence Directorate (GRU) gained access to the Democratic National Committee (DNC) networks; in March 2016, they began cyber operations aimed to compromise email accounts of Democratic Party officials; and in June 2016, they released content of the stolen data using WikiLeaks and DCLeaks.com (*Assessing Russian activities and intentions in recent US elections*, 2017, 2).

When planning this influence campaign, Moscow likely contemplated between its value and cost. Its lowest value was "undermin[ing] public faith in the U.S. democratic process" and its highest value was "harm[ing Secretary Clinton's] electability and potential presidency" that could have resulted

[7] For information on user statistics, see: https://www.statista.com/statistics/867539/top-active-social-media-platforms-in-germany/

in the change in the U.S. foreign policy (*Assessing Russian activities and intentions in recent US elections*, 2017, ii). Even its lowest value was far greater than any costs Russia could envision paying.

If caught, Moscow knew that it faced potential retaliation by the world's military and cyber power. This retaliation, aligned with U.S. foreign policy tools, could have taken one of the following forms: diplomacy, economy, nuclear, and/or military (cyber and non-). Diplomatic retaliation was the least of Russia's worries considering the existing tension between the two countries. Moscow expected that, if elected, Clinton would only exacerbate this tension. The cost of additional economic sanctions — in addition to those that the country already faced for the Ukrainian conflict — was marginal. Moreover, nuclear and military responses to cyber and information operations were off the table.

At the time when the Russian government was about to start its influence campaign, it was not clear whether the U.S. government had the political will to retaliate against the Kremlin using cyber means. If the Kremlin was accused of interference, it could simply cite the difficulty of attributing the origin of cyber operations to deny their involvement or to blame patriotic hackers for executing these operations, as it has done in the past (Rid, 2013). In this case, a U.S. cyber response against the Kremlin might not have been justifiable. Moreover, Washington might have been hesitant to retaliate because of the high U.S. Internet connectivity that created a vast cyber attack surface providing plenty of targets if Russia chose to respond. Lastly, U.S. cyber defenses were not strong enough to deter Russia by prevention because, in 2014, the U.S. government was working on protecting its own network and critical infrastructure objects from cyber operations, and had yet to realize the danger of information campaigns.

This evidence demonstrates that there was a significant gap between the value and cost of the Kremlin's influence campaign at its start in 2014. In the following two years, the Kremlin likely felt that military and economic options were unlikely to add any additional costs because Washington was unlikely to change its view on these foreign policy tools. Because Moscow was aware that Washington was building its cyber institutions to increase its defenses and improve its offense, the Kremlin must have contemplated the additional cyber costs that it would incur during the influence campaign.

There were a few possible sources of additional costs. The first source was better defenses from cyber and information operations. Washington's cyber institutions implemented prior to 2014, meant to deter by prevention, could have given the Kremlin an idea of Washington's best possible defense. *Presidential Policy Directive 21* (2013), for instance, which focused on critical infrastructure protection might have sent two signals. First, because it did not cover voting machines as part of critical infrastructure, Moscow might have interpreted this as a signal that the government was not paying attention to election infrastructure protection. Second, because critical infrastructure protection was a rather new direction in U.S. cyber policy, Russia might have assumed that Washington, with its vast bureaucracy, would continue working in this direction over the next few years. Because a swift change in U.S. cyber policy was quite unlikely, the Kremlin was, to some extent, confident that Washington was not going to spend significant resources on educating political campaign leaders and the public about the danger of cyber threats and disinformation operations. But even if it did, the short-term impact of these efforts were quite uncertain, only slightly raising the already relatively low costs of the Kremlin's information operations.

The second source is cyber retaliation in the form of information operations. Both options were rather costly and ineffective and would have resulted in relatively minimal costs for the Kremlin.

Washington was unlikely to attempt information operations because of Kremlin's tight control of Russia's print, online, and social media and because of Russia's treatment of information as a weapon, allowing its military to respond to such information threats (*Conceptual Views Regarding the Activities of the Armed Forces of the Russian Federation in the Information Space*, 2011).

The third source is the targeting of Russia's critical infrastructure. Even if Washington spent the necessary time and resources for such an attack, the damage that Russia experienced would have been limited. For example, it is impossible to hack Russia's entire power grid system at once because most stations are manually controlled and can be restored by flipping a switch. Moreover, the Kremlin periodically tests the "cyber robustness" of all potential targets and does not use U.S. equipment to avoid the risk of Washington's remote-access backdoor (Ryabikova, 2019).[8] In all these hypothetical scenarios, the main question remains: was Washington willing even to consider any of these options for cyber retaliation, given the U.S. high Internet connectivity?

In short, the potential worst-case scenario costs that Moscow faced were lower than the value it would gain by election interference. Thus, the Kremlin's interference campaign was never going to be deterred by U.S. cyber institutions.

