Cyber Deterrence by Punishment:
Role of Different Perceptions

Nasser S. AlAzwani and Thomas M. Chen
City, University of London
Northampton Square
London EC1V 0HB
United Kingdom
Email: nasser.al-azwani@city.ac.uk, tom.chen.1@city.ac.uk

Abstract:
Nuclear deterrence based on mutual assured destruction seems to have successfully prevented a global nuclear war for decades. Can deterrence be effective for cyber attacks between nation-states? The cyber environment is drastically different from the nuclear case. A major difference is the possibility of different perceptions by the states which may lead to a failure of cyber deterrence. In this paper, we compare differences between nuclear deterrence and cyber deterrence. We adapt a game theoretic model from the nuclear case to the cyber environment and show that differences in perceived payoffs can lead to attack strategies where deterrence fails in cyberspace.

- **Keywords:** Cyber security; deterrence theory; cyber deterrence; game theory; cyber defense.

1 Introduction

States around the world have become more dependent on technology and integrated systems [1]. Recognizing potential vulnerabilities in critical infrastructures, cyber security is now a top national priority for many states. Undoubtedly, ICT (information and communication technologies) brings societies and multinationals closer culturally [2] but has introduced serious challenges at the same time. Technologically advanced states are more at risk of enemies exploiting their vulnerabilities to gain unauthorized access to network resources or to cause harm to systems or people.

It is well known that security was not a high priority in the original design of the Internet. Since the Internet was opened to public services, cyber crime and cyber attacks have become commonplace [3]. Cyber attacks now threaten national security, and policy makers are challenged with dealing with threats from enemy state actors.

Cyber attacks are possible because of vulnerabilities in critical infrastructure [4, 5]. Control systems are increasingly connected to the Internet which allows adversaries from anywhere to carry out reconnaissance and remotely scan for vulnerabilities. An example incident was the malware attack on Ukraine's electrical grid that brought down approximately 75 percent of its electricity service [6]. Unfortunately, the complexity of modern systems and networks make them difficult to manage in terms of identifying and mitigating vulnerabilities [7].

Detecting and responding to cyber attacks can be extremely costly. Actively responding to attacks with offensive counter attacks has been proposed but the problem is the difficulty of attributing cyber attacks to the real attacker [8]. Active responses also incur a risk of escalating conflicts to more serious levels (e.g., to military confrontations).

Clearly, it is preferable to deter cyber attacks in the first place. It is mutually beneficial for all states to maintain a peaceful and cooperative cyberspace. However, it is not entirely clear how cyber deterrence can work. Although deterrence theory is well understood for nuclear weapons, the cyber environment is much different.

2 Traditional Deterrence

There is an extensive literature on the effectiveness of deterrence strategy as practiced in international relations [9]. During the Cold War, the threat of global nuclear war was a widesspread concern. The main idea behind nuclear deterrence is that any attack by one state on another state would be met with a devastating retaliatory response. This response will minimize any expected gain to the first attacker. Mutual assured destruction was the basis of many national policies for national security [10].

History has provided examples of different international cases where deterrence was successful [11]. Moreover, deterrence was considered the main reason for the prevention of another nuclear world war as well as prevention of chemical or biological attacks [12].

Deterrence theory is based on the assumption that people will make rational choices. It has been used as the basis for social policies, for instance, to discourage people from committing a broad range of crimes [13]. One of the best references that explains the development of general deterrence theory and its applications is Steff [14].

For successful deterrence, there is a need for defensive capabilities for observability, attribution, and readiness for retaliation [15]. Specifically, successful deterrence is predicated on three premises: (1) the deterrent should have a sufficient capability (2) the deterrent threat should be credible and (3) the deterrent threat should be communicated clearly to the adversary [16].

