15 Understanding transnational cyber attribution

Moving from “whodunit” to who did it

*Brenden Kuerbis, Farzaneh Badiei, Karl Grindal,
and Milton Mueller*

Like the broader field of cyber security, cyber attribution is a socio-technical endeavor. Accordingly, we can expect that any transformation of cyber attribution will “be co-constituted by technological possibilities, political choices, and scientific practices” (Dunn Cavelti and Wenger 2020). This chapter examines some of the current practices of attribution, scientific developments in the field, and possibilities for its transnational institutionalization.¹

First, we provide some background on the role of attribution in deterrence and accountability, and the challenges of attributing, particularly to nation-state actors. We then analyze attributions made from 2016 to 2018. We characterize the actors involved and types of attributions, finding a shift toward private actor attributions and mix of approaches by states. Next, we explore some technical advances in attribution. Better algorithmic-driven attribution, seemingly possible by the collection and analysis of numerous artifacts left on networks by threat actors, could certainly help push attribution forward although it raises issues of how state and non-state actors cooperate. We also look at attempts to understand behavioral aspects of attribution, exploring one game-theoretic attempt to model when states will or will not attribute an attack to another state, and use our dataset to explore certain predictions of the model. This exercise allows us to understand more clearly which state and non-state actors make or avoid making attributions, and the institutional conditions under which their agreement on attribution might occur.

In light of the above analysis, it is unlikely that attribution made by a nation-state (or even allied states) will be accepted as neutral and authoritative by another state, especially if those states are rivals or hostile. Given political fragmentation and socio-technical uncertainty around the current practice of attribution, we review proposed models for institutionalizing transnational attribution. The initial models offered have dramatically different structures and actor participation. Recognizing the shortcomings in them and the strategic use of attribution by states, a group of university-based and independent researchers are seeking to build independent, transnational attribution capabilities grounded in scientific method. Such a collective approach, if recognized, could address credibility

issues and result in more stable outcomes and ultimately help with accountability. We conclude with a brief agenda for future research.

The role of cyber attribution in deterrence and accountability

One can defend against a cyberattack, but without attribution, attackers lack a deterrent. At best, secure systems increase the time needed to find a vulnerability to a point beyond that which the attacker is willing to spend. Without proper incentives to restrain malicious attacker behavior, be they state or non-state, it's unreasonable to expect the present situation to change. As a deterrent, attribution has several advantages over other responses: Unlike strategies such as hack backs, it might not result in the militarization of the internet and it might even prevent it (Dunn Cavelty 2012).

Accurate attribution requires experienced threat intelligence and digital forensics experts. While governments and threat intelligence groups will attribute attacks to specific intrusion sets, sometimes even linking these to specific actors, there is no internationally recognized forensic process with an evidentiary based level of confidence. Rather, attribution is more often than not based on limited evidence and the reputation of the attributing entity. Considering that both attributing groups and attackers could be based anywhere in the world, without a recognized standardized and institutionalized process for attribution, can we expect a global coalition to implement sanctions or otherwise deter the attacker?

There is an important distinction between identifying intrusion sets and assigning them to an adversary or "threat group" on the one hand and linking this adversary with a known state or non-state actor on the other. Robert Lee refers to the latter as "true attribution" (Lee 2016). This two-part distinction can be compared to Herb Lin's model, developed in the article "Attribution of Malicious Cyber Incidents", which uses three levels of attribution: Machines, human operators, and the ultimate party responsible (Lin 2016). In Mandiant's 2013 attribution of an Advanced Persistent Threat (i.e., "APT-1") to the China PLA Unit 612398 all three levels of Lin's model are described (Wittes 2013). At the lowest level would be IP addresses associated with command and control (C&C) servers. Next, is attribution to a human operator; the Mandiant report identifies a persona who went by the alias "ugly gorilla", but associated this with the real person, Wang Dong. Ultimately though, the report is attributing APT-1 to China's People's Liberation Army and hence the Chinese state.

Defining an ultimate responsible party can be particularly challenging when it comes to state involvement. Even when a person is clearly identified as being in the attributed country, it is not necessarily clear from the forensics whether that person was a contractor or an employee, or whether they were operating under express instructions or on their own. Jason Healey's Spectrum of State Responsibility acknowledges that states employ hackers, contract out hacking,

encourage hacking, or permit its use within their jurisdiction, each level representing a different degree of state responsibility (Healy 2013).

The challenge of attribution to nation-state actors

The practice of attribution can be cumulative, grouping information from incidents to create intrusion sets. Intrusion sets are adversarial behaviors, what is sometimes abbreviated as “tactics, techniques, and procedures (TTPs)”, and technical resources with common properties from previous attacks that are grouped together (e.g., a “campaign”) and associated with a common actor (e.g., a “threat group”). This process has some general standardization by convention and predictive success, but there is no one correct method. Accordingly, SANS in 2010 noted:

There is no rule of thumb or objective threshold to inform when linked intrusions should become a campaign. The best measure is results: if a set of indicators effectively predict similar intrusions when observed in the future, then they have probably been selected properly.

