Strategic aspects of cyberattack, attribution, and blame

Benjamin Edwards[a,1], Alexander Furnas[b], Stephanie Forrest[c,d], and Robert Axelrod[e,1]

[a]Information Security Group, IBM Research, Yorktown Heights, NY 10598; [b]Department of Political Science, University of Michigan, Ann Arbor, MI 48109; [c]Department of Computer Science, University of New Mexico, Albuquerque, NM 87131; [d]Santa Fe Institute, Santa Fe, NM 87501; and [e]Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI 48109

Contributed by Robert Axelrod, January 11, 2017 (sent for review October 14, 2016; reviewed by Ross Anderson, Matthew Bishop, and Bruce Schneier)

Cyber conflict is now a common and potentially dangerous occurrence. The target typically faces a strategic choice based on its ability to attribute the attack to a specific perpetrator and whether it has a viable punishment at its disposal. We present a game-theoretic model, in which the best strategic choice for the victim depends on the vulnerability of the attacker, the knowledge level of the victim, payoffs for different outcomes, and the beliefs of each player about their opponent. The resulting blame game allows analysis of four policy-relevant questions: the conditions under which peace (i.e., no attacks) is stable, when attacks should be tolerated, the consequences of asymmetric technical attribution capabilities, and when a mischievous third party or an accident can undermine peace. Numerous historical examples illustrate how the theory applies to cases of cyber or kinetic conflict involving the United States, Russia, China, Japan, North Korea, Estonia, Israel, Iran, and Syria.

cyber conflict | attribution | blame | Bayesian game theory | strategy

When Israel attacked and destroyed a nuclear facility under construction in Syria, the Syrians did not immediately acknowledge the attack, let alone blame Israel (1). Likewise, when China pressured Japan to release a detained fishing boat captain by halting exports of rare earths needed for Japanese electronics, Japan did not accuse China directly (1). Thus, even when the victim can readily establish the identity of the perpetrator, the victim does not necessarily choose to assign blame. In both the Syrian and Japanese cases, direct retaliation was problematic and potentially escalatory, and therefore, the victim chose not to respond. In cases where the victim lacks an appropriate response, publicly blaming the perpetrator without backing it up only makes the victim look weak. Strategic issues of attribution and choice of response are present in many problems of the contemporary era involving nations, nonstate actors, and sometimes, individuals.

In the cyber domain, assigning blame for an attack or intrusion is complicated by both technical factors and lack of agreement on basic definitions (e.g., what constitutes an attack or what counts as critical infrastructure). Sources in or close to the US Government assert that its ability to trace back a cyber operation to its geographic origin (e.g., an urban neighborhood in China) is excellent (2). However, unlike its response to aggression in the physical world, the United States has been surprisingly restrained in responding to incidents, such as the Chinese theft of databases containing the personal information of 21.5 million federal employees (3) or intellectual property (4). Similarly, the Russian data theft from JP Morgan Chase (5, 6) and Iranian cyberattacks against the United States (7) have not provoked a public retaliatory response. The US Government was surprisingly slow to blame Russia for the compromise and leak of documents from the Democratic National Committee, an attack seemingly aimed at influencing the 2016 Presidential election (8). Such restraint is controversial, and some government officials argue publicly that cyber issues should be treated similarly to their physical analogs (9).

There are several reasons why the technical ability to attribute the origin of an attack to a precise location is insufficient alone for the political and strategic purposes of deterring and responding to cyber events:

Technology Is Not Politics. There is often uncertainty about whether the perpetrators are acting for themselves or as agents of another entity (e.g., government). For example, the large Distributed Denial of Service against Estonia's internet infrastructure in 2007 was originally blamed on Russia, which denied responsibility (10). (An Estonian citizen of Russian ethnicity was held responsible for the attack and fined.) In a world where nonstate actors can readily acquire the ability to conduct cyberattacks, holding a government responsible, even for attacks originating within its borders, is not easy.

If the originating location is a military facility, that may suffice to attribute blame to a government. However, even this case can leave open questions of whether the operation was sanctioned or rogue. A related issue is the need for public acceptance. If the methods of attribution are classified or proprietary, security experts or international bodies may be skeptical that the attribution is correct. The United States' attribution to North Korea of the 2014 cyberattacks on Sony is a recent example (11), where experts were skeptical of the initial attribution until additional evidence was discovered (12, 13) and provided by the Federal Bureau of Investigation (FBI) (14) to support the claim.

Evidence Can Be Spoofed. Many of the known attribution methods for cyber can be spoofed (15). Unlike biological or nuclear events, digital records can be copied, altered, created, or deleted; identities can be faked; and attacks can be disguised as accidents or incompetence. Although it is challenging to disguise an attack completely, it is equally challenging to guarantee the provenance of the digital information required for attribution.