3 Cyber Deterrence

With the end of the Cold War and the escalation of cyber attacks between states, a natural question is whether deterrence strategies used to prevent nuclear war can be equally effective to prevent cyber attacks on critical infrastructure (possibly so-called "cyber warfare") [17, 18]. The concepts of deterrence may be similar for cyber space, but in terms of practice, major differences between the natures of cyber and nuclear domains should be recognized.

Cyber deterrence is a proactive strategy rather than a reactive defensive strategy. There is a difference between cyber defense and deterrence. Defense happens after an attack has been initiated in order to mitigate damage from the attack or win the conflict. In contrast, deterrence aims to prevent the conflict altogether and maintain peace within the cyber space. Clearly, cyber deterrence should be preferred over defense, just as in healthcare, prevention of disease in the first place is better than curing disease after it happens. By implication then, cyber deterrence should have high priority from the perspective of national security policies.

As mentioned earlier, successful deterrence depends on three essential pillars [19]:

- A credible defense meaning that the defender will be able to force the attacker to give up ultimately (the gain for the attacker will be less than the loss).
- Readiness to retaliate in the event of an attack.
- Willingness to retaliate against the attacker.

The challenges of cyber deterrence raised in the literature such as attribution, retaliations, and escalation [20] can be addressed in the context of these three pillars.

Cyber deterrence depends on a strong defense. Technologies for cyber defense have made great advances for attack detection, mitigation and recovery, but technology offers limited defense [21]. For example, Stuxnet was able to compromise an Iranian nuclear power plant despite Iranian precautionary controls [22]. Another example, malware hit the western part of Ukraine bringing down the electricity for more than six hours on December 23, 2015 [23].

In traditional deterrence, there are two types of deterrence strategy: deterrence by denial and deterrence by punishment. These two strategies also rely on credibility, capability, and communication with opponent [24, 25]. Both strategies are discussed below in the cyber space context.

3.1 Cyber deterrence by denial

The objective of deterrence by denial in cyber space is to develop a strong cyber defense that will make it very difficult for cyber attacks to succeed. Typical defenses (so-called defense in depth) consist of multiple layers including firewalls, intrusion detection systems, unified threat management, and encryption. The human element includes cyber security training and raising awareness of best practices.

In terms of the three factors mentioned earlier:

- Capability: states need to harden and strengthen their systems, particularly in terms of testing for vulnerabilities and patching.
- Communication: national and international cooperation may lead to agreement on norms or treaties between states.
- Credibility: investments in defenses must be convincing to potential attackers.

3.2 Cyber deterrence by punishment

Cyber deterrence by punishment is an alternative to deterrence by denial. In this strategy, a defending state threatens retalation against any attacking enemy state. The retaliation should be perceived by the attacker to inflict more cost than the perceived gains. In order for retaliation to work, it must be possible to attribute the attacker. This is straightforward in the case of nuclear weapons but not that easy in the cyber domain where cyber attacks may be stealthy.

- Capability: states must be able to attribute cyber attacks, presumably easier with robust military or law enforcement agencies. International cooperation is often required for attribution, but the lack of an international framework for cooperation is a major challenge.
- Communication: states need to clearly advertize their readiness to retaliate against any attacker.
- Credibility: the threat of retaliation must be convincing to potential attackers.

4 Problem and Approach

Nuclear deterrence by punishment is straightforward in terms of capability, credibility, and communication. All states are aware of each other's capabilities and the consequences of attacking each other. Moreover, all states are aware of the readiness and willingness of other states to retaliate.

In contrast, the cyber domain involves more uncertainties. A cyber attack may cause damage perceived differently by the attacker and defender. A critical target chosen by the attacker may actually be less valuable to the defender. Depending on the perceived damage, retaliation may or may not happen. In this paper, we investigate the consequences of different perceptions about the value of assets.

Game theory has proven to be a useful tool for analyzing strategic and competitive situations like deterrence. A game models the possible actions within a conflict and helps the players understand their best choice of action. Different types of games have been studied, depending on deterministic or random, complete or incomplete information, precommitments, signaling or no signaling, cooperation or no cooperation, and so on [26].