[8] A *backdoor* is an undocumented portal that allows an attacker to enter the system.

References

Assessing Russian activities and intentions in recent US elections. 2017. *Unclassified Version* .

Auchard, Eric and Tpby Sterling. 2017. "Google and sister company to offer cyber security to election groups.".

Beuth, Patrick, Kai Biermann, Martin Klingst and Holger Stark. 2017. "Merkel and the Fancy Bear." *ZEIT Online* .

Brady, Kate. 2017. "German government plans cyberattack 'hackback' ahead of election.".

Brattberg, Erik and Tim Maurer. 2018. *Russian Election Interference: Europe's Counter to Fake News and Cyber Attacks*. Vol. 23 Carnegie Endowment for International Peace.

Conceptual Views Regarding the Activities of the Armed Forces of the Russian Federation in the Information Space. 2011. Ministry of Defence of the Russian Federation, Moscow.

Cunningham, Fiona. 2018. "Maximizing Leverage: Explaining ChinaâĂŹs Cyber Force Posture.".

Cyber Security Strategy for Germany. 2016.

Estonian Annual Review. 2009.

Estonian Annual Review. 2017.

Fudenberg, Drew and Jean Tirole. 1991. "Game theory, 1991." *Cambridge, Massachusetts* 393(12):80.

Galante, Laura and Shaun EE. 2018. *Defining Russian Election Interference: An analysis of select 2014 to 2018 cyber enabled incidents*. Atlantic Council.

Gannon, Andres, Gartzke Erik and Jon Lindsay. 2017. "After Deterrence: Explaining Conflict Short of War.".

Gerden, Eugene. 2016. "Russia to spend 250 million US dollars strengthening cyber-offensive capabilities.".

Herpig, Sven. 2017. "Cyber Operations: Defending Political IT-Infrastructures.".

Interview. 2018. "Interviews on Cyber Institutions as a Deterrent.".

Kostyuk, Nadiya and Carly Wayne. 2019. "Communicating Cybersecurity: Citizen Risk Perception of Cyber Threats.".

Laudrain, Arthur P.B. 2019. "FranceâĂŹs New Offensive Cyber Doctrine." *Lawfare* .

Lupovici, Amir. 2018. "Toward a Securitization Theory of Deterrence." *International Studies Quarterly* .

Matsubara, Mihoko. 2018. "How JapanâĂŹs Pacifist Constitution Shapes Its Approach to Cyberspace." *Council on Foreign Relations* .

Mueller, RS. 2019. "Report on the investigation into Russian interference in the 2016 presidential election." *US Department of Justice* .

Nimmo, Ben. 2017. "The KremlinâĂŹs Amplifiers in Germany." *Digital Forensic Research Lab* .

Presidential Policy Directive 21. 2013.

Public Elements of the Doctrine on Military Cyber Offensive. 2019.

Reuters. 2017. "Russia sets up information warfare units - defence minister." *Reuters* .

Rid, Thomas. 2013. *Cyber war will not take place*. Oxford University Press.

Rosenberger, L and JM Berger. 2017. "Hamilton 68: A New Tool to Track Russian Disinformation on Twitter.".

Ryabikova, Victoria. 2019. "How Russia protects critical infrastructure from cyber attacks." *Russia Beyond* .

Schulze, Matthias. 2019. "2017 German Elections.". Email interview.

Schulze, Matthias and Sven Herpig. 2018. "Germany Develops Offensive Cyber Capabilities Without A Coherent Strategy of What to Do With Them." *Council on Foreign Relations* .

Schwirtz, Michael. 2017. "German Election Mystery: Why No Russian Meddling?" *New York Times* .

Scott, Mark. 2017. "Ahead of election, Germany seeks fake news antidote.".

Segal, Adam. 2014. "What Briefing Chinese Officials On Cyber Really Accomplishes." *Forbes* .

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

U.S.DepartmentOfDefense. 2016. "Consolidated DoD FY17 Budget Fact Sheet." https://dod.defense.gov/Portals/1/features/2016/0216_budget/docs/2-4-16_Consolidated_DoD_FY17_Budget_Fact_Sheet.pdf. Accessed: 2019-07-13.

USDepartmentOfJustice. 2018. "United States of America v. Internet Research Agency LLC et al.".

US DoD Active Defense. 2019. https://www.militaryfactory.com/dictionary/military-terms-defined.asp?term_id=37. Accessed: 2019-07-13.

Vanderburg, Eric. 2017. "Ransomware developers learn from the mistakes of WannaCry, NotPetya." *Carbonite* .

Vavra, Shannon. 2017. "The world's top cyber powers." *Axios* .

White Paper. 2016.