Zeitschrift für Friedens- und Konfliktforschung (2023) 12:31–58
https://doi.org/10.1007/s42597-023-00099-7

AUFSATZ

Preventing the escalation of cyber conflicts: towards an approach to plausibly assure the non-involvement in a cyberattack

Thomas Reinhold · Christian Reuter

Received: 10 January 2022 / Revised: 3 March 2023 / Accepted: 10 March 2023 / Published online: 17 May 2023

Abstract While cyberspace has evolved into a commonly shared space vital to our individual lives and societies, malicious cyber activities by state actors as part of espionage operations, regarding defense strategies, or as part of traditional conflicts have strongly increased. In contrast, attributing the origin of such activities remains problematic. The ambiguity of digital data raises the problem of misinterpreting available information, increasing the risk of misinformed reactions and conflict escalation. In order to reduce this risk, this paper proposes a transparency system based on technologies which usually already exist for IT security measures that an accused actor in a specific incident can use to provide credible information which plausibly assures his non-involvement. The paper analyses the technical requirements, presents the technical concept and discusses the necessary adjustments to existing IT networks for its implementation. Intended as a measure for conflict de-escalation, the paper further discusses the limitations of this approach, especially with regard to technical limits as well as the political motivation and behavior of states.

Keywords Attribution · Cyberwar · Cyberpeace · Escalation prevention · Transparency · Trust building

✉ Thomas Reinhold · Christian Reuter
PEASEC—Science and Technology for Peace and Security, TU Darmstadt, Pankratiusstraße 2, 64289 Darmstadt, Germany
E-Mail: reinhold@peasec.tu-darmstadt.de

Christian Reuter
E-Mail: reuter@peasec.tu-darmstadt.de

Zur Verhinderung der Eskalation von Cyberkonflikten: Ein Ansatz zum plausiblen Nachweis der Nichtbeteiligung an einem Cyberangriff

Zusammenfassung Während sich der Cyberspace zu einer gemeinsam genutzten Domäne entwickelt hat, die für unser individuelles Leben und unsere Gesellschaften notwendig ist, haben zeitgleich auch böswillige Cyberaktivitäten staatlicher Akteure im Rahmen von Spionageoperationen, im Hinblick auf Verteidigungsstrategien oder im Kontext traditioneller Konflikte stark zugenommen. Gleichzeitig bleibt die Herkunftszuordnung (Attribution) solcher Aktivitäten weiterhin schwierig. Die Interpretation digitaler Spuren ist anfällig für Fehlinterpretationen und erhöht damit das Risiko von Konflikteskalation, insbesondere wenn andere Wege der Krisenkommunikation zwischen Gegnern fehlen. Um die Wahrscheinlichkeit derartiger Fehlreaktionen zu verringern, wird in diesem Papier ein Transparenzsystem vorgeschlagen, das auf Technologien basiert, die in der Regel bereits für IT-Sicherheitsmaßnahmen vorhanden sind, und mit Hilfe dessen ein, in einem bestimmten Vorfall beschuldigter Akteur glaubwürdige Informationen liefern kann um seine Nichtbeteiligung an dem Vorfall plausibel zu belegen. Das Papier analysiert die technischen Voraussetzungen, stellt das zugrundeliegende technische Konzept vor und diskutiert die für die Umsetzung notwendigen Anpassungen an bestehende IT-Netzwerke. Als Maßnahme zur Deeskalation von Konflikten gedacht, werden auch die Grenzen dieses Ansatzes diskutiert, insbesondere im Hinblick auf technische Möglichkeiten sowie die politische Motivation und das Verhalten von Staaten.

Schlüsselwörter Attribution · Cyberkrieg · Cyberfrieden · Eskalationsprävention · Transparenz · Vertrauensbildung

1 Introduction

Disruptive cyber operations, influencing campaigns and espionage are increasingly a daily occurrence in this domain. This is underscored by the fact that states have included such operations in their domestic security and defense doctrines and understand this as an additional field of military operations (UNIDIR 2016). Offensive activities such as cyberattacks can be launched and performed indirectly or with counterfeited digital footprints, i.e. incriminating uninvolved, innocent actors (Warrell and Foy 2019). However, due to a complex and often time-consuming process that impedes effective counter activities, many have declared the secure identification of sources of malicious cyber activities to be impractical (UNIDIR 2018). These technical, political and organizational challenges of the so-called attribution problem increase the risk of misunderstanding, miscalculating and misinterpreting malicious cyber activities. The precise role and influence of military cyber activities in open conflicts is still ambiguous. However, particularly the increasing attacks against critical infrastructures (Noguchi and Ueda 2017; Lunden et al. 2021), the fear of interference by foreign states in vital public services in peace times, or situations where immediate responses are necessary to counter and mitigate cyber threats could lead to misguided responses and create a momentum for the escalation of

conflicts. Although mostly considered a last resort, some states have even reserved the right to respond to cyberattacks with physical military means if deemed necessary as an immediate defense against hazards. This scenario and its increasing probability have recently been expressed, e.g., by US President Biden, with regard to the ongoing cyber tension between the US and other major powers (The White House 2021). These problems may be exacerbated if other ways of crisis communication or security and trust-building measures between adversaries are missing. Whereas the history of technical peace research has addressed these issues for other military technologies with conflict prevention and de-escalation measures, suitable approaches for peacekeeping in cyberspace are currently lacking and their development is strongly recommended (Wissenschaftsrat 2019, p. 60ff).

Against this background, this paper focuses on the research question of how technical measures can enable state actors to mutually control their cyberspace activities of military forces and intelligence services by providing verifiable data that can be used to assess and verify a state's non-involvement in a specific previous or ongoing cyber incident. This paper does not aim to provide a ready-made implementable measure but rather to provide a food-for-thought as well as a technical foundation for such an approach and to examine the possibilities for establishing such measures based on existing IT infrastructures. This paper further discusses the necessary technical adjustments required to provide this kind of information while maintaining an appropriate level of secrecy. Given the sensitive nature of internet traffic surveillance, the paper discusses the limits of such a measure with regard to its implementation in order to uphold human rights principles and avoid their violation. Concerning the aspect that cyberattacks are often carried out by hacker groups or other so-called proxies that are not directly associated with or under the control of state institutions, the paper also analyses the political preconditions and limitations. Regardless of the quite specific and small application scope and its requirements and restrictions, we hope to provide an impulse for cyber conflict de-escalation measures that help to mitigate the escalation risks inherent to the status quo. Finally, given the relatively new domain of cyberspace, the paper aims to connect computer sciences with peace and conflict research and hopes to provide valuable impulses for dialogue between them (Reuter 2019).