Significance

Attribution of cyberattacks has strategic and technical components. We provide a formal model that incorporates both elements and shows the conditions under which it is rational to tolerate an attack and when it is better to assign blame publicly. The model applies to a wide range of conflicts and provides guidance to policymakers about which parameters must be estimated to make a sound decision about attribution and blame. It also draws some surprising conclusions about the risks of asymmetric technical attribution capabilities.

Author contributions: B.E., A.F., S.F., and R.A. designed research, performed research, and wrote the paper.

Reviewers: R.A., University of Cambridge; M.B., University of California, Davis; and B.S., Harvard University.

The authors declare no conflict of interest.

[1] To whom correspondence may be addressed. Email: benjamin.edwards@ibm.com or axe@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1700442114/-/DCSupplemental.

Incentive to Exaggerate. To promote deterrence, countries have an incentive to overstate their cyber capabilities for both attacking and attribution in contrast with nuclear conflict, where attribution of the source of a massive nuclear strike is easy and reliable. There was no doubt, for example, that the Soviet Union was capable of destroying a US city with a nuclear armed ballistic missile, because previous tests had shown their capabilities.

Lack of Appropriate Response. Holding another party responsible by painful punishment (rather than symbolic actions) risks escalation rather than contrition (2). For example, when North Korean attackers compromised Sony Pictures Entertainment and exfiltrated and leaked confidential emails and intellectual property, the United States had no comparable (in-kind) target within North Korea (16). The United States' options included ignoring the attacks; retaliating with a potentially disproportionate response, leading to additional escalation; or retaliating in a different domain. In this case, the United States' response publicly blamed the North Korean government and imposed largely symbolic economic sanctions (13). This kind of case is challenging for the United States, because the victim was a US company but not the government, and the United States has historically been reluctant to assume the responsibility of and authority for protecting the networks of commercial enterprises. Also challenging is the risk that an in-kind cyber response will legitimize behavior that most believe is not legitimate. Thus, an appropriate response is one that is painful but not escalatory, proportionate, and legitimate—difficult constraints to meet in the cyber domain.

Here, we consider how one's technical ability to attribute a cyber operation interacts with strategic considerations, such as the availability of a proportional response or knowledge about the ties between an attacker and a sponsoring organization. Although our analysis focuses on the case of cyber conflict, we note that many of the properties characterizing cyber conflicts are increasingly relevant to physical conflicts. Today, nonstate actors assert their role on the world stage [e.g., the Islamic State of Iraq and the Levant (ISIL)], mingle with civilians (complicating proportional response), use suicide bombers to compensate for the lack of sophisticated weapons, and have nontransparent ties to legitimate states.

Most earlier work on attribution of cyberattacks focused on the technical problem of tracing an attack back to its point of origins (17–20) or the problem of how to assign responsibility to an individual or organization (21, 22). In particular, the diamond model provides a framework for organizing and reasoning about information surrounding a cyber incident (23) but fails to consider the larger geopolitical context in which an attack occurs. Bishop and Goldman (24) argue that attribution is possible not based on effects but based on attacker capability. Schneier (25) discusses the different levels of technical evidence required for attribution, and Clark and Landau (21) explore how the internet might be restructured to enhance attribution of attacks. Technical attribution might be used outside of the context of an attack, and the policy implications of such applications are explored in ref. 26. The strategic component of attributing cyberattacks has been studied in the context of single attacks against specific systems using game theory (27–30), but these models do not account for the larger sociopolitical landscape. This broader context has been studied qualitatively using a variety of models (15, 31, 32). Attribution from an economic perspective was studied by ref. 33, concluding that deterrence is effective for high-value targets and that denial and defense are a wiser investment for lower-value targets. This work provides a formal model that includes both technical and nontechnical aspects of cyber attribution. Fearon (34) provides a model for crises in which actors may quit a conflict and suffer a public cost (as we do here), but in this model, attribution is assumed.

The formal model presented here addresses four questions.

i) What are the conditions under which mutual cooperation (i.e., no attacks) or mutual defection (i.e., reciprocal attacks) is stable?

ii) When is it rational for a player to tolerate attacks by an adversary rather than risk escalating a conflict?

iii) What are the consequences of asymmetric technical attribution abilities among adversaries?

iv) Under what conditions can a third party (such as a nonstate actor) or an accidental attack undermine cooperation between two players?

The Blame Game

We present a Bayesian game-theoretic model that captures many of the features observed in the examples given earlier. The game consists of two players (A and B). Player A is the attacker, and player B is the victim. A and B play a two-part sequential Bayesian game. Player A first chooses whether to attack B. After an attack, A receives gain G, B suffers loss L, and player B chooses whether to blame player A for the attack. (We do not consider false flag incidents, in which B blames A when B knows that A did not attack.) Table 1 summarizes the important features of the game.