(Cloppert 2010)

This predictive modeling creates important questions about degrees of confidence, and how the practice of threat intelligence responds to novelty. Assuming an incident is correctly associated with an intrusion set, how is this intrusion set linked to a specific actor? Information like common language, activity during specific hours, the choice of targets, and level of complexity are often used to associate an incident group with a specific responsible threat actor. But this type of attribution extends beyond a purely technical association. The reuse of certain TTPs can complicate this attribution. For example, the vulnerability EternalBlue is reported to have been developed by the NSA, but was later exploited by Russia, North Korea, and Iran (Segal 2018).

Attribution conceptual frameworks help digital forensics to structure collected information and compare it to known intrusion sets. Examples of these include, the Diamond Model of Intrusion Analysis developed by Caltagirone and Pendergast (2013), and the “Q-model” developed by Rid and Buchanan (2015). Both the Diamond Model and Q-model acknowledge the need for a nontechnical dimension to attribution. In the Diamond Model, the nontechnical dimension is described by the relationship between the victim and adversary. The strategic dimension of the Q-Model is described as a “function of what is at stake politically” (Rid and Buchanan 2015).

While the political dimension of attribution might be quantified, it is necessarily relational, a product more of political science or intelligence studies than computer science. As sanctions or other disincentives are used to punish offensive cyber operations, we might expect cyber operations to adjust by taking steps to disguise their identity. The CIA’s leaked Marble Framework, for example, has

been described as providing the capability to change the language of the source code from English to another language like Russian or Farsi (Burgess 2017). Meanwhile, cyber tools invented by one country are being reused by another. This suggests a technical race between forensic experts and counter-forensic obfuscation. While obfuscation might serve powerful states well in the short term, it does little to mitigate the long-term damage of offensive cyberattacks. There is also the inequity of state attribution capability. This is said to have played a role in the breakdown of the UN Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (UN GGE) (Schmitt and Vihul 2017).

Attribution processes today

Our preliminary research has started to categorize the origin and characteristics of publicly attributed incidents. This work builds on the Council on Foreign Relations (CFR) dataset of state-sponsored cyber incident² (Segal and Grigsby 2018). Reviewing 82 incidents identified by CFR between 2016 and the first quarter of 2018 (Table 15.1), we coded each case, identifying whether states and/or private actors made a public attribution, as well as details related to the attribution,³ including timing and outcome.

We understand that publicly disclosed incident databases can be criticized as being just the tip of the iceberg, and that two years of data based on a single dataset is not conclusive. However, this data, which has been supplemented with some of our own observations, is one of the most complete data sources available, and is superior to the anecdotal treatment attribution usually gets. Several interesting initial observations can be made. First, the vast majority of incidents, 70 (85%), resulted in some form of public attribution, with only 12 incidents (15%) not being attributed to a perpetrator. A small number of incidents, 7 (9%), included attributions involving both government(s) and private actor(s). These public attributions may have involved coordinated action between states (e.g., NotPetya) or states and non-state actors (e.g., WannaCry), or attributions published by non-state actors citing anonymous government sources, or what appeared to be separate attributions made independently by private actors and states (e.g., Democratic National Committee hacks).

Table 15.1 Incident attributions made by actor type

<i>Actor type</i>	<i>Year</i>			
	<i>2016</i>	<i>2017</i>	<i>2018 1Q</i>	<i>Grand Total</i>
No attribution made	6	5	1	12
Both government(s) and private actor(s)	4	3		7
Government(s)	7	7	1	15
Private actor(s)	12	26	10	48
Grand Total	29	41	12	82

Fifteen incidents (18%) were attributions made by government(s), including where identified government officials informally “named and shamed” alleged perpetrators, or formally accused them in official statements, reports, sanctions, or indictments. The largest number of attributions have been made by private actors, a category that includes threat intelligence organizations, network security companies and news media organizations. The importance of these actors in attribution is evident from the number of attributions made by them, which seems to be nearly doubling over the past three years. It also highlights the need for a standardized attribution process.