This paper is structured into the following sections: After the introduction and the definition of the research question in the first section, current research on the attribution problem based on related work from a technical as well as political perspective is discussed. Section 3 presents the cases that have been selected to illustrate this work's background and motivation. Section 4 explains the required technical principles of network-based digital data transfer and the problem of the ambiguity of digital data. Afterward, Sect. 5 discusses the motivation of states to join a risk reduction measure and comprises the central arguments of this paper with a conceptual as well as a technical outline for a system that can provide evidence to verify an accused state's non-involvement in a cyberattack. Finally, in Sect. 6, the developed measures, their limitation, and potential pitfalls are discussed in relation to the research question and their practical application. The section ends with an outlook on further extensions of this approach as well as future research opportunities.

2 Related work

The problem of attributing malicious cyber activities to their perpetrators has gained much attention in recent years. As attribution is both a technical and a political process, many different approaches have been proposed to solve the so-called attribution problem.

2.1 Technical challenges

A common argument for most approaches is that attribution is a complicated task, as information needs to be collected and interpreted forensically (Koch and Golling 2019). To support and improve this time and resource-consuming process, one area of research focuses on a more detailed and standardized implementation of data collection and storage based on frameworks (Lilly et al. 2019). This is also partly linked to advanced intrusion detection systems (Rubio et al. 2019), which are able to detect attack attempts for early data acquisition of the attacker's activities (Ni et al. 2016). Other investigations focus on the topology of IT networks and the question of where intrusion detection systems should best be placed, and how the processing of physical signals alone can indicate uncommon behavior and possible attacks (Giraldo et al. 2019), sometimes alongside methods of artificial intelligence and machine learning to detect intrusions and to identify the attacker's location (Siva Kumar et al. 2017). The approach aims to create complex data sets by including threat intelligence sharing platforms to gain a broader information basis (Perry et al. 2019). This allows for classifying and filtering the aggregated information on cyber incidents or applying data-analysis methods to detect similarities in attackers' behavior (Hoon et al. 2018; Shute et al. 2017). Other approaches focus on the inclusion of publicly available open-source information (Lemay et al. 2018) or social media (Bargar et al. 2019; Kumar and Carley 2016). In summary, current research on attribution as a measure of security concentrates on the tasks of data acquisition, optimization and analysis to provide a better and faster answer to the question: *Who did it?* In contrast, the perspective of this paper focuses on the question: *How can I credibly assure that I was not involved/that I am not the perpetrator?* This aspect is essential in order to improve security by preventing misattribution. This is a novel approach that has not been scientifically researched so far but lies in the tradition and demands of technical peace and conflict research (Reuter et al. 2020).

2.2 Political and security challenges

Attribution is and "should not be an aim in itself" (Broeders et al. 2020), as it pursues an intent and the "decision to attribute a cyber operation to another actor should be strictly linked to a broader policy objective(s) that a state or a group of states wishes to achieve" (Broeders et al. 2020). This broader objective has been discussed in different ways. First, attributing an attack towards a designated attacking country is a prerequisite for any legal military response in accordance with the UN Charter (Wingfield and Wingo 2021). In other words, without sufficient evidence, there is no target to refer to legitimately. Rowe (2015) focuses on this aspect and further

discusses (1) possible measures to achieve attribution, (2) the complexity of this task from a technical and legal standpoint, and (3) the influence this can have on an attacked state's political and military decision-making processes. In most cases, attribution requires cooperation between states in order to collect technical evidence of an attack, thus requiring a mutual understanding of the problem, legal common ground and suitable corresponding processes (UN-GGE 2015) for the collection as well as the exchange of threat intelligence (Riebe et al. 2019). This is discussed by Bendiek and Schulze (2021), with a focus on the development and enforcement of a harmonized cyber sanction regime of EU member states.

A further aspect of attribution is discussed regarding its applicability and limitations for deterrence and how "scaling of exploitation and retaliation costs lead to different degrees of coverage and effectiveness for deterrence by denial and punishment" (Lindsay 2015). A related aspect considers the role of public attribution—naming and shaming of the accused country—and the trust and accountability problem it faces regarding national interests (Egloff and Wenger 2019). A prominent approach to overcome this problem whilst fostering transparency in attribution processes is the establishment and empowerment of independent, supranational institutions that could be in charge of attribution based on evidence provided by attacked countries (Droz and Stauffacher 2018; Davis II et al. 2017). Other debates address the role of non-state actors behind cyberattacks, the responsibility and due diligence of states, and how attribution and the right to self-defense apply under these conditions (Starski 2015). Other research explores the question of whether or to what extent military cyber activities have shaped the battlefield and how this new military domain influences strategic or tactical military approaches in open conflicts (Kostyuk and Zhukov 2019). Blagden (2020) even argues that the technical challenges of revealing the concealment of attackers can be neglected, as an "attacker must necessarily reveal a set of interests that it values", putting the question of interests at the center and declaring attribution a primarily political task.

3 Case examples

The research question is motivated by real-world cyber incidents and the associated escalation risks. In order to illustrate them and the constraints of a measure focused on state actors, we have selected two examples of actual critical cyber-based conflicts that happened in the context of strong regional political and military tensions between actors that have developed strong offensive cyber capabilities over the last years. The examples are used to derive an exemplary model for the attribution process and to define requirements for a technical system that can provide plausible information that provides evidence and assurance for non-involvement in a specific cyber incident. The findings are put together into a model for a technical measure and its applicability. The security gain, as well as its limitations and necessary further extensions, are discussed both for the technical and the political context.

3.1 Selected case examples

The following two brief examples illustrate the different scenarios and variations of cyber-based incidents in the context of regional tensions and the problems of unclear or misleading attribution. They have been selected because they prototypically exemplify the ambiguity of digital data as well as the situational pressure for action, whereas the overall political situation suggests an obvious answer to the question of the origin of the attacks, thus increasing the risk of misinterpretation and the danger of misreaction. Sources about such incidents, the political considerations regarding countermeasures and the actual measures taken are often rare or incomplete—partly because the public (non-)communication towards the suspected attackers is often already part of the reaction. In most cases, the only available sources are press reports and other anecdotal public media coverage. Compared to other incidents, the selected case studies are based on credible sources that at least provide enough details to analyze the course of events.

3.1.1 The cyberattack against chemical plants in Saudi Arabia, August 2017

In August 2017, a cyberattack was detected at a petrochemical plant in Saudi Arabia that targeted the industrial control systems which monitor, control and regulate all different aspects of the industrial process (Perlroth and Krauss 2018). In contrast to former attacks against such systems, which have often tried to silently manipulate the controlled processes, the aim of the detected attack was presumably to deliberately sabotage the industrial hardware by triggering explosions. This would most likely have resulted in significant damage to the facility as well as possibly human injuries or casualties. This was prevented by a programming mistake made by the attackers. Investigators blamed Iran for the attack due to previous incidents against governmental institutions and industrial facilities (Tarakanov 2012) as well as overall political tensions between the two countries (Marcus 2019; Baezner 2019). However, cyber capabilities from other opposing nations, such as China (The State Council Information Office of the People's Republic of China 2019), Russia (Karaganov and Suslov 2019), or Israel (Dwyer and Silomon 2019) would also have been sufficient to conduct this attack. Press reports suggested that Saudi Arabia feared an immediate second attempt of sabotage after the failed first attempt. This required a quick decision on whether and how to respond. As official communication channels between the nations have been scarce since an attack on the Saudi Embassy in Tehran in 2016 (Saudi Gazette 2016), the situation highlights the dangerous scenario of the lack of significant evidence and political crisis communication channels.