Player A is one of two types: vulnerable or not vulnerable. When A is vulnerable, it is vulnerable to B's blame. If B blames a vulnerable A, the result is a loss l to A and a gain g to B. Substantively, we can interpret vulnerability in one of several ways. In a cyber security context, A might be vulnerable, because it is technically susceptible to counterattack from B (and knows it). More relevant to our strategic questions, A could also be vulnerable, because it knows that it is in a tenuous geopolitical position, and it would be detrimental if a high-profile cyberattack that it conducted came to light. The recently documented decrease in Chinese economic espionage against US companies might be an example of this latter case (35). We can think of B's gain g from its response to A as either utility from a counterattack of its own or its increased reputational strength from exposing the attacker.

When A is not vulnerable, B cannot respond to A's attack in a way that lowers A's utility. That is, A receives G, the benefit of its attack, regardless of B's response. A might be not vulnerable, for example, if its attack was so sophisticated that it cannot be definitively traced or if B is reluctant to legitimize A's attack by retaliating. Alternatively, B might not have an in-kind cyber response available. For example, the United States has industrial secrets that could be valuable to China or North Korea, but the converse may not always be true, in which case these countries would be unafraid of the United States retaliating with an industrial espionage counterattack. To summarize, B can hurt a vulnerable A (and gains from doing so), but B cannot hurt a not vulnerable A and pays a cost for trying. The notion that a player can be vulnerable or not vulnerable arises from circumstances that are exogenous to the game.

Player B also has two types: knowledgeable and not knowledgeable. B may or may not be able to distinguish the type of player A (i.e., whether A is vulnerable or not vulnerable). Knowledgeable B knows enough about A's technical capability, the nature of the attack, and the geopolitical context to know whether blaming A will hurt A. When B is not knowledgeable, either it cannot convincingly attribute an attack to A or it cannot determine A's type. We assume that B has prior beliefs about A—the probability that A is vulnerable is v, and the probability that it is not vulnerable is $1 - v$. If B is not knowledgeable, then it must play a strategy based on its prior beliefs about A's type, but if B is knowledgeable, then it can play a strategy conditioned on A's type. In Fig. 1, this situation is indicated by the horizontal dotted line for B's information set, where not knowledgeable B cannot distinguish A's type.

Table 1. Blame game summary

Game element	Description
Actions	Player A chooses to attack or not; if player A attacks, player B may or may not blame it for the attack.
Player types	Player A is or is not vulnerable to blame; player B is or is not knowledgeable about whether player A is vulnerable.
Payoffs	The payoffs to the players depend on their actions and types as shown in Fig. 1.
Beliefs	Player A's belief reflects how hopeful it is that player B cannot determine A's type (vulnerable or not), and it is in the form of a probability estimate; player B's belief is a probability estimate reflecting how confident it is that player A is vulnerable.
Outcomes	The players' types, payoffs, and beliefs determine the equilibria of the game, which are no attack, attack and no blame, or attack and blame.
Analysis	The conditions under which cooperation (i.e., no attacks) is stable, when attacks are tolerated, the consequences of asymmetric capabilities for technical attribution, and when a third party or an accident can undermine cooperation.

When A cannot distinguish B's type, it uses its prior beliefs about these probabilities, denoted as k and $1-k$, respectively, and indicated in Fig. 1 by the dotted lines for A's information set.

A and B play a sequential game. Depending on A's type, the $\{$blame, attack$\}$ strategy has different payoffs. L is the magnitude of the loss suffered by B when it is attacked, G is the magnitude of the gain that A receives from attacking B, l is the magnitude of the loss suffered by a vulnerable A when it is blamed by B, and g is the magnitude of the gain that B receives from blaming a vulnerable A.

If B chooses not to blame A for the attack, B pays the cost of inaction—N if B is knowledgeable and n if B is not knowledgeable, with $0 \leq n \leq N$. For example, after the recent release of private Democratic National Committee emails, there was public outcry over the US Government's inaction when many claimed that there was evidence that the attack came from Russia (36). When B is knowledgeable, inaction is viewed more negatively (e.g., "why can't such a powerful nation respond to a known attacker?") than when it is not knowledgeable.

If B blames a not vulnerable A, B suffers additional cost. If B is not knowledgeable, then B's cost is C, and if B is knowledgeable, the cost is c, with $0 \leq c \leq C$. These variables can be interpreted as the reputation cost incurred by issuing an ultimatum and then not following through, the reputation cost of publicly revealing one's vulnerability/powerlessness, or the cost from direct retaliation by A for being blamed. This interpretation of c and C allows us to consider a one-shot game rather than a sequential game, because the payoff for future rounds is incorporated into these two parameters; c is less than C, because we assume that, if B is knowledgeable, it will suffer less reputation cost than it would for blaming with little evidence. Also, a one-shot sequential game reflects the reality of many cyber conflicts, where even when the same parties have repeated interactions, the parameters of each round can vary [e.g., different attribution certainty, a player's type (knowledgeable or not knowledgeable), or the cost of blaming], making analysis of an iterated game in which each round has different payoffs intractable.