The incident data also allow important distinctions to be made. Table 15.2 shows attributions made to threat group(s) or state sponsor(s) by the actor type making the attribution. The total number of attributions made differs from the number of incidents (Table 15.1) as more than one entity may be implicated per incident by one or more actor type. Consistent with the incident observations above, private actors made substantially more attributions to both threat groups (31 versus 5) and state sponsors (38 versus 13) than governments. The majority of attributions made by government(s) were made to a state sponsor. These attributions included the United States and allied countries accusing Iran, Russia, and North Korea, as well as the United States implicating itself. As noted previously in Table 15.1, governments made attributions in 15 incidents. Table 15.2 shows that governments attributed those incidents to state sponsors 13 times. Governments (in this case, the United States) attributed an attack to a threat group five times; three of those times the attribution was to both a threat group (APT28, APT 29, Lazarus) *and* an alleged state sponsor (Russia, North Korea). Only twice did a government (in this case, Switzerland) not attribute to a state sponsor, but limited its accusation to a threat group (Turla) although a state sponsor was suspected. However, despite the appearance, a Chi-Square test concludes there is no significant difference between actor type (i.e., governments or private actors) with regard to whom (threat group or state sponsor) they attribute incidents. Neither actor type is more likely, or perhaps better suited, to make attributions to a threat group or state sponsor.

An evaluation of the collected attribution documents, namely Executive Orders (in the United States), criminal complaints, indictments, sanctions, and government statements, reveals that the United States’ current attribution practice is possibly the most elaborate compared to other countries, using various attribution methods and judicial processes. Table 15.3 shows states use the judicial system and forensic

Table 15.2 Attributions made by actor type to actor type

<i>Attribution made by (actor type)</i>	<i>Incidents attributed to threat group</i>	<i>Incidents attributed to state sponsor</i>
Both government(s) and private actor(s)	4	7
Government(s)	5	13
Private actor(s)	31	38
Grand Total	40	58

Table 15.3 Attribution documents

<i>Document type</i>	<i>Year</i>			<i>Grand Total</i>
	<i>2016</i>	<i>2017</i>	<i>2018 1Q</i>	
Criminal complaint			1	1
Executive order	1			1
Indictment	1	2	2	5
Sanction	1		2	3
Statement		3	7	10
Media report	5	6	6	17
Press release	1	1		2
Technical report	14	30	11	55
Grand Total	23	42	29	94

technical evidence to carry out the attribution. All the collected indictments (5) are issued by the United States. The US government's approach is different from other countries when issuing official statements (such as White House Press Secretary announcements). It usually collaborates with the Department of Justice (DoJ); after the DoJ receives the indictment, the Office of Asset Control (OFAC) imposes sanctions on the indicted individuals (US Department of Treasury 2018). While other countries such as New Zealand, Australia, the United Kingdom, and Canada have special agencies that might get involved with attribution, the outcome of attribution is usually announced by government agencies and ministries and no national court is involved in charging the attackers.⁴

Various actors including the private sector and government agencies issue technical reports containing indicators of compromise or threat intelligence that attribute cyberattacks. These reports are more common than other forms of attribution documents. This might be due to the fact that issuing technical reports might be easier than going through complex and lengthy judicial processes and issuing alerts based on technical reports by government agencies might prevent further damage and stop the cyberattack from scaling.

The evolution of the US approach

Until recently, the US approach to attribution was as follows: The prosecutor gathered technical and circumstantial evidence about the identity of the adversary and as well as their direct or indirect links to the responsible state. Then the prosecutor would file an indictment against the alleged attacker(s) in the federal court, which grants indictments through the grand jury. This process was lengthy, and documents would not be unsealed until many months after the filing of the indictment. The grand jury would then issue the indictment and the Department of Justice would release a statement along with sanctions being imposed on the alleged attackers through OFAC.

Over time, US prosecutors' filings have become more complex, relying on more evidence to receive the indictment. However, a "complaint" was filed against the alleged perpetrator of WannaCry and not an indictment. A complaint is different in that it is not issued by a grand jury and the prosecutor can decide to file the complaint, naming the individual to be arrested. It should include an affidavit by a prosecutor or a law enforcement official familiar with the case (US Department of Justice 2014, 2018). The prosecutor in the criminal procedure decides to file a complaint if a crime is imminent and it proves "probable cause". After the arrest is made based on a criminal complaint, the federal prosecutor must secure an indictment (with a limited amount of time) to proceed with the felony charge(s).

In an indictment, a grand jury hears evidence and testimony from witnesses presented by the prosecution. It also has the power to subpoena witnesses. But grand jury proceedings are closed to the public and secret, the defense has no opportunity to present evidence or challenge the prosecution evidence. The probable cause standard is one of the lowest in criminal law; only enough evidence that convinces a reasonable person to believe that a crime has been committed must be established. Once an indictment is issued, there is a very small chance that it will be dismissed. Hence it provides higher certainty in the case of attribution that the charges are based on strong grounds.

Despite the procedural pitfalls of an indictment relative to a complaint, it is stronger and might be procedurally more just. The US Treasury's quick reaction to the conspiracy complaint that was filed against the WannaCry alleged attacker and North Korea and the imposition of immediate sanctions based on that and not an indictment reduces the procedural standards of attribution even further.