3.1.2 The cyberattack against the Ukrainian power grid, December 2015

The second case is the cyberattack against five power supply companies in Ukraine that took place on December 23, 2015 (Zetter 2016). The attack itself targeted the control systems of the power plants and their supply infrastructure, as well as the companies' call-center services so that customers could no longer receive any information. In total, up to 230,000 people were cut off from electricity for one to

six hours. The attack occurred in the aftermath of the Russian occupation of Crimea as part of the ongoing crisis between Ukraine and the Russian Federation. The attack was immediately attributed to hackers from Russia based on the geolocation of the attackers' IP addresses. From a technical perspective, this is not itself valid or sufficient evidence and could have been a false track laid by third-party attackers, demonstrating the ambiguity of the available information. So far, it is not entirely clear if the cyberattack was a sabotage operation, a test run of its own capabilities, or a political demonstration of power. In any case, it was an attack on critical infrastructure and, thus, a violation of established rules of international humanitarian law (IHL).

3.2 Comparison of the example cases

The selected cases, with their different presumable motivation of sabotage and social disruption and insecurity, have in common that the involved actors are in a situation of high political tension and ongoing small-scale military conflicts or are on the brink of tensions erupting. Although the disputes have so far remained below the threshold of a declared war, tensions have increased due to the decline in communication relations and the loss of commonly shared fora as well as diverging interests. The available sources do not allow for any conclusions as to whether a technical incident analysis had been performed. Given the presented expense for a valid attribution and the short time available in all cases, it is highly unlikely that forensically robust results had been available. As a result, the attribution of cyber incidents was strongly influenced by the overall situation and previous incidents. The probability of blaming the *usual suspect* when other information is missing or incomplete is very high in such cases, increasing the risk of misattribution and escalation of the incidents into a full-fledged military confrontation even further.

4 Technical properties of the cyberspace and the ambiguity of digital data

The difficulties related to attribution in cyberspace and the risk of false attribution due to misinterpretation of the available information are based on some specific technical features of this domain and the way in which cyberattacks are performed. The following section will highlight these particularities as a prerequisite for the research question.

Cyberspace is a virtual domain by design that abstracts a space from a specific real existing geographic location. It consists of autonomous, self-contained networks that integrate and connect groups of different IT systems or sub-networks. The networks are connected via gateway servers at their point of contact (*borders*). To perform any kind of data transmission between two IT systems, it is necessary to identify both systems, which is done by using a technique named *Internet protocol addressing* or, in short, *IP addressing* (Scaglia 2007). Figure 1 presents a simplified model of such a data transfer and the involved IT systems. It is important to understand that an address of an IT system—in the following called *(A)*—is not inevitably

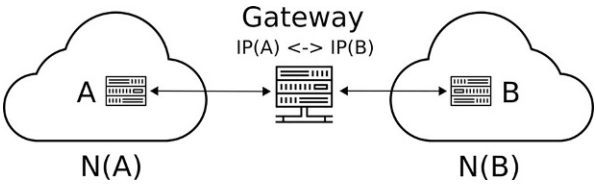

Fig. 1 Simplified model of data transfer between two computers in separate networks (*Source*: own illustration)

unique. It must only be distinct within the network to which the system is directly connected, hereafter referred to as *N(A)*. Any connection of *(A)* to an external IT system *(B)* that is not part of *N(A)* is transferred over the gateway server that—in the simplest case—directly connects the networks *N(A)* and *N(B)* but in actual scenarios involves multiple networks. The gateway server handles the necessary *network address translation* (NAT) to ensure that data can be transferred between two networks with possibly non-unique IP addresses (Juniper 2022). This means that the effective sender address that *(B)* can identify while receiving data from *(A)* is not a unique address of this IT system but rather an address provided by the gateway server of *N(A)*, which means that there is no clear and directly visible and re-traceable path to the origin of the connection. This aspect also entails that any kind of geographical localizing based on IP addresses will reveal the involved networks and not necessarily the specific IT system *(A)* itself. Over the last years, the current protocol—the technical rules of how network traffic is handled by the different IT devices—has been shifting towards a more modern approach (Internet Protocol version 6 or short IPv6 instead of IPv4) where devices have worldwide unique IP addresses (Juniper 2022). Nevertheless, for reasons of data protection or security, these unique addresses are often reduced to their network part, and the identification of a single IT system is taken over by gateway servers.

Another significant aspect of the technological basics of cyberspace is that it abstracts the process of data transmission between IT systems over different structural and conventional layers and generalizes specific functionalities with common technical protocols (Scaglia 2007). All IT systems that communicate over cyberspace have to use these protocols. There is no close connection between the observed usage of an IT system—like a cyberattack—and its real-world and intended purpose. Even a forensically *waterproof* identification of an attack's origin cannot exclude the possibility that an identified IT system had been taken over by adversaries. A popular but mere theoretical but conceptually valid example is the misuse of IT systems in a hospital that had been hacked to carry out cyberattacks.

A third important principle of cyberspace concerns the aspect that any data transmitted during connections between two distant IT systems *(A)* and *(B)* is split up into a large number of small packets that are sent separately and merged at their destination. Each transmitted packet can potentially take a different path, i.e., *route*. This principle guarantees that disruptions of a route can be balanced by other transmission paths. In the context of cyberattacks, this means that retracing the steps of attacks to their origin equals finding the path back via potentially different and numerous routes.

Drawing from these specific technical features, many real-world cyberattack scenarios involve multiple steps of intermediary hubs that are used to blur tracks. This

often involves the usage of one or more so-called *command and control* servers (C&C) that are used by attackers to coordinate progress and collect stolen data. C&Cs are either hijacked systems or rented servers that do not belong to the attackers themselves. Considering the explanation of IP addresses provided above, this means that even if a victim of a cyberattack can identify a unique IT system via its IP address, it is probably not the actual system of the attacker. Therefore, the task of attributing such attacks typically involves the analysis of at least some of the IT systems used as hubs and the C&C infrastructure. Aside from the time required to perform these actions, each step potentially relies on the cooperation of other actors to gather information from affected systems within their jurisdiction, as well as the availability of such data samples (Clark and Landau 2010).