For simplicity, all variables are greater than or equal to zero. *SI Appendix*, Table S1 summarizes the variables and provides a brief description of each one.

Analysis of Outcomes

Analysis of the blame game reveals rational strategies under different conditions and player types (*SI Appendix* shows the proofs). We use the term equilibrium in the game-theoretic sense to mean Nash equilibrium [i.e., a strategy in which neither player has incentive to deviate from its current strategy (37)]. We consider a single round, where A has an opportunity to attack and B has an opportunity to respond by blaming or not blaming. It is sufficient to consider the one-shot sequential game, because future outcomes are incorporated into the C and c parameters.

B's Strategy. We examine B's strategy as a function of whether it is knowledgeable or not knowledgeable.

Theorem 3.1. If B is knowledgeable and if A is vulnerable, B always blames.

Theorem 3.2. If B is knowledgeable and if A is not vulnerable, B will blame if $N > c$.

If B is not knowledgeable, it can use its belief, v, that A is vulnerable. It can then calculate that it pays to blame if $v > v^*$,

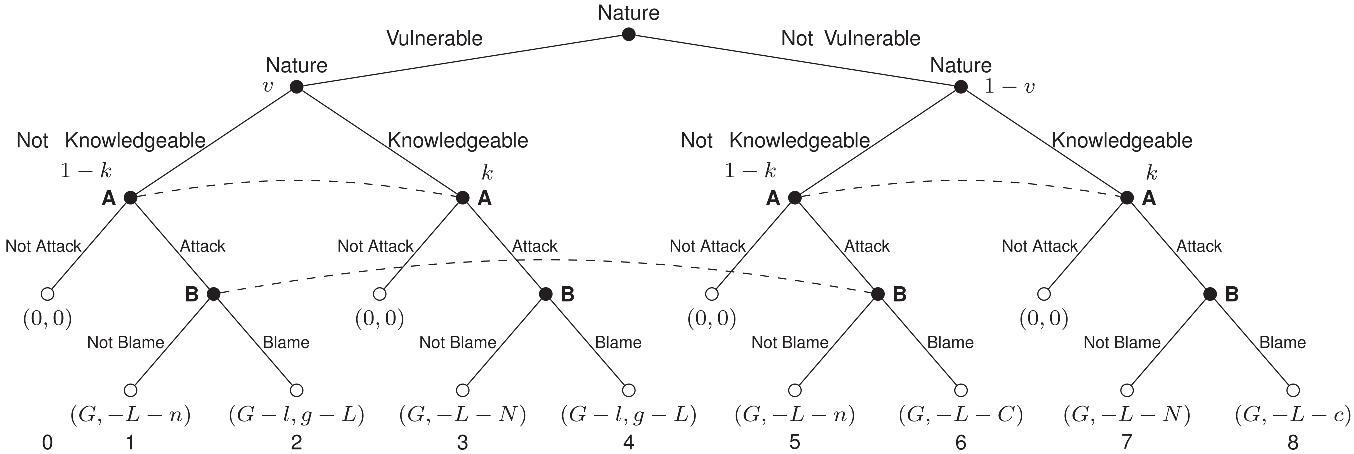

Fig. 1. Extensive form representation of blame game with player types. Numbers at the bottom are used to reference particular outcomes.

where $v^* = \frac{C-n}{g+C}$. When $v > v^*$, we say that B is confident (i.e., certain enough that A is vulnerable to take the risk of blaming A). This expression is a direct result of analysis of the blame game's payoffs but has an intuitive interpretation. More details are in *SI Appendix*. We assume that, when B is knowledgeable, it always seems confident. (This assumption does not affect the analysis or results. An analysis of the game without this assumption is given in *SI Appendix*.)

Theorem 3.3. If B is not knowledgeable but confident, it will blame A.

A's Strategy. In this section, we consider whether A should make the choice to attack B.

Theorem 3.4. If A is vulnerable and if B is known to be confident, A should attack only if $G > l$.

We say that A is worried if A believes that there is a sufficiently high probability that B is knowledgeable, namely $k > \frac{G}{l}$. Conversely, A is not worried if $k < \frac{G}{l}$.

Theorem 3.5. If A is vulnerable and if B is known to be not confident, A will attack if A is not worried.

Theorem 3.6. If A is not vulnerable, then it should always attack.

Equilibrium. From the above theorems, we can establish the conditions for three different equilibria in the blame game.