Other national approaches to attribution

Countries other than the United States also attribute cyberattacks to nation states, either by supporting another state's statement or action or by carrying out their own cyber attribution through their national cyber security agencies. New Zealand's National Cyber Security Center is a government center that has been involved with attribution, and its most common target of attribution is to states:

The NCSC's most common form of attribution occurs when an incident is detected or discovered that contains indicators or technical artefacts previously associated with a state-sponsored actor. These indicators and artefacts come from numerous sources including the NCSC's own analysis and partner and open source reporting.

(National Cyber Security Center of New Zealand 2016)

The UK's National Cyber Security Centre (NCSC) similarly gets involved with attributing cyber-attacks to states actors. It issues technical alerts in collaboration with other countries' government agencies, for example the US DHS and FBI in the case of NotPetya (UK National Cyber Security Centre 2018), as well as the

assessment of whether a state actor has been involved with a cyberattack. The UK Foreign Office minister relies on those assessments to issue statements condemning such cyberattacks (UK Foreign Office 2018).

New developments in advancing attribution

Within the private sector and academia, research into attribution has advanced on technological and behavioral fronts. Promising technologies are emerging to significantly improve the forensic confidence in attribution. New areas of research include improved monitoring of infrastructure and application of machine learning to identify anomalous network traffic possibly indicative of adversaries (e.g., Radford et al. 2018). Our colleagues at Georgia Tech's Institute for Internet Security & Privacy are also investigating attribution as part of the Rhamnousia project (Toon 2016). The work is sponsored by the United States' Defense Advanced Research Project Agency's Enhanced Attribution program, which seeks to "develop technologies to associate the malicious actions of cyber adversaries to individual cyber operators and then to enable the government to reveal publicly the malicious actions of individual cyber operators without damaging sources and methods" (DARPA 2018). At a high level, the Rhamnousia project seeks to connect large sets of disparate data artifacts to fuel new algorithmic attribution methods that will expedite the process of attribution. As such, the process of conducting a cyberattack leaves numerous observable data artifacts on adversary-controlled and victims' networks, as well as on networks in-between. Data includes, but is not limited to, behavioral biometrics from user devices, network traffic, and intrusion detection logs, as well as Domain Name System (DNS) use and registrations (Keromytis 2016).

In some cases, this data can be used to help identify what are presumably nation-state adversaries. For instance, researchers at ETH Zurich were able to reliably determine C&C infrastructure used by APT campaigns by examining web query data (Lamprakis et al. 2017). Applying machine learning techniques to detect and cluster data observed across multiple networks and associate it with APT threat actors continues to advance (Ghafir et al. 2018; Rubio et al. 2020). As mentioned earlier in the case of APT-1, these technical data, when merged with other data like open source and other intelligence can be linked to adversary personas, real-world identities, and in some cases, responsible state entities. The above research efforts represent steady improvements that will continually evolve in response to changing adversarial tactics, and may increase the speed, confidence, and breadth of attribution. But these efforts also raise questions about data collection and sharing between private actors and/or governments, methodological transparency and reproducibility of analysis, effective public communication, and interaction with other legal and political attribution processes.

Behavioral understanding of when and how actors engage in public attribution of nation-state attacks is also advancing. Edwards et al. (2017) study the strategic aspects of attribution and blame in the context of cyber conflicts between attacking and victim states. They present a Bayesian game-theoretic model,⁵ in which

the decision to blame an attacker “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”. In their model, vulnerability refers to an attacking state being technically susceptible to counterattack (e.g., in the case of states with low cyber capabilities, or large attack surface) or being in a tenuous geopolitical position, where it would be detrimental if a high-profile cyberattack that it conducted came to light (e.g., in the case of states with offensive capabilities). Knowledgeable victims are able to distinguish the type of its attacker (vulnerable or not) and have the requisite technical capability and understanding of the nature of the attack as well as geopolitical context to know whether blaming will hurt the attacker. Unknowledgeable victims cannot determine its adversary’s type or convincingly attribute an attack.

While their analysis focuses on states, it draws several interesting conclusions which we relate to our dataset. First, Edwards et al. recognize it may be rational for a victim state to tolerate attacks rather than risk escalation through blaming (i.e., attribution), especially when attacks are mild, and no appropriate response is available. Citing the case of Chinese-sponsored economic espionage against US industry, they note the US government’s inability to respond with in-kind attacks and refusal to blame China publicly given the importance of the countries’ broader relationship, instead pursuing diplomacy resulting in the US–China 2015 cyber agreement (US White House 2015). While this strategy apparently worked initially, analysis suggests the underlying intergovernmental negotiation has been unsuccessful in stemming China’s PLA-backed espionage (Segal et al. 2018).