These discussed features of cyberspace complicate the attribution and create a strong character of ambiguity. Available information on attacks and traces towards attackers is, in most real-world cases, either incomplete or inconclusive in terms of its interpretation. In addition, this information is easy to manipulate or counterfeit, and attackers might have created false tracks by forging misleading evidence, commonly described as *false flag operations* (Steffens 2020). On the other hand, cyberattacks against critical systems often require immediate decisions to be made about countermeasures to stop the attack's ongoing threat. In combination, the current lack of internationally binding norms for responsible state behavior in cyberspace leads to the situations described, in which misunderstandings, miscalculations, and misinterpretations could cause wrong and potentially destructive responses.

5 Reducing the escalation risk: outline of a system to plausibly assure non-involvement in a cyberattack

Based on the previous assessment, this section proposes a concept that, while it cannot help to diminish the *burden of proof* of the cyberattack victim, aims to help to reduce the threat of accidental escalation of a conflict. The concept is understood in the sense of the CSCE Helsinki final act that recognized "[...] the need to contribute to reducing the dangers of armed conflict and of misunderstanding or miscalculation of military activities which could give rise to apprehension, particularly in a situation where the participating States lack clear and timely information about the nature of such activities" (CSCE 1975, p. 10). The measures are based on the idea that a reduction in the escalation risk or even its prevention can be achieved if an accused actor is able to credibly and plausibly assure they were not involved in a specific cyber incident by providing verifiable information. Although it is ultimately the decision of an attacked state how to react, which is often determined by various aspects, political goals or political signals, such information could alter its judgment and assessment of the situation. By being as transparent as deemed necessary about current cyber activities within their military or intelligence services networks, accused states can further signal to other states that the risk of an imminent cyber threat is unlikely to exist.

5.1 Political incentives and motivation of states to establish and comply with such a measure

States as actors often have diverging interests and are—besides treaties or other binding commitments—able and sometimes willing to act in contradiction (e.g., by cheating) to their commitments. The resulting limitations will be analyzed and discussed in more detail in Chap. 6.1. Regarding this context, the main motivation of a state to establish measures that can plausibly assure their non-involvement in a specific cyber threat is their self-interest in preventing the escalation of a conflict or of being falsely held accountable for malicious activities. Being able to provide the described information and to make their own cyber activities transparent could even be a measure of confidence-building. In addition, the system will be completely operated by and, therefore, under the control of the establishing state or authorities that established the measure, and—in the case of an incident—only the state decides which information is disclosed. All this, of course, does not diminish the possibilities of carrying out malicious operations covertly nonetheless or of even being the actual perpetrator of the cyberattack in question (e.g., by using a proxy). The core element, in terms of the measure's political effects, is the credibility of an accused state. The de-escalation effect of the measure is directly linked to this credibility and, therefore, its other military or intelligence cyberspace activities and—in the best case—its refrain from using covert or proxy operations. In conclusion, this means that it is in a state's interest to create opportunities to de-escalate and avoid cyber conflicts and to maintain its credibility.

5.2 Requirements for a system for the plausible assurance of non-involvement

In order to assure non-involvement in a cyber incident, an actor—besides their political credibility—needs to supply verifiable data that fulfill the following requirements:

- *Relevance*: The information must contain all incoming and outgoing relevant network connections of the accused actor: (a) to or from all networks of the attacked actor, (b) to or from the IT systems that had been targeted and (c) about connections to networks or IT systems of third parties that are suspected of having been used as C&C infrastructure or any other kind of indirect attack controlling measures.
- *Sufficiency*: The above information must be supplied for the incoming and outgoing connections of a defined scale of networks that are under the accused actor's jurisdiction to a technical degree that allows plausible verification and ensures non-involvement.
- *Timeliness*: The information must be supplied in a timely manner for the entire period of the cyberattacks and/or the malicious activities.

The information could either be supplied voluntarily by an actor or as a response to a request by an accusing actor or entrusted authority. The provided information can be anonymized to the degree that assures non-involvement in a specific attack while filtering out irrelevant data or disguising secret information. This is possible,

e.g., by reducing the logged number of connections to such an extent that the subnets can be identified, but not the actual machine. For IPv4 as well as the newer protocol IPv6, this can be done by cutting the IP address of each logged system, IP (*A/B/...*), down to a subnet address. This would successfully hide the actual amount of different IT systems within this subnet as well as the individual connection information of each of these IT systems. On the other hand, it allows the mapping and comparison of all outgoing and incoming connections of this network that could be associated with a cyberattack that occurred at the same time, based on the logged timestamps, destinations and type of traffic.

Based on this information, either the accusing actor or a neutral third party would be able to assess the provided data. Instead of tracing back the path to the alleged attacker, the validation would be able to directly focus on the supposed origin of the attack path and therefore be able to validate a statement of non-involvement. As has already been pointed out, this will not reveal the identity of the actual attacker but can help to relieve the alleged attacker.

5.3 A conceptual cyber incident model as a potential use case scenario

Before discussing the technical implementation, the following section presents a schematic model of a cyber incident, as illustrated in Fig. 2. It is set in a crisis situation between two previously introduced actors—*(A)* and *(B)*—with an assumed strong degree of mistrust, negative expectations, and non-communication, as well as current political or military tensions. In this situation, actor *(B)* was the victim of a cyberattack. The example further involves the actual attacker actor *(C)* and the uninvolved but exploited actor *(D)*.

1. Actor *(B)* detects an incident, in the following referred to as (x), within its networks *N(B)*.
2. Entitled authorities of Actor *(B)* check the logged information as well as the technical integrity of the affected IT systems and detect unauthorized access to these systems from a source outside of *N(B)* over a time frame called T(x).
3. Actor *(B)* identifies the unauthorized access from an IP address IP(x) that is registered to a party within the jurisdiction of actor *(D)*. Actor *(D)* is assumed to be

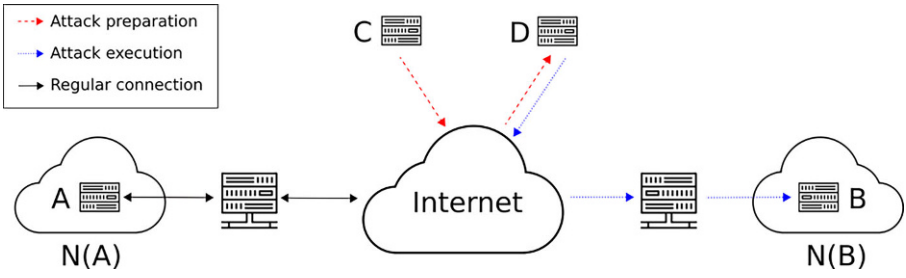

Fig. 2 Schematic model of common cyberattacks via an intermediary third party (*Source*: own illustration)

uninvolved in the current conflict and has no history of aggressive behavior against actor *(B)*.
4. Due to specific circumstances, the authorized agencies of actor *(B)* are not able to trace back the path from (x). Possible reasons for this situation, as discussed before, are:
   – The short reaction time that is available to decide on countermeasures by actor *(B)*.
   – A refusal of actor *(D)* to provide further information.
   – The absence of valid logging information either on the identified systems of actor *(D)* or on further intermediary steps.
5. Actor *(B)* accuses actor *(A)* of being the agent behind the incident (x) with reference to the political background, ongoing tensions, former incidents, or aggressive announcements by actor *(A),* or due to similar false-flag operations that had been tied to actor *(A)* in the past. To bring the harmful cyber activities to an end, actor *(B)* signals the willingness to use strong political or economic measures (the most likely reaction in espionage scenarios) or military force (the likely reaction in cyberattack scenarios).