Our approach is to start with the traditional deterrence game (developed for nuclear deterrence) and then adapt the game model to the cyber domain [27]. There is a large literature discussing cyber deterrence but little of it uses game theory. Analysis of the game model will help to understand conditions leading to success or failure of cyber deterrence.

5 Game model for traditional deterrence

Figure 1 shows a simple two-player deterrence game in extensive form representing the traditional (nuclear) conflict. As usual, players are assumed to be rational and always seeking the best strategy to maximize their payoff. In this game, the players represent two nation-states in possible conflict. State A is the challenger threatening to attack State B, while B wants to deter the attack.

Since A is the challenger, A has the choice of first move which can to attack B or not. If A does not attack, B can attack pre-emptively or maintain the status quo. If A attacks, B can retaliate or not respond. It makes no sense for B to do nothing if A attacks. For deterrence to be effective, B should pre-commit to retaliate immediately if A attacks; thus, A is certain about mutual destruction if A attacks.

The payoffs for A and B, respectively, are noted as (A_i, B_j). It is assumed that payoffs are ordered: $A_1 < A_2 < A_3 < A_4$ and $B_1 < B_2 < B_3 < B_4$. The payoff for status quo should be no change, so $A_3 = B_3 = 0$. This game has three possible strategic scenarios as shown in Fig. 1:

1. status quo (no attack by either A or B);
2. A loses and B wins (no attack by A followed by a preemptive attack by B);
3. mutual destruction (A attacks followed by retaliation by B).

The fourth outcome (A attacks and B does not respond) is not possible because B pre-commits to retaliate if A attacks. If A attacks

Hypothesis: *State B can deter State A by a threat of sufficient and certain retaliation (deterrence by punishment).*

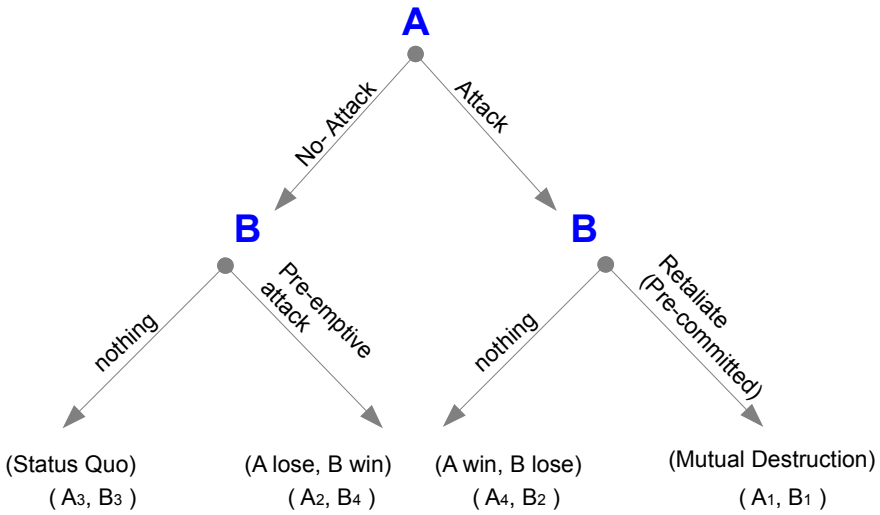

Figure 1: Deterrence game [28].

Games in extensive form are usually solved by backward induction. First, consider the subgame where A does not attack B. In the subgame, B has a choice to do nothing or pre-emptively attack A. The payoff for pre-emptive attack is higher, so B should choose to attack A. In this subgame, A should expect a loss of A_2 (a negative payoff).

Next, consider the other subgame where A chooses to attack B. B is pre-committed to retaliate, and the result is mutual destruction. The payoff to A is A_1 (a negative payoff).