Moreover, the USG did eventually also file an indictment, publicly attributing espionage activity to individuals affiliated with China’s PLA (US Department of Justice 2014). So, it may be more precise to describe the strategy as one that evolves over time. Tolerance of attacks in the near-term may be explained by Edwards et al.’s logic, but restraint allows the opportunity for sufficient evidence to be marshaled. Perhaps more importantly, the substantially higher number of attributions made by private actors to state sponsors observed in our data suggests that the extent of state’s use of the strategy may be dramatically understated. States may be knowingly refraining from blaming other states far more often and for more reasons than are evident.

Second, Edwards et al. “somewhat surprisingly” conclude that it is rarely beneficial for a victim to increase its own attribution capability. Why? They suggest that a non-vulnerable state will attack regardless of the victim’s capability to blame. And if an attacking state is vulnerable, a knowledgeable victim’s confidence in its ability to accurately attribute an attack will increase its incentive to counterattack. In both cases, the equilibrium outcome is unstable (i.e., attack, no blame). To the contrary, they argue the likelihood of stable equilibrium(s) (i.e., no attack; attack, blame) increases if both attackers and victims become knowledgeable through improved symmetric technical attribution capabilities. As explained below, the data illustrate the limited usefulness of individual attribution capability, and how collectively determined attribution methods and outcomes have evolved and arguably encourage restraint but also suffer from shortcomings as currently conceived.

A clear example of the former is Ukraine's public attribution of numerous incidents to Russia. Subject to so many attacks allegedly from Russia that Ukrainian officials have called their country "Russia's cyber-attack testing ground". The Ukrainian government "has managed to directly link Russia to most cyberattacks, citing the characteristics of the attacks and their timing; many occur on historically significant dates in Ukraine, or just before or during holidays, thus maximizing the effect" (Miller 2018). But, despite condemning Russia publicly for the alleged attacks in addition to substantial financial and other support from the United States and NATO to bolster its cyber security, attacks have persisted.

Another example is the Democratic National Committee incident. In December 2016, the White House and US Dept. of Treasury separately leveled sanctions against five Russian entities (including two Russian intelligence organizations) and six Russian individuals in response to attacks on the Democratic National Committee (Federal Register 2017). They were based in part on a technical report issued by the Dept. of Homeland Security and Federal Bureau of Investigation that provided many already reported indicators of compromise (e.g., IP addresses, domain names), as well as classified USG intelligence information (US Department of Homeland Security 2016; Office Director of National Intelligence 2017). There was no detail supporting attribution in the White House statement, and the DHS/FBI report was criticized by a former NSA security expert for failing to provide any evidence of attribution (Lee 2016). In short, the veracity of the attribution suffered, given the absence of publicly available evidence (or explicit linkages to evidence which had already been published by a threat intelligence company). Moreover, while the sanction and indictment processes clearly attributed alleged activities to individuals and organizations, questions remain as to their enforceability and effectiveness as a deterrent.

To the contrary, the WannaCry and NotPetya incidents were followed by attribution efforts coordinated between multiple allied states, and seemingly to a lesser degree, private actors. The coordination of public attribution among the states took place through various means. States that support other states' attribution results have been mainly subjected to the same cyberattack or are allies of the attributing and attacked countries.⁶ Some states that have supported US attribution announcements clarify that they have done their own investigation. For example, in the case of WannaCry, New Zealand endorsed the US claims of attribution to the North Korean government, while relying on its own evidence.

The United Kingdom also assessed that WannaCry was carried out by North Korea, not directly mentioning the assessment of the United States but saying: "we are committed to strengthening coordinated international efforts to uphold a free, open, peaceful and secure cyberspace" (UK Foreign Office 2017). Multiple states also supported attribution of NotPetya to Russia, including Australia, Estonia, Ukraine, the United Kingdom, Denmark, Lithuania, Japan, and Canada. Again, the United Kingdom carried out its own investigations and condemned the attack (UK Foreign Office 2018). Canada was not attacked by NotPetya, but condemned the attack to show its support for other allies. Australia declared that

Russia was behind NotPetya, based on advice from its own intelligence agencies and consultations with the United States and the United Kingdom.⁷

The apparent success of these efforts was institutionalized, with a ministerial and communique expressing that the states “would coordinate on appropriate responses and attribution” (Department of Homeland Affairs Australian Government 2018). However, this agreement was only among the Five Eyes, which raises the question whether or not collective attributions made by those states will be accepted more broadly. Moreover, this initiative focuses on “coordinating technical attribution and operational response policies to mitigate significant cyber incidents” and does not discuss the attribution process.