In terms of this described scenario, the rephrased research question is: By which technical measures can Actor *(A)* credibly and plausibly assure that no IT systems within *N(A)* had any connection to the identified attacking system IP(x) for the time frame T(x) of the incident?

5.4 Necessary technical capabilities and specifications

The credibility of the provided information and thus the plausibility of the argumentation of a state's own non-involvement rests upon two technical properties of the data: (a) the tamper-proof collection and storage, and (b) the level of detail, the coverage and the verifiability of the information collected and provided by an accused actor. Therefore, based on the conceptual assessment of Sect. 5.2, this data and its collection measures need to fulfill the following technical specifications:

- Cover a time frame that is long enough to satisfy the accusing actor.
- Collect data from all relevant IT networks without hindering their functionality.
- Contain details on the endpoints of all connections that had been established and should rely on information that is always accessible during the network-based data transmission.
- Use a tamper-proof, non-circumstantial data acquisition and storage measure that is considered trustworthy even in non-cooperative actor relationships.

Notably, all IT network technologies gather and potentially store information on the established or performed network connections. This information analysis and storage already takes place in most cases as an IT security measure to monitor connections, identify malicious activities, detect or track hacking attacks and attempts (Gür 2015; Ekran Systems 2022), and control the access to IT systems and networks (so-called *firewalling*), and the measures we propose mostly require no changes apart from immutable storage. Under these prerequisites, the capability to

gather the necessary information is taken for granted and will not be described any further. In terms of the proposed context as a risk-reduction measure and with regard to the research question, this focus emerges into the following aspects:

- In which IT networks and where within a network does the data need to be collected?
- What kind of data needs to be stored and to which level of detail?
- For how long should the data be stored?
- How can tamper-proof storage be performed?

These questions will be discussed in detail in the following sections. Moreover, Chap. 6 will discuss the limitation of this approach and the possibility of state cheating.

5.4.1 In which IT networks and where within an IT network does data have to be gathered?

Connected IT systems are always topologically organized in network structures on the physical level, where the gateway servers between networks *know* about any outgoing and incoming connections for a specific network. With regard to the network-sub-network topology, it is only necessary to store the information about network activity at the logically *outermost* gateways, where data leaves the IT system of an actor and is transferred to external systems. In terms of the context of an application, only networks of institutions under direct state legislation or control and likely to be responsible for cyber activities in foreign IT networks, such as military networks or the IT systems and networks of intelligence services, need to implement the measure. While this has its limitations (see Chap. 6 for further discussion), it prevents the establishment of unlawful surveillance measures. Therefore, this storage needs to be performed on all gateways that connect these specific networks to the *outside world*—like private or commercial networks—to prevent *hidden channels*. The measure itself may need additional capacities for data storage but does not affect the functionality of the gateways nor the question of who runs them. As argued, in most cases, the relevant information has already been gathered and is ready to use.

5.4.2 What kind of data must be gathered and stored, and to which level of detail?

A typical connection between two IT systems consists of the exchange of multiple data packets with different purposes that establish the connection, transfer the data in multiple single packets, acknowledge the successful transmission of the packets, and finally close the connection. Besides the actual payload, each packet contains the so-called metadata of the information on the data packet sender and its destination—both identified via IP addresses. In terms of the proposed measure, the following information needs to be stored to ensure that for a given time slot, no data was transmitted to a specific IT system or network:

- The timestamp of when connections were established and closed, either from within the network or by request from external IT systems.

- The destinations (for outgoing connections) or origins (for incoming connections) of connections.
- The amount of data that has been transferred and, if available, for which kind of application the transferred data is meant to be processed.

For an effective application of the proposed measure, it is therefore not necessary to store the transferred packets but only the mentioned metadata on the connections. These kinds of data are simple text-based information that can be stored and transmitted without difficulty.[1] Although the actual logging rules for each implementing system are highly system-dependent, the above information should be gathered and stored for each connection that passes through the system without exception. In terms of secrecy, the stored information could potentially reveal sensitive data, as it contains details on the quantity and types of IT systems and services within the network, as well as the quantity and locations of systems that the specific gateway usually connects to. In order to maintain confidentiality, IP addresses can be partly anonymized to contain only information on the sender and destination networks, as it still contains sufficient information to assure non-involvement. This aspect also addresses upcoming IT security measures of *moving target defense* (MTD) (Carvalho and Ford 2014; Dishington et al. 2019), which aim to obfuscate the identity of IT systems within a network to confuse attackers by, for instance, randomly changing IP addresses.

5.4.3 How long does data need to be stored?

The question of the storage duration cannot be answered uniformly and is rather a task of considering a trade-off between essential storage resources, secrecy, and evidential value; notably, this parameter is easily adjustable. A solid basis for an estimated storage time can be provided by studies that are regularly performed by IT security companies that analyse hacking incidents. For example, a report by Mandiant Consulting (2016) estimated that in 2016, cyberattacks had lasted 146 days on the worldwide average before they were detected (Mandiant 2016). The same report calculated the average detection life span of hacking attacks for Europe and the Middle East to be up to 469 days. The analysts calculated a decreasing worldwide average of 99 days for 2017 and 56 days for 2019 and concluded that the life span of attacks significantly dropped due to higher sensitivity for IT security (Mandiant 2017, 2020).

Another approach to further specify the necessary logging time frame could be taken from recommendations (IBM 2019) for the size and time frame of logging data structures for IT security reasons. This system-specific assessment is performed to determine how far back in time a hacked target can retrace steps within its own systems via logged digital forensic information. Further inferences can be drawn from

[1] For illustration, storing such simple text-based information, one gigabyte of storage can hold up to 678,000 pages of text. (https://www.digitalwarroom.com/blog/how-many-pages-in-a-gigabyte). The King James Bible (Old and New Testaments), which is stored at Project Gutenberg in plaintext ASCII format (7 or 8 bytes per character) has a size if around 4.2 Megabytes.

national and international data retention policies to track criminal cyber activities (EGNYTE 2021). Furthermore, in the case of an ongoing attack, providing the current gateway activities of an accused actor can be an important measure to provide evidence for their non-involvement in the communication of the actual attacker with their command-and-control infrastructures. In the optimistic circumstances that our proposed measure is established as part of a treaty, the logging time frame should be commonly agreed upon and defined for all treaty members, as any past activity that is older than the time frame will not be reflected in the logs.