No Attack.

A is vulnerable, B is knowledgeable, and $G < l$.

A is vulnerable and worried, and B is not knowledgeable and not confident.

A is vulnerable, B is not knowledgeable but confident, and $G < l$.

A attacks, and B does not blame.

A is vulnerable and not worried, and B is not knowledgeable and not confident.

A is not vulnerable, and B is not knowledgeable and not confident.

A is not vulnerable, B is knowledgeable, and $N < c$.

A attacks, and B blames.

A is vulnerable, B is knowledgeable, and $G > l$.

A is vulnerable, B is not knowledgeable and confident, and $G > l$.

A is not vulnerable, and B is not knowledgeable and confident.

A is not vulnerable, B is knowledgeable, and $N > c$.

Equilibria that include attacks occur even when B is knowledgeable. That is, even with high technical attribution abilities and sophisticated knowledge of the political landscape, B cannot always deter attack, because knowledge alone is not sufficient in the game. Moreover, the no attack equilibrium exists only when A is vulnerable.

Discussion

Analysis of the blame game reveals several interesting conclusions. First, if A is not vulnerable, then B cannot prevent A from attacking. If B increases its attribution ability or decreases A's gain from an attack (say by establishing stronger defenses), it will not change A's best strategy. In this situation, B's best option is to minimize its damage by lowering N (e.g., by reducing public outcry) or reduce c, the cost of blaming A. Surprisingly, it is rarely beneficial for B to increase its own attribution ability. For example, if A is vulnerable, increasing B's confidence only increases its certainty of retaliation, not its ability to deter future attacks. B only has an incentive to increase its attribution ability (v), when C increases or g decreases. The only effective way to deter A is by decreasing the gain (G) that A receives for attacking or increasing the loss (l) that A experiences when blamed.

If A is vulnerable to blame, then B has other options, especially if it has an appropriate response available. An in-kind cyberattack could be the most appropriate and could lower the cost of C or c, the future cost of not blaming A. That is, if B can carry out its own attacks in the near future, then the cost for not blaming in the current round is reduced. Recalling that an appropriate attack must be legitimate, B may be reluctant to respond in kind because of the risk of legitimizing a form of attack that it views as illegitimate. For example, the United States asserts that economic espionage is illegitimate and may, therefore, be reluctant to punish China "in kind" for infiltrating US companies.

Next, we return to the four motivating questions and discuss how the blame game answers them.

In the context of the game, the first question asks when mutual cooperation (no attacks) and mutual defection (attack and no blame) are stable. We assume, in cyber conflict, that both players have the ability to play the role of A by initiating an attack and that the attacked party plays the role of B. Thus, mutual cooperation is the condition when neither player chooses attack, a condition that we analyze once, because the conditions hold reciprocally (and similarly for mutual defection). Our analysis shows that (*i*) the stability of mutual cooperation requires that both players be vulnerable and (*ii*) if the victim is not knowledgeable, then high confidence about the vulnerability of the other player and low reward for attacking lead to mutual cooperation.

In the case of recent US/China interactions, many attacks likely fail the first condition if an in-kind response is inappropriate. This failure could explain why indicting specific individuals was the best available option, even if it had little effect on the individuals themselves or the government (38, 39).

Turning to the second question, which asks when it is rational to tolerate an attack, the attack/no blame equilibrium addresses this case. We note that this equilibrium is stable regardless whether B is knowledgeable. If A is not vulnerable, $c > N$ (perhaps the case for Chinese industrial espionage), and N is small (low public outcry), then it may not be rational for B to reveal its technical abilities by identifying the cyberattacks (large c). In this case, tolerating the attacks is rational. A contrasting example comes from the recent hacking campaign against the Democratic National Committee and various other targets during the 2016 US presidential election, which has been attributed to the Russian Government. Mounting public pressure (growing N) led the United States to publicly blame Russia, promising "...a strong diplomatic, political, cyber and economic response" (8), even when President Obama acknowledged that "...the idea that somehow public shaming is gonna be effective, I think doesn't read the—the thought process in Russia very well" (40). This statement is consistent with our interpretation of a not vulnerable Russia.

Another example could be the Stuxnet attack against Iran, where Iran did not immediately blame the attackers publicly (presumably the United States and Israel), which would be rational if Iran believed the attackers were not vulnerable.

The third question asked about the consequences of asymmetric attribution capabilities. Recalling that the game is symmetric, meaning that either player can decide to attack, if one player is not knowledgeable and has sufficiently low belief (v) in the opponent's vulnerability, then it is likely to be a victim. Under these circumstances, it might instead decide to attack preemptively, because it has little to lose. This analysis shows that stability could be increased if both players become knowledgeable through improved technical attribution capabilities.

The fourth question asked when cooperation between two parties can be undermined by a third party, such as a nonstate actor

Table 2. Equilibria in the blame game as a function of player types

	B is Knowledgeable		B is Not Knowledgeable	
A is Vulnerable	🟧 $G > l$		🟧 B is confident and $G > l$	⬜ A is not worried and B is not confident
	🟪 $G < l$		🟪 B is confident and $G < l$	🟪 A is worried and B is not confident
A is Not Vulnerable	🟧 $N > c$	⬜ $N < c$	🟧 B is confident	⬜ B is not confident

Key			
Key	🟧 Attack, Blame	⬜ Attack, No Blame	🟪 No Attack

or an accidental attack. If B is attacked and believes that A is responsible, then B will blame A under the conditions elaborated above. If B blames A, we can turn the game around and ask about the conditions under which A (as the victim) will attack B. When all of these conditions are met, a third party or an accident can undermine mutual cooperation between two players.

Beyond answering the four questions for cyberattacks, the blame game is applicable to some noncyber situations. Returning to an earlier example, it explains why it was reasonable for Syria to not blame Israel for attacking its nuclear facility. Syria (B) likely knew that Israel (A) was responsible for the attack, but Israel was not vulnerable, and Syria was knowledgeable. Moreover, if Syria blamed Israel but did nothing about it, it would suffer a reputation cost ($N < c$). In terms of the blame game tree in Fig. 1, Syria preferred outcome 8 to outcome 7 and, therefore, tolerated the attack. As Table 2 shows, when A is not vulnerable, B is knowledgeable, and $N < c$, the predicted outcome is that A attacks and that B does not blame—exactly what happened.

The case of China pressuring Japan for release of a detained sea captain by halting China's exports of rare earths is analogous to the Israel–Syria case. China (A) was not vulnerable to a Japanese response, and Japan (B) knew it. Japan would suffer a reputation cost if it blamed China without punishing ($N < c$). As predicted by the model, when China halted exports of rare earths, Japan did not blame China.

As a third example, in 2008, Hamas (A) governed Gaza. A rogue group in Gaza, Palestinian Jihad (A') attacked Israel (B) with rockets. Israel responded with a warning: "Hamas controls the Gaza Strip and they are accountable for every active aggression against Israel. We will not allow Hamas to subcontract out terrorism" (41). In effect, Israel was saying that, because Hamas could control Islamic Jihad, Israel would blame Hamas if there was another attack from Gaza. Given that Hamas was vulnerable and Israel was knowledgeable, Israel's threat was credible. Table 2 predicts that Hamas would try to restrain Islamic Jihad to prevent another attack on Israel if $G < l$, and Hamas has usually done so.

Our model is general enough that it could be applied to areas outside of nation state-level conflict. Children bullied by peers or abused by adults may need to decide whether to blame a potentially invulnerable attacker in a position of power. Individuals may be unable to blame large corporations for wrongs if excessive costs have rendered litigation impossible and corporations invulnerable.

Conclusion

This paper studies the strategic aspects of attribution and blame, especially in the context of cyber conflicts. We define a game-theoretic model called the blame game, and its analysis shows that, in many cases, it may be rational for nations to tolerate cyberattacks, especially if they are relatively mild and if no appropriate response is available (42). Tolerance may even be rational in the face of strong public criticism (43). Highly unbalanced attribution capabilities between adversaries can increase the risk of conflict. Although we emphasized examples from the cyber domain, the game is also relevant to some kinetic conflicts, especially those involving nonstate actors.

The analysis provided here is not intended to suggest specific cyber policies. Rather, our model provides concepts and parameters that can be helpful in formulating the questions that a policymaker might want to ask in a particular setting to predict the outcomes of specific choices.

For the attacker, A, the questions proceed as follow. The first question is "am I vulnerable to blame?" If the answer is no, then A should attack, because there is likely to be no consequence to being blamed. If A is vulnerable, the next question is "is B confident that I am vulnerable ($v > \frac{c-n}{G-c}$)?" If the answer is yes, then the next question is "is the gain from attack higher than the cost of blame ($G > l$)?" (or are the payoffs at nodes 2 and 4 greater than that at zero in Fig. 1). A yes answer suggests that attack is the right action. If B is not confident, then the last question for A is "am I worried that B is knowledgeable ($k > \frac{G}{l}$)?" If not, then A should attack, and if so, it should keep the peace. Determining whether to initiate an attack requires estimation of the other players' belief in one's own type.

For the victim, B, the questions proceed as follows. The first question is "am I knowledgeable about A's type?" A knowledgeable player will know its attacker's type. If A is not vulnerable, B should ask "is $N > c$ (is node 8 preferable to node 7 in Fig. 1)?" (that is, is the cost of doing nothing higher than the cost of blaming), and if yes, then B should blame. Knowledgeable B should always blame attacks from A if A is vulnerable. If B is not knowledgeable, the second question to be asked is "am I confident that B is vulnerable ($v > \frac{c-n}{G-c}$)?" If yes, then B should blame.

Although the questions above are straightforward, the answers are not. Determining reasonable values for any of the parameters is clearly challenging for policymakers. For example, the confidence of an adversary can be difficult to determine, requiring estimation of the public outcry for not blaming what may appear to be an obvious attack.

Our model quantifies these strategic and technical aspects of attribution and how they interact, and it highlights the questions that actors in cyber conflicts should try to answer before taking action or responding to the actions of others. We hope that the model will help policymakers identify gaps in their knowledge and focus on estimating parameters in advance of new cyberattacks.

ACKNOWLEDGMENTS. We thank Alton Worthington, Brandon Kaplowitz, Herb Lin, Steve Crocker, Joseph Nye, Vinh Nguyen, Cari Martinez, Susan Landau, and Sam Liles for their comments on early drafts of this paper. S.F. and R.A. were partially supported by National Science Foundation (NSF) Grant 1444871, and S.F. acknowledges NSF Grant 1518878, Defense Advanced Research Projects Agency Grant FA8750-15-C-0118, Air Force Research Laboratory Grant FA8750-15-2-0075, and the Santa Fe Institute.

1. Axelrod R, Iliev R (2014) Timing of cyber conflict. *Proc Natl Acad Sci USA* 111(4):1298–1303.
2. Clarke RA, Knake RK (2011) *Cyber War* (HarperCollins, New York).
3. Davis JH (July 10, 2015) Hacking exposed 21 million in U.S., government says. *New York Times*, Section A, p 1.
4. Wong E (May 23, 2013) Hackers find China is land of opportunity. *New York Times*, Section A, p 1.
5. Riley M, Robertson J (August 27, 2014) FBI said to examine whether Russia tied to JPMorgan hacking. *Bloomberg*. Available at https://www.bloomberg.com/news/articles/2014-08-27/fbi-said-to-be-probing-whether-russia-tied-to-jpmorgan-hacking. Accessed February 4, 2017.
6. Silver-Greenberg J, Goldstein M, Perlroth N (October 3, 2014) Hackers' attack struck systems at 10 companies. *New York Times*, Section A, p 1.
7. Nakashima E, Zapotosky M (March 24, 2016) U.S. indicts 7 in connection with cyberattacks linked to Iranian government. *Washington Post*, Section A, p 14.
8. Nakashima E (October 7, 2016) U.S. officially condemns Russia over hacking. *Washington Post*, Section A, p 1.
9. Waterman S (June 30, 2016) US needs to publicly attribute cyberattacks, former House intel chair says. *Fedscoop*. Available at https://www.fedscoop.com/fmr-rep-rogers-u-s-needs-to-publicly-attribute-cyberattacks/. Accessed February 4, 2017.
10. Schmidt A (2013) The Estonian cyberattacks. *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*, eds Healy J (Cyber Conflict Studies Association, Arlington, VA), pp 174–193.
11. FBI National Press Office (December 19, 2014) Update on Sony investigation. Available at https://www.fbi.gov/news/pressrel/press-releases/update-on-sony-investigation. Accessed February 4, 2017.
12. Appelbaum J, et al. (December 28, 2014) Inside the NSA's war on internet security. *Der Spiegel Online*. Available at http://www.spiegel.de/international/germany/inside-the-nsa-s-war-on-internet-security-a-1010361.html. Accessed February 6, 2017.
13. Sanger D, Fackler M (January 18, 2015) Tracking the Cyberattack on Sony to North Koreans. *New York Times*, Section A, p 1.
14. Scannel K (January 8, 2015) FBI details North Korean attack on Sony. *Financial Times*. Available at https://www.ft.com/content/287beee4-96a2-11e4-a83c-00144feabdc0. Accessed February 4, 2017.
15. Rid T, Buchanan B (2015) Attributing cyber attacks. *J Strat Stud* 38(1-2):4–37.
16. Sanger DE, Perlroth N (December 18, 2014) U.S. is said to find North Korea behind cyberattack on Sony. *New York Times*, Section A, p 1.
17. Wheeler DA, Larsen GN (2003) *Techniques for Cyber Attack Attribution* (Institute for Defense Analysis, Alexandria, VA), Tech Rep ADA468859.
18. Mandiant (2013) APT1: Exposing One of China's Cyber Espionage Units. Available at https://www.fireeye.com/content/dam/fireeye-www/services/pdfs/mandiant-apt1-report.pdf. Accessed February 6, 2017.
19. Kaspersky Labs (2015) Equation group: Questions and answers. Available at https://securelist.com/files/2015/02/Equation_group_questions_and_answers.pdf. Accessed February 4, 2017.
20. McVey SB (2015) Cyber attribution: Useful evidence in attributing malware and cyber attacks. PhD thesis (Utica College, Utica, NY).
21. Clark DD, Landau S (2011) Untangling attribution. *Harvard Law School National Security J* 2(2):323–352.
22. Hunker J, Gates C, Bishop M (2011) Attribution requirements for next generation internets. *Proceedings of the IEEE International Conference on Technologies for Homeland Security* (IEEE, Piscataway, NJ), pp 345–350.
23. Caltagirone S, Pendergast A, Betz C (2013) *The Diamond Model of Intrusion Analysis* (Center for Cyber Intelligence Analysis and Threat, Hanover, MD), Tech Rep ADA586960.
24. Bishop M, Goldman E (2003) The strategy and tactics of information warfare. *Contemp Security Policy* 24(1):113–139.
25. Schneier B (January 5, 2015) We still don't know who hacked Sony. *The Atlantic*. Available at https://www.theatlantic.com/international/archive/2015/01/we-still-dont-know-who-hacked-sony-north-korea/384198/. Accessed February 4, 2017.
26. Bishop M, Gates C, Hunker J (2009) The sisterhood of the traveling packets. *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, eds Ford R, Heydari MH, Somayaji A (Association for Computing Machinery, New York), pp 59–70.
27. Van Dijk M, Juels A, Oprea A, Rivest RL (2013) FlipIt: The game of stealthy takeover. *J Cryptology* 26(4):655–713.
28. Roy S, et al. (2010) A survey of game theory as applied to network security. *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences* , eds Sprague RH (IEEE, Los Alamitos, CA), pp 1–10.
29. Manshaei MH, Zhu Q, Alpcan T, Bacşar T, Hubaux JP (2013) Game theory meets network security and privacy. *ACM Comput Surv* 45(3):25.
30. Yan G, Lee R, Kent A, Wolpert D (2012) Towards a Bayesian network game framework for evaluating DDoS attacks and defense. *Proceedings of the 2012 ACM Conference Computer and Communications Security* (Association for Computing Machinery, New York, NY), pp 553–566.
31. Brenner SW (2007) "At light speed": Attribution and response to cybercrime/terrorism/warfare. *J Crim Law Criminol* 97(2):379–475.
32. Libicki MC (2009) *Cyberdeterrence and Cyberwar* (Rand Corporation, Santa Monica, CA).
33. Lindsay JR (2015) Tipping the scales: The attribution problem and the feasibility of deterrence against cyberattack. *J Cybersecurity* 1(1):53–67.
34. Fearon JD (1994) Domestic political audiences and the escalation of international disputes. *Am Polit Sci Rev* 88(3):577–592.
35. Nakashima E (June 21, 2016) Economic cyberespionage by China has dropped steeply, security firm says. *Washington Post*, Section A, p 3.
36. Sanger DE, Schmitt E (July 26, 2016) Spy agency consensus grows that Russia hacked D.N.C. *New York Times*, Section A, p 1.
37. Maschler M, Solan E, Zamir S (2013) Game Theory (Cambridge Univ Press, Cambridge, UK).
38. Nakashima E (December 1, 2015) Following US indictments, Chinese cybertheft waned after U.S. indictments. *Washington Post*, Section A, p 3.
39. Sanger D (September 19, 2015) U.S. and China seek arms deal for cyberspace. *New York Times*, Section A, p 1.
40. Obama B (December 15, 2016) Press conference by the president. *White House*. Available at https://obamawhitehouse.archives.gov/the-press-office/2015/12/18/press-conference-president-121815. Accessed February 4, 2017.
41. al Mughrabi N (March 13, 2008) Islamic Jihad rockets hit Israel after West Bank raid. *Reuters*. Available at http://www.reuters.com/article/us-palestinians-israel-idUSL1278499020080313. Accessed Feb 4, 2017.
42. Alexander KB (July 13, 2016) Digital acts of war: Evolving the cybersecurity conversation before the subcommittees on information technology and national security of the Committee on Oversight and Government Reform. *Testimony to the House Oversight Committee*. Available at https://oversight.house.gov/wp-content/uploads/2016/07/Gen-Alexander-Statement-Digital-Acts-of-War-7-13.pdf. Accessed February 4, 2017.
43. Washington Post Editorial Board (August 12, 2015) The U.S. has been complacent and lazy in responding to cyberattacks. *Washington Post*. Available at https://www.washingtonpost.com/opinions/the-us-has-been-complacent-and-lazy-in-responding-to-cyberattacks/2015/08/12/d10040c2-3d4d-11e5-8e98-115a3cf7d7ae_story.html. Accessed February 4, 2017.