Not all states are willing to participate in collective public attribution. Germany, one of the most affected countries by NotPetya, surprisingly did not join the collective action of states condemning Russia. Some relate Germany’s inaction to its close ties to Russia or its lack of a capability to coordinate a response (Koch 2018). But it was clear that Germany was not willing to publicly attribute the attack. The European Union has also followed a similar approach and does not engage with public attribution. In response to a question from the Council of the European Union as to why it has not joined its allies to publicly attribute NotPetya to Russia, the Council said:

In its conclusions on malicious cyber activities of 16 April 2018, the Council expressed the EU’s serious concern about the increased ability and willingness of third states and non-state actors to pursue their objectives by undertaking malicious cyber activities, [...] It is not for the Council to comment on national governments’ decisions, based on all-source intelligence, to publicly attribute cyber-attacks to a state actor.

(Council of European Union 2018)

The European Union also in the conclusions on malicious cyber activities emphasized the importance of cyber norms.

Institutionalizing transnational attribution

Both technological developments and better understanding of how states act strategically highlight the need for institutionalizing neutral, transnational attribution. At some point, the evidence has to be assessed and independently reviewed, and that cannot be carried out through technological means alone. A decision to blame a responsible party has to take place through a recognized collective attribution process. Such a process has not been implemented, nor have current processes been studied in detail.

Proposals for institutionalizing transnational attribution

A transnational attribution institution could serve as a neutral global platform to evaluate and perform authoritative public attributions. It would be an independent

entity or set of processes whose attribution decisions would aspire to be widely perceived as *unbiased*, *legitimate*, and *valid*, even among parties who might be antagonistic (such as rival nation-states). Various proposals have been put forward with different scopes of activity, organizational structures, levels of stakeholder involvement, and evidentiary standards to potentially achieve such a process. Four of the leading attribution proposals have markedly different descriptions. Microsoft describes their proposal as “a public-private forum to address attribution” (Charney 2016); the Atlantic Council called for a multilateral “attribution and adjudication council for cyber-attacks rising to the [legal] level of ‘armed conflict’” (Healy et al. 2014); a RAND study called for a “Global Cyber Attribution Consortium” of non-state actors (Davies et al. 2017); a Russian think tank called for an “independent, international cyber court or arbitration method that deals only with government-level cyber conflicts” (Chernenko et al. 2018). A more recent initiative builds on two of these proposals.

The International Attribution Organization proposed is one such proposal that has been widely touted in the Microsoft Digital Geneva Convention, and in its subsequent articulation (see Charney 2016, also Charney et al. 2016). This proposal included language that suggested that an independent attribution organization should (1) span the public and private sectors while including civil society and academia, (2) both investigate and serve an information sharing role, and (3) resemble the International Atomic Energy Agency (IAEA). The initial proposal contained significant ambiguity as to whether or not this is describing a multi-stakeholder or multilateral model.

The Atlantic Council’s 2014 *Confidence Building Measures in Cyberspace* report proposes a multilateral “attribution and adjudication council for cyberattacks rising to the [legal] level of ‘armed conflict’” (Healy et al. 2014). While the scope is only limited to incidents that rise above an international legal threshold, Healey et al. suggest that these assessments should result in the application of an enforcement mechanism. The organization, like the Digital Geneva Convention draws on the IAEA for inspiration, but also the Biological Weapons Convention and Nuclear Nonproliferation Treaty.

RAND’s Stateless Attribution Report draws on both the Atlantic Council’s and Microsoft’s work, but suggests that “an attribution organization should be managed and operated independently from states”. Their report also differs from the Atlantic Council report in suggesting that an enforcement role is not needed. While the RAND Report classifies the Atlantic Council proposal as including non-state actors in collaborative investigations, this seems to confuse organizational management and support. As the Atlantic Council’s proposal makes use of private sector data and expertise as a multilateral entity, the RAND proposal does not explain how non-state actors would assist targeted states without their involvement.

The work by Chernenko et al. paper presents an interesting contrast to the IAEA model for attribution. While not denying the significance of private sector actors, the Chernenko et al. proposal is explicitly state based, recommending an “independent, international cyber court ... that deals only with government-level

cyber conflicts” (Chernenko et al. 2018). This scoping is less expansive than the Microsoft proposal, but more inclusive than the Atlantic Council’s, covering government-level cyber conflict which would include those below the threshold of armed conflict.

Each proposal offers different scopes of activity for a cyber attribution organization and pushes for dramatically different structures (e.g., multilateral vs. nongovernmental, or hierarchical vs. networked). And while the RAND Report makes powerful arguments as to why states have conflicting incentives to participate in an attribution organization and cautions against their membership in any Consortium, none of the above proposals explicitly consider the incentives for private actors to participate in the forensic process. The authors are tracking the aforementioned proposals and critiquing their viability, but believe more research is needed before a consensus can form.

Over the past two years, a handful of organizations, including the Internet Governance Project (with which the authors are affiliated) and Swiss-based ICT4Peace, have built upon the ideas presented in both Microsoft and RAND proposals. After socializing the idea of transnational, independent cyber attribution in fora like RightsCon, the UN Internet Governance Forum and the North American Network Operators Group, an initial workshop bringing together university-based and independent researchers took place in May 2020 (Internet Governance Project 2018). Together the workshop participants continue to develop a global network of researchers based in academia, civil society, and business who want to cooperate to develop attribution capabilities that are considered scientific and credible by the broader community. If successful, this could effectively counter state-sponsored or state-affiliated cyberattacks and the strategic use of attribution, and complement other efforts like the CyberPeace Institute (<https://cyberpeaceinstitute.org/>) to build and enforce cyber norms through accountability.

Challenges of collective action in attribution

Three major challenges are likely to present themselves in institutionalizing transnational attribution; these include geopolitical conflict, building independent capability, and private sector participation. These challenges overlap with, but are more institutional than, the challenges identified by the RAND study: Effective attribution and persuasive communication. Efficacy and communication will be contingent on the breadth of participation of public and private entities and their willingness to be transparent with the evidence. As with any political challenge, getting collective action from actors with competing interests presents a challenge.

Adversarial geopolitical relationships are likely to extend to any attribution organization. The advantage of such an organization is that by joining it participants agree to adhere to the constitutive as well as procedural rules, even when they disagree over the particulars. Neutrality of international bodies is often established through the professionalism of participants: Either a technical independence as described in the RAND study or a judicial independence might claim to

embody this ethos. Should states as political actors be involved, as described by the Atlantic Council proposal, a majoritarian ethos might be needed to result in collective action. A consensus-based solution proposed in the Microsoft Digital Geneva Convention research could certainly face challenges acquiring unanimity.

In addition to the geopolitical challenges of managing an organization are those of creating trustworthy assessments. The Organization for the Prohibition of Chemical Weapons (OPCW) manages to maintain global trust in its forensics with an independent laboratory, whose work it supplements with a network of over 20 certified laboratories distributed across numerous national jurisdictions. While the same strategy might help to supplement the capability of an attribution-based organization, building this capability will require financial resources. Finding dedicated financial resources for transnational attribution might create its own challenges. Would a government finance an organization tasked with rooting out its espionage operations, what incentives are there for the private sector, particularly those who sell services to multiple governments?

The cyberspace domain is uniquely defined by private sector participation and ownership of the core infrastructure. In this respect, Microsoft's Digital Geneva Convention is served well by including the private sector but creates a potential contradiction by drawing on the example of the International Atomic Energy Agency. It is possible to imagine an independent, member state-funded international organization, like that of the IAEA. Or by empowering the private sector, academia, and civil society is Microsoft suggesting a multistakeholder model? At face value, it appears that governments will set the rules, while private actors will lend their services and data, but nothing is stated about how these interests might be aligned. If a subset of private sector cyber security firms has advanced forensic capability equaling or exceeding that of most states, why would they participate in a monopsony attribution organization? Presumably, benefits to them would need to outweigh costs. Alternatively, if access to the internet's infrastructure allows an investigation to backtrack the origins of an attacker, what process should enable the acquisition of relevant evidence? Should this layer of attribution include partnerships with national law enforcement or permit international inspections? Either way, this potentially burdens the private sector and has implications for global privacy.

Conclusion

This chapter has briefly described the state of play in cyber attribution and number of competing visions for its future. At present, threat intelligence firms and national security agencies are the primary producers of forensic data and attributions. While reliance on algorithms to cluster observed data and identify infrastructure and adversaries is advancing this introduces socio-technical uncertainties around how data is collected, shared, and analyzed. Coupled with political fragmentation and strategic behavior by states there is need to focus on the institutionalization of credible, independent attribution. While ideal models for making attribution were described, too little is known about the current state of affairs.

As Edwards et al. (2017) suggest, understanding behavior in attribution clearly needs to incorporate the role and incentives of private actors. A research agenda going forward should attempt to better understand the practice of attribution and provide novel institutional designs and processes grounded in scientific method that go beyond merely replicating international organization approaches in other fields. To achieve this further exploration is needed around research questions like:

- How does the public and state response to attribution differ based on whether the forensic assessment comes from the private sector, state intelligence, law enforcement, or secondhand media reporting?
- How can scientific concepts and practices of empirical data and methodological transparency, reproducibility, and falsifiability be incorporated into and improve the practice of attribution?
- What data and methods are used in attribution to threat actors and ultimately to responsible parties?
 - When it comes to findings, are there different accepted levels of confidence?
 - How do geopolitical rivalries undermine the confidence placed in attribution?
- Is a hierarchically organized institution really needed to align participant incentives, or can a more loosely organized form of networked governance suffice?
- How would different visions for attribution address the concerns and incentives of stakeholders, distribute costs, and get off the ground?

Future work should continue to seek a better understanding of how governance models, including an independent network of researchers based in academia, civil society, and business might help resolve the issues flagged above so that responsible parties can be held accountable. Despite the capacity of advanced and persistent threat actors, the need to protect intelligence sources and methods, and conflicting nationalistic approaches we believe that movement toward transnational, independent, credible attributions to “who did it” is possible.

Notes

- 1 This chapter is based in part on an earlier work by the authors, “Cyber Attribution: Can a New Institution Achieve Transnational Credibility?”, *Cyber Defense Review*, 4(1), 2019.
- 2 The Council on Foreign Relations is not the only entity collecting and publishing cyber incident data. Another example is the Dyadic Cyber Incident and Dispute Dataset by Valeriano and Maness (2015), as well as incident data collected by the New America Foundation. Methodological questions can be raised where differences occur between these datasets, e.g., in what is considered a state-sponsored “incident”, or an attribution to a specific perpetrator.
- 3 For example, we have linked technical reports published for each incident from the APT Notes repository available at <https://github.com/aptnotes/data/blob/master/APTnotes.csv>

- 4 Other countries make statements and announcements through their cyber security centers or foreign ministries. See, e.g., Canada's announcement condemning NotPetya (Communications Security Establishment 2018) and UK's Foreign Office announcement on WannaCry (UK Foreign Office 2017). One factor that might explain the difference in approach between the United States and the rest of the world is that the United States is mainly the main target of the attacks while others might indirectly suffer.
- 5 A game in which the players have incomplete information on the other players (e.g. on their available strategies or payoffs), but, they have beliefs with known probability distribution.
- 6 For example, Estonia and Canada condemned cyberattacks that did not harm their countries; see Estonian Ministry of Foreign Affairs (2018).
- 7 "Based on advice from Australian intelligence agencies, and through consultation with the United States and United Kingdom, the Australian Government has judged that Russian state sponsored actors were responsible for the incident". Statement by Minister of Law Enforcement and Cyber security (no longer available online).

References

All links checked on August 23, 2021.

- Burgess, M. (2017, May 7). WikiLeaks Drops 'Grasshopper' Documents, Part Four of Its CIA Vault 7 Files. *Wired*. Retrieved from: <https://www.wired.co.uk/article/cia-files-wikileaks-vault-7>.
- Caltagirone, S., Pendergast, A., and Betz, C. (2013). The Diamond Model of Intrusion Analysis. Working Paper. Retrieved from: <https://www.activeresponse.org/wp-content/uploads/2013/07/diamond.pdf>.
- Charney, S. (2016, June 23). Cyber Security Norms for Nation-States and the Global ICT Industry. *Microsoft on the Issues* (Blog). Retrieved from: <https://blogs.microsoft.com/on-the-issues/2016/06/23/cyber-security-norms-nation-states-global-ict-industry/>.
- Charney, S., English, E., Kleiner, A., Malisevic, N., McKay, A., Neutze, J., and Nicholas, P. (2016). From Articulation to Implementation: Enabling Progress on Cyber Security Norms. Microsoft Corporation Working Paper, June. Retrieved from: <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/REVMc8>.
- Chernenko, E., Demidov, O., and Lukyanov, F. (2018, February 23). Increasing International Cooperation in Cyber Security and Adapting Cyber Norms. Council on Foreign Relations (Blog). Retrieved from: <https://www.cfr.org/report/increasing-international-cooperation-cybersecurity-and-adapting-cyber-norms>.
- Cloppert, M. (2010, June 21). Security Intelligence, Defining APT Campaigns. *SANS* (Blog). Retrieved from: <https://www.sans.org/blog/security-intelligence-defining-apt-campaigns/>.
- Communications Security Establishment [CSE]. (2018, February 15). CSE Statement on NotPetya Malware. Retrieved from: <https://www.cse-cst.gc.ca/en/media/2018-02-15>.
- Council of European Union. (2018, April 16). Council Conclusions on Malicious Cyber Activities – Approval. Retrieved from: <https://data.consilium.europa.eu/doc/document/ST-7925-2018-INIT/en/pdf>.
- Davis II, J. S., Boudreaux, B., Welburn, J. W., Aguirre, J., Ogletree, C., McGovern, G., and Chase, M. S. (2017). *Stateless Attribution*. Santa Monica: RAND Corporation.
- Defense Advanced Research Project Agency [DARPA]. (2018). Department of Defense Fiscal Year (FY) 2018 Budget Estimates. Retrieved from: https://www.darpa.mil/attachments/DARPA_FY18_Presidents_Budget_Request.pdf.