5.4.4 How can the process of data gathering and storage be technically tamper-proof?

The acquisition and storage of logging information is not new and is a common feature of IT security measures. In the context of this proposal, it is the credibility of such data that decides whether a targeted victim believes the *digital facts* that an accused actor provides. Credibility can be reached by technically ensuring that neither the process of the logging data acquisition has been tampered with (for example, connections to some specific endpoints get excluded from logging) nor that logged information can be manipulated afterward. Preventing and ensuring tamper-proof data storage is an issue that can be solved by cryptographically and incrementally signing the logged data, a method that is known as *immutable data storage* (Rovnyagin et al. 2021). This kind of technical verification for streams of logging data is a concept that has already been described as an *audit log* or *audit trail* (Schneier and Kelsey 1998, p. 1ff) for use cases in safety or secrecy critical scenarios (Putz et al. 2019). An additional degree of credibility can be achieved by ensuring that the mechanism which collects the logging information (commonly defined by so-called logging rules) itself has not been modified. This is possible by including the logging rules or hashes (*digital fingerprints*) of the logging code as part of the cryptographically secured logged data, as it provides tamper-proof copies of the logging process and its configuration for a retroactive validity comparison.

Creating and securing logging data with an immutable storage mechanism results in an increase in necessary processing and storage capacities as the information has to be encrypted and digitally signed. However, recent developments show that this can be done highly efficiently by using the GPU (graphics processing unit), a commonly used component in modern computers and that the storage scales 1:1 with the stored information items without creating overhead.

The actual storage capacities that are necessary for this proposed measure are strongly influenced by the actual network usage, the topology of the network and the storage time frame. Nevertheless, this is mitigated by the fact that the proposed measure only requires the storage of simple textual information, which is usually only a few hundred bytes, as opposed to complex binary information, such as images. Furthermore, the amount of storage required is reduced to the defined time frame of data storage and can be further cut down by stripping the network transaction information down to only necessary information in the sense of its meaningfulness, as discussed earlier. Even if a complex network usually has multiple sub-networks and the storage is therefore required for more than one gateway server, this is

quite negligible given the current prices of storage devices and the speed of their increasing capacity development (Coughlin 2020). The storage itself does not affect the functionality of the systems.

5.5 Outlines for a system of plausible assurance of non-involvement

The previous conceptual and technical requirements analysis, as well as the analysis of already existing technical capabilities of IT systems and their network components, show that a system for the plausible assurance of non-involvement in a cyber incident could be established based on already existing IT networking components. As these provide all the necessary information and nearly all necessary tools, the following aspects need to be taken into account during implementation:

- Ensure that information about network connections is collected (discussed in Sect. 5.4.2) and gathered on all relevant network and subnetwork gateway servers (discussed in Sect. 5.4.1).
- The existing logging system is—if not already in place—extended by storage of these logs for a sufficient amount of time (discussed in Sect. 5.4.3).
- The storage is performed tamper-proof, for instance, via immutable data storage technology (discussed in Sect. 5.4.4) and keeps the balance between the required secrecy of the actor collecting information as well as necessary details for a potential accusing actor (discussed in Sect. 5.4.2).
- The storage remains under the control of the establishing actor and is managed by the IT service personnel that already runs the IT security measures of the affected IT networks.

Regarding the cyber incident model presented earlier (see Chap. 5.3) as well as the model of common cyberattacks (as outlined in Fig. 2), the de-escalation measure could work as follows:

- After the entitled authorities of Actor *(B)* detected and analyzed the incident coming from an IT system within the networks of actor *(D)*, they request information about the gateways of this specific network $N(D)$.
- If actor *(D)* is cooperating and has the logging mechanism in place, it detects unlawful access coming from within the networks of actor *(A), w*hich can be linked to the cyberattack activities against actor (*B*). It provides this information to actor *(B).*
- As actor *(B)* suspects actor *(A)* to be the origin of the attack, it requests information from actor *(A)* about all connections for the time frame of the attacks on $N(B)$ for all gateway servers of $N(A)$. This suspicion is usually based on overall political circumstances, bilateral political tension, recent or former events and also influenced by political considerations of the accusing state (*B*) (Broeders et al. 2020). If there are multiple suspected actors, these steps would need to be taken with them accordingly.

- Based on the information provided by actor *(A)*[2] and potentially supplemented by additional information from actor *(D)*, actor *(B)* checks for connections from $N(A)$ to $N(D)$ that match the attacks in time and scale as seen from $N(D)$ to $N(B)$ to trace back the full attack path. This matching can be performed highly effectively by simple text search and comparison algorithms in no time to at least look for indicators of the presumption of actor *(B)*. There also exist tools capable of visualizing network connections[3] to help human analysts quickly get an impression of the situation. As actual cyberattacks require many network connections, an actual attack by actor *(A)* would be immediately visible in the data set. If no indicator is found to support this assumption, the information must be checked again more closely to rule out the possibility that connections have been overlooked.
- A de-escalation of the exemplarily assumed tension would be achieved if the data provided by actor (A) either contain no information about outgoing connections from its networks $N(A)$ directly towards the attacked system of actor (B) for the time of the attack and if the connections from $N(A)$ towards the neutral networks $N(D)$ cannot be linked to the attacks against $N(B)$.

The data collection of actor *(A)* is highly dependent on its IT network structure and the institutional organization that is in charge of operating the gateway server. However, as mentioned above, the information is usually already logged and is available in structured log files (usually one file per day per gateway) due to common IT security measures. This means they can be collected and presented immediately if the proposed measure is in place and the organizational structures are working. The data sharing between the actors can either be realized via secret communication channels between (A) and (B); be published if this supports the political signaling effect of the accused actor (A); or (B) could create public pressure that requires this step by publicly requesting the information from *(A)*. As network log data is only text-based information, log files are rarely larger than a few gigabytes (1 gigabyte can store 1 billion characters) and transmitted without any delay within minutes[4]. The provided data could either be validated and checked by actor *(B)* as demonstrated above, or by a neutral party whose conclusions would be accepted by both actors, such as a UN institution.