Working backwards, A has the choice to attack with payoff A_1 or not attack with payoff A_2. Since it is assumed $A_1 < A_2$, it is better for A to not attack. Thus, A is effectively deterred from attacking B because of the threat of punishment.

It might seem in this case that B will win and A will lose, because B will pre-emptively attack A with payoff B_4. However, it should be pointed out that the game is symmetric. In other words, A will retaliate against B if B chooses to attack, so B is equally deterred from attacking A. Since A and B are mutually deterred from attacking each other, the status quo is maintained [29].

Why should B pre-commit to retaliate against A? Suppose that retaliation against A is not certain as shown in Fig. 2. If A attacks B, then B will not respond with probability P or retaliate with probability $1 - P$.

Again consider the subgames and work backwards. The first subgame is not changed. However, the subgame where A attacks has different expected payoffs now. The expected payoff for A is

$$E(A's\ payoff) = PA_4 + (1 - P)A_1 \tag{1}$$

Working backwards, A has the choice to attack with payoff $PA_4 + (1-P)A_1$ or not attack with payoff A_2. The incentive to attack is greater if

$$PA_4 + (1 - P)A_1 > A_2 \tag{2}$$

or the probability that B will not retaliate is

$$P > \frac{A_2 - A_1}{A_4 - A_1} \tag{3}$$

In this case, A may be tempted to attack B if A believes that there is a sufficient chance of "getting away with it" (i.e., B will not respond). For effective deterrence then, it is important for B to: (1) establish credibility for retaliation with nuclear capabilities and (2) communicate willingness, readiness, and pre-commitment to retaliate.

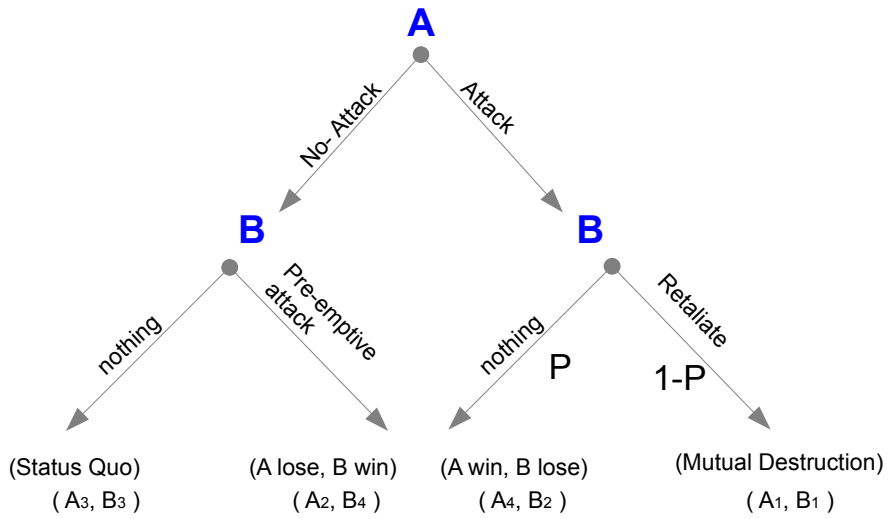

Figure 2: Deterrence game with uncertain retaliation.

6 Game model for cyber deterrence

The previous section described how states are deterred from attacking each other because of mutual assured deterrence. However, the situation is clearly different for cyber attacks. It is well known that states conduct ongoing cyber campaigns against each other. Why does deterrence work for nuclear war but not for the cyber domain?

The targets and damages from cyber attacks are different from nuclear attacks. Nuclear attacks are obviously devastating, cyber attacks are much more varied. Some cyber attacks are aimed at data theft while others more serious attacks are aimed at critical infrastructures [30]. Thus, the payoffs (gains and losses) in a cyber deterrence game model are more difficult to ascertain [31].

Critical infrastructure typically encompasses energy, telecommunications, financial services, water, and transportation, but there is not a universal agreement. These can span both public and private sectors. In the U.S., the definition of critical infrastructure has been expanded to include systems and assets, whether physical or virtual, so vital to the nation that the incapacity or destruction of such systems and assets would have a debilitating impact on securing, national economic security, national public health or safety [32]. They have different levels of importance in economic, social and military terms [33].

The problem with cyber deterrence by punishment arises because state A does not fear retaliation from state B. This might happen because:

- A does not its loss from retaliation as much as B perceives the loss to be;
- B may not retaliate because the loss from A's attack is not serious enough to merit retaliation.

Both might happen when the two states have different perceptions about payoffs in the game.

In theory, cyber deterrence should be based on the same game model shown in Fig. 1. State A is certain of a loss from attacking B if B pre-commits to retaliation. The expected loss from attacking B should be greater than the expected loss from doing nothing. By symmetry, both states see the status quo as the rational strategy.

In the previous section, it was established that state A will be deterred if $A_1 < A_2$, which was assumed. However, this may not be A's actual perception of the payoffs if the specific target of retaliation by B was unintentionally chosen to be less valuable to A than believed. What if A perceives that the loss from mutual destruction, A_1, is actually less (i.e., more positive) than the possible loss from doing nothing, A_2? Then A's best strategy would be to attack B and risk mutual destruction.

Another possibility is that B may not choose to retaliate if A attacks. A pre-commitment to retaliation was assumed for mutual assured destruction. However, let us reconsider the second subgame in Fig. 1. If state A attacks, B has a choice to retaliate or do nothing. Retaliation incurs a loss of B_1 while doing nothing will be a loss of B_2. It was assumed that $B_1 < B_2$, that is, B's actual

best strategy is to do nothing if A attacks. However, this strategy was ruled out because it could encourage A to attack; a pre-commitment to retaliate is a prerequisite for deterrence. However, if B_2 is a small loss that B can tolerate, then B may actually choose to do nothing. In any case however, B should "signal" (communicate clearly) its pre-commitment to retaliate in response to an attack by A, in order to deter A, even if B does not actually follow through on retaliation.

7 Deterrence Strategy: Target Selection)

In the previous section, we discussed that the payoffs in the cyber deterrence game can affected by the choice of targets. That is, the value of targets may not be perceived as expected by the other state. This can change the strategic choices of states leading to failure of deterrence.

Fig. 3 is a visualization of the value calculation that both states A and B carry out to decide on their best strategies. If A chooses to attack, it must weigh the perceived value of the target, and whether damage to the target will elicit a retaliation from B. In turn, B should weigh the perceived value of the target to retaliate against. For successful cyber deterrence, this target should be valuable enough to discourage A from attacking in the first place.

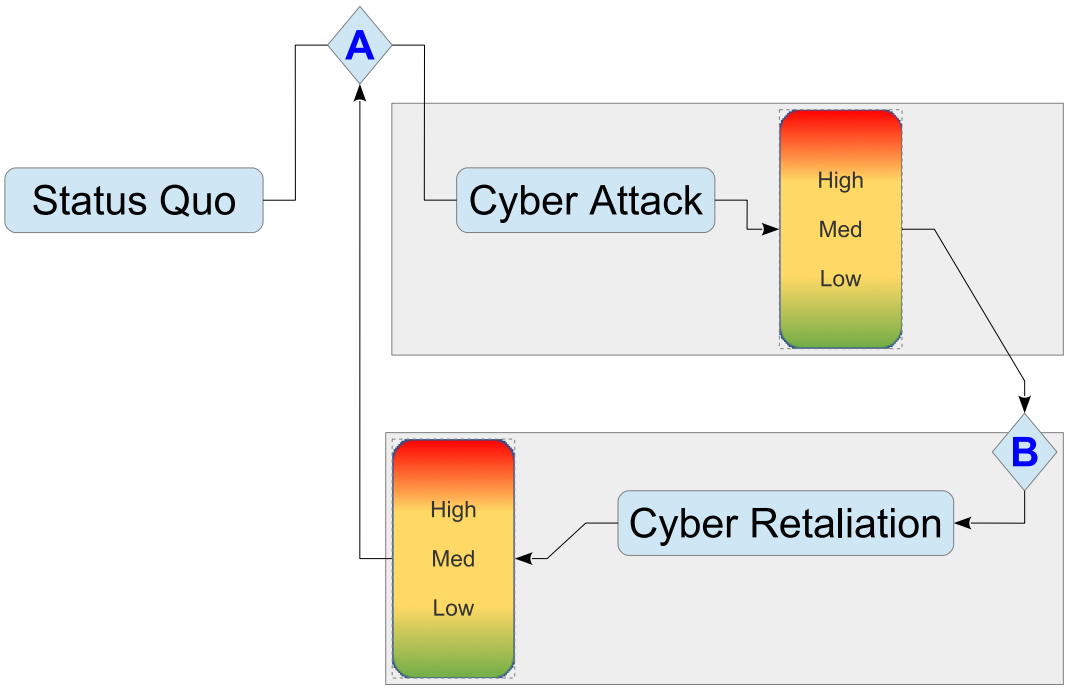

Figure 3: Attack and retaliation perceptions.

As the previous section established, misunderstandings of the game can arise from different perceptions of target values. For successful deterrence, it is important to minimize the differences in perceptions [34]. It is also important, as in nuclear deterrence, for each state to signal its pre-commitment to retaliate to a valued target of its enemy, even if the retaliation is not actually carried through. Deterrence by punishment can work only if the enemy fears a serious retaliation.

8 Conclusion

This paper has analyzed the traditional deterrence game model and has attempted to explain the role of threat of retaliation strategy in the success of nuclear deterrence. So far nuclear deterrence has seemed to work to avoid global nuclear war, but cyber deterrence has not worked. We have applied the deterrence game model to the cyber domain to explore reasons for the failure of cyber deterrence. One of the reasons may be different perceptions of target values, i.e., payoffs in the game model. Target selection plays a vital role in affecting the states' perceptions of the payoffs and ultimately their best strategic choices. It is important that targets are chosen suitably, and pre-commitment to retaliation is signaled clearly between states, in order to minimize the difference in perceived payoffs.

If both states understand the game clearly, then the principle of deterrence by punishment should work for cyber deterrence.

References

[1] Rinaldi, Steven M., James P. Peerenboom, and Terrence K. Kelly, 2001. Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE Control Systems 21, no. 6, pp. 11-25.

[2] Betz, D.J., 2017. Cyberspace and the State: Towards a Strategy for Cyber-power. Routledge.

[3] Carter, W.A., Sofio, D.G. and Alperen, M.J., 2017. Cybersecurity Legislation and Critical Infrastructure Vulnerabilities. Foundations of Homeland Security: Law and Policy, pp. 233-249.

[4] Hughes, J. and Cybenko, G., 2014, June. Three tenets for secure cyber-physical system design and assessment. In Cyber Sensing 2014 (Vol. 9097, p. 90970A). International Society for Optics and Photonics.

[5] Stoneburner, G., Goguen, A. and Feringa, A., 2013. Risk management guide for information technology systems. NIST.

[6] Sullivan, J.E. and Kamensky, D., 2017. How cyber-attacks in Ukraine show the vulnerability of the US power grid. The Electricity Journal, 30(3), pp. 30-35.

[7] Foreman, P., 2009. Vulnerability Management. CRC Press.

[8] Rid, T. and Buchanan, B., 2015. Attributing cyber attacks. Journal of Strategic Studies, 38(1-2), pp. 4-37.

[9] Langlois, J.P.P., 1989. Modeling deterrence and international crises. Journal of conflict resolution, 33(1), pp. 67-83.

[10] Morgan, P.M., 1983. Deterrence: A conceptual analysis (Vol. 40). Sage Publications.

[11] Huth, P. and Russett, B., 1984. What makes deterrence work? Cases from 1900 to 1980. World Politics, 36(4), pp. 496-526.

[12] Powell, R., 1990. Nuclear deterrence theory: The search for credibility. Cambridge University Press.

[13] Morgan, P.M., 2003. Deterrence now (Vol. 89). Cambridge University Press.

[14] Steff, R., 2016. Strategic Thinking, Deterrence and the US Ballistic Missile Defense Project: From Truman to Obama. Routledge.

[15] Schelling, T.C., 2008. Arms and influence: With a new preface and afterworld. Yale University Press.

[16] Paul, T.V., Morgan, P.M. and Wirtz, J.J. eds., 2009. Complex deterrence: Strategy in the global age. University of Chicago Press.

[17] Elliott, D., 2011. Deterring strategic cyberattack. IEEE Security & Privacy, 9(5), pp. 36-40.

[18] Kugler, Richard L., 2009. Deterrence of cyber attacks. Cyberpower and national security 320.

[19] Quackenbush, Stephen L., 2011. Understanding general deterrence. In Understanding General Deterrence, pp. 1-20. Palgrave Macmillan.

[20] Wei, Maj Lee Hsiang, 2015. The Challenges of Cyber Deterrence. Journal of the Singapore Armed Forces 41, no. 1.

[21] Multari, N.J., Singhal, A. and Miller, E., 2017, October. SafeConfig'17: Applying the Scientific Method to Active Cyber Defense Research. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 2641-2642). ACM.

[22] Farwell, J.P. and Rohozinski, R., 2011. Stuxnet and the future of cyber war. Survival, 53(1), pp. 23-40.

[23] Zetter, K., 2016. Inside the cunning, unprecedented hack of Ukraine's power grid. Wired.

[24] Bendiek, Annegret, and Tobias Metzger, 2015. Deterrence theory in the cyber-century. INFORMATIK 2015.

[25] Lowther, A. ed., 2012. Deterrence: rising powers, rogue regimes, and terrorism in the twenty-first century. Springer.

[26] Wang, Yuan, Yongjun Wang, Jing Liu, Zhijian Huang, and Peidai Xie, 2016. A Survey of Game Theoretic Methods for Cyber Security. In IEEE International Conference on Data Science in Cyberspace (DSC), pp. 631-636.

[27] Do, Cuong T., Nguyen H. Tran, Choongseon Hong, Charles A. Kamhoua, Kevin A. Kwiat, Erik Blasch, Shaolei Ren, Niki Pissinou, and Sundaraja Sitharama Iyengar, 2017. Game Theory for Cyber Security and Privacy. ACM Computing Surveys (CSUR) 50, no. 2: 30.

[28] Brams, S.J. and Bramj, S.J., 1985. Superpower games: Applying game theory to superpower conflict (p. 1985). New Haven: Yale University Press.

[29] Cimbala, S.J., 1998. The past and future of nuclear deterrence. Greenwood Publishing Group.

[30] Shackelford, S.J., Sulmeyer, M., Deckard, A.N.C., Buchanan, B. and Micic, B., 2017. From Russia with Love: Understanding the Russian Cyber Threat to US Critical Infrastructure and What to Do about It. Neb. L. Rev., 96, p. 320.

[31] Philbin, M.J., 2013. Cyber deterrence: An old concept in a new domain. Army War College, Carlisle Barracks PA.

[32] Moteff, J., Copeland, C. and Fischer, J., 2003, January. Critical infrastructures: What makes an infrastructure critical?. Library of Congress Washington DC Congressional Research Service.

[33] Moteff, J. and Parfomak, P., 2004, October. Critical infrastructure and key assets: definition and identification. Library of Congress Washington DC Congressional Research Service.

[34] Libicki, Martin C. Cyberdeterrence and cyberwar. Rand Corporation, 2009.