Beyond this conflict scenario, an attacked state could also decide not to go public with the incident, not blame another state, or request information. In these cases, our proposed measure would not provide any support. This said some experts argue that transparency in political relations is not always the best solution. However,

[2] It is important to mention that the requested information by actor (A) has to get transferred to actor (B). The time required for this is, as already mentioned, very dependent on the networks of actor (A) as well as on the transmission capacities and is an important factor despite the very effective data storage. This can be estimated in the case of a concrete implementation and on the basis of network information then available and taken into account in the political processes. Alternatively, however, direct access by actor (B) to actor (A) 's databases would also be technically possible. Even if this is associated with increased political costs, it could offer a way out in the extreme case of a danger that needs to be averted immediately.

[3] See e.g. this list of open-source tools for network monitoring and traffic visualizations, especially the tool "Nagios Core". https://www.comparitech.com/net-admin/open-source-network-monitoring-tools/.

[4] see footnote 3.

following this line of argument would go beyond the scope of this paper (Brown and Fazal 2021).

The above-presented outline is not to be understood as a ready-made blueprint for a measure to be implemented immediately. Rather, it is a theoretical concept. However, the concrete considerations depend to a large extent on the specific circumstances in the individual networks (e.g., which IT systems are used), including questions about technological developments (e.g., the storage space required and, therefore, the required additional IT hardware to store the logged information changes quickly, so that a certain fixed reference becomes obsolete very quickly (see e.g. Coughlin 2020)). Such considerations also depend on the political considerations of each implementing actor (e.g., for what time period logs are stored). Additionally, a real-world attack could involve many different proxies and an attacked state could suspect multiple states to be the origin. Nevertheless, the immediate attribution of attacks is largely a matter of political considerations and is based on previous events and situations of tension. However, even if further analysis of these parameters and their interconnection is valuable, this goes beyond the scope of this work.

6 Discussion and outlook

6.1 Limitations and potential pitfalls

The developed procedure faces some potential pitfalls that will be discussed in this section. The most relevant limitation is the measure's limited use case for interacting states and their necessary motivation to join and comply with the measure. Section 5.1 has already discussed this prerequisite, pointing out that a state's participation and compliance to the measure directly influence the state's credibility and thus, its possibilities to reduce the risks of the escalation of cyber conflicts. Regarding implementation, there are various ways to cheat, as will be discussed below. Nevertheless, the main motivation for states to prevail from cheating is the plausibility of the information provided and the effectiveness of the measure in the long term.

6.1.1 Use of proxies

In terms of the plausibility of the measure and the information it can provide, a particularly critical limitation is based on the use of proxies for actions of military or intelligence services in cyberspace. A proxy could either be IT systems within the state's jurisdiction that are not associated with the state-owned military or intelligence services and are not officially under its control. Russia, e.g., is known to use so-called *patriotic hackers*, which are officially civilian groups that are suspected of being under the unofficial influence and command of Russian national intelligence services. However, a proxy could also be one or more IT systems used to perform cyberattacks that are located in another state. This scenario is quite common and one of the previously discussed reasons for the risk of miscalculation. Unfortunately, there is no technical solution against states acting contrary to agreements and avoiding

attribution in this sense (Wingfield and Wingo 2021). To a certain degree, a strong civil society could reveal such behavior, and international intelligence services cooperation of states could help to uncover such operations by sharing information or—in the best case—the proxy system is controlled from IT systems where the proposed measures are in place and would leave traces in the logged information. Besides this, cyberattacks carried out by non-state actors without the direct or indirect involvement of state institutions cannot be mitigated by the proposed measure. This is an inherent limitation, as only IT networks directly under the control and management of military forces or intelligence services are to be logged. Any logging beyond this would pave the way for censorship or surveillance, which is neither desired nor intended by this paper. In the best-case scenario, a state commits to the due-diligence principle of state liability and prevents malicious activities that are performed under its jurisdiction. However, this goes beyond the scope of this work. Another aspect might be that a state uses the commitment to the proposed measures as a pretense to establish complete civilian surveillance. Apart from the fact that such a state would probably use any circumstances to justify such an undertaking (like fighting criminal or terrorist activities), this is neither desirable nor necessary or useful for the proposed application context.

6.1.2 Dishonest participation

A participating actor could decide to avoid attribution by not logging all relevant information, e.g., by leaving out incriminating connections. Although defining rules that omit some connections is highly system-dependent, this can easily be achieved either by excluding a specific range of IP addresses to not get logged or even specific for specific types of transferred data. In addition, an implementing state could also decide to leave out some gateways completely, which we have previously referred to as *hidden channels*. This could be mitigated to a certain degree by immutably logging the logging rules themselves, which can then be verified against transmitted information to check for inconsistencies by using special verified logging systems, such as the so-called *Trusted Execution Environment* (TEE) (Felton 2019). However, this kind of cheating cannot be completely prevented because as long as the logged information is under the control of a state, it is possible for it to cheat. On the other hand, this method of non-compliance can be detected when an attacked state sees network connections from the accused state in its own log files that did not show up in the logged information presented by the accused actor. This discrepancy between the provided information, on the one hand, and the information of the attacked state, on the other hand, can be revealed immediately with a simple 1:1 check of both data sets.

If an attacker uses proxies to perform the attack or if the attacker uses obfuscation measures to hide the attack path while using the proposed measures to signal non-involvement, this form of cheating is also not fail-safe. Though complex and time-consuming, the process of attribution and analysis of attacks in the aftermath can probably reveal the origin of an attack or at least hint to the performing state, which—if it contradicts the initially provided proof of non-involvement—will undermine the state's credibility. In any case, although this dishonest behavior may

benefit a state in the short term, as it seems to prove its alleged non-involvement in the incident, in the long term, it undermines plausibility and credibility and renders the measure and its de-escalating effect useless. Each implementing state must therefore decide which path to follow. Finally, it must be recognized that states and their institutions often have divergent, sometimes contradictory interests and that political decisions and intentions sometimes contradict concluded agreements. From a pragmatic perspective, although such behavior undermines the value of a specific attempt to de-escalate, it does not undermine the value of the proposed approach in general, given the tense situation in cyberspace, the high potential for misinterpretation and the necessity for peace-sustaining measures.

6.1.3 Limited feasibility and range of applicable networks

As discussed, the radius of a possible implementation is limited to specific networks like military and intelligence services networks to prevent the establishment of a surveillance system. This is not considered problematic in terms of the research question as the measures directly aim at the implementation by government institutions, which already have a special role that is usually associated with high responsibility and legal obligations. Although highly dependent on the political will, establishing the proposed measure within their networks is therefore considered applicable and legally indisputable if the political will to provide de-escalation measures is given. Furthermore, in most cases, military and intelligence operations follow specific orders, strategic and tactical planning, and have a chain of command and management of their activities. In particular, cyber operations are often the result of years of planning, building technological resources and personal know-how, and therefore do not occur isolated and *out of the blue*. Such long-running activities often use a continuously used channel to collect sensitive information, and the backdoor to the attack system is kept active, well hidden and up to date with the potentially evolving security measures in the target system. This highly increases the probability that at least some traces of network activities are logged as indications of harmful activities in the named systems for the time frame of the storage measure (as discussed in Sect. 5.4.3), even if the attack is carried out indirectly or via proxies.

6.1.4 Limitations of the logged information, secrecy and confidentiality

A double-edged aspect is an extent of collected, stored, and potentially committed information about network activities. On the one hand, more detailed logging has a higher informative value regarding the intended effects, but on the other hand, it might contain secret information that can prevent actors from establishing such measures. This can be diminished by anonymizing the stored information from a level that would allow identifying a unique IT system to a level where only the fact that the connections originate from the network would be logged. In addition, the information remains secret with the actor that deployed the measure until needed and is deliberately used only to prevent an imminent crisis with political tensions necessitating such means of de-escalation. Besides the already discussed relevance of a sufficient time frame of data storage and the importance of including all relevant

gateway servers that directly influence the credible argumentation, another limitation is given when cyber activities involve anonymization services, such as the TOR (an abbreviation for *The Onion Routing*) network (Tor Project 2019). Such services remove any information from packages that allow their assignment of an endpoint of a connection to its origin. Even if this effectively undermines the approach of attributing cyberattacks to perpetrators, the proposed measure can nonetheless provide plausible information that there were no connections between the networks of the accused actor and the servers of the anonymizing services during the specific time frame of the attacks.

Another limitation concerns the aspect that the data, which can show actor *(A)* was not involved in a specific attack against actor *(B)*, could at the same time contain critical information that would reveal cyber activities committed by *(A)* against another actor *(C)* that, until then, would not have been discovered. This could potentially discourage actors from implementing this measure but can partly be mitigated by the mentioned anonymization measures. In addition, the signatures (e.g., the amount and timing of network connections, as well as the extent and type of transferred data) differ between cyberattacks. This means that even if the provided data revealed ongoing cyber operations that are not part of the actual incident, the logged information and its characteristics could provide plausible information to argue in favor of the accused state.

6.1.5 Necessary technical and organizational adjustments

At last, it needs to be pointed out that the proposed measures need an adjustment of existing IT network infrastructures with an extension of the necessary processing and storage capabilities as well as the associated costs to sustain these capabilities. However, as argued above, such capabilities often already exist for IT security measures, thus limiting the need for complex IT infrastructural changes. Nevertheless, any additional hardware and software need maintenance and skilled personnel. As the storage could contain sensitive information, implementing the measure could also require the establishment of technical access control mechanisms alongside the organizational structures and permissions. With regard to the required storage capacities, it has already been argued that it highly depends on the actual IT network topology, size and activity. Nevertheless, as the proposed measure is thought first and foremost as a political advance and needs to be backed up by national legislation, it is worth pointing out that the actual technical requirements have not played a role in the past in similar legal initiatives such as data retention (EUFRA 2017) and were considered a necessity for IT providers to adapt.[5]. Section 6.3 will, nevertheless, discuss how simulations could help to estimate the technical dimension.

[5] A rough indication of the costs required by this measure is provided by an estimate of the German "Bundesnetzagentur" (Biermann 2015; Bundesnetzagentur 2015). In 2015, in the course of the discussion on the introduction of data retention, in which connection data should be retained for 10 weeks and location data for 4 weeks, they estimated the costs for telecommunications companies at approx. 100,000 euros for companies with up to 1000 telecommunications subscribers and up to 400,000 euros for companies with up to 30 million telecommunications subscribers. The estimates were based on surveys of German ICT companies.

6.2 Conclusion

The technical analysis presented above introduces a system of network connection logging and tamper-proof storage that enables an actor to provide network activity information that can plausibly assure their non-involvement in a given cyber incident to an accusing actor or a neutral intermediary. The analysis shows that, besides the political credibility of an accused actor, the keys for any line of argument depend on completely establishing the logging mechanism on all relevant gateway servers, on the time frame in which logging data is stored and kept, and on its tamper-proof acquisition and storage. The possibility to anonymize logged information to the network level without losing its evidential value provides the technical requirement of secrecy as well as making the measure ready for upcoming IT security measures like MTD. The conceptual outline shows that such systems can be built upon existing technologies and often already available IT hardware while including a few adjustments relating to storage capacities and maintaining the capacities to keep up with newer developments like the current shift of the IP. To create incentives for implementation, the measure of providing data allows different approaches that can disguise irrelevant and sensitive information. Despite the presented limitations, the proposed measure can provide a significant tool to circumvent the inherent problems of cyberattacks of missing data, their interpretation, miscalculations, and their attribution. With regard to the overall goal of arms control to provide tools for reducing the risk of armed conflicts, the measure can support this objective and provides a tool that allows overcoming the current insecure and unstable status quo.

In the examples of cyber incidents presented, it became clear that a state's perception of an imminent cyber threat and the need to respond to it can lead to escalation. The analysis of the attribution problems highlighted how the ambiguity of data could lead to miscalculations regarding the scope, the origin and the intention of a cyberattack—be it espionage that has gone wrong, sabotage, or an actual open, disruptive attack—and therefore lead to likely misresponses. This analysis also shows that the potential for escalations exists even when there has been no actual cyberattack but merely a fear of an imminent attack on critical systems or a perception of preparation for an attack, when there are no means for involved actors to provide information that can preclude this, e.g. by demonstrating that no cyber activities that usually precede such attacks have been taken. This fear has been fuelled by ongoing successful or attempted cyberattacks against critical infrastructures over the last years (Weinberg 2021) and the recent demonstration of some states not prevailing from such prohibited measures. Such political situations with smoldering conflicts, a high degree of mistrust, and political and military tensions exist, for instance, between Pakistan and India (Baezner 2018; Hess 2021). These tensions have been affecting regional civil communities in the border regions for decades, while political communication channels are scarce. For such situations, the proposed measure can potentially provide information that can be used to de-escalate a tense political situation by showing the absence of a cyber threat.

Concerning the context of this paper and the research question, the proposed measure can provide means to reduce the risk of conflict escalation in the aftermath or during ongoing cyber incidents due to misattribution or misinterpretation

of information, thus helping to refrain from the use of force (UNIDIR 2018). The measure is also capable of providing the required degree of transparency of current cyber activities within an actor's military networks to show that there is no imminent threat of malicious cyber activities. The necessary degree of commitment of involved actors recommends such measures for situations between actors with a high degree of political tension, where no other communication channels for crisis reduction exist, which relates to the commitments for stronger international cooperation in cyberspace (UN-GGE 2015). With regard to the consequences for personal rights and data privacy, the proposed measure should be limited to an application in highly critical scenarios only and only for potentially affected networks of military forces or intelligence services. Notably, the stakes for private, commercial and public IT systems, their protection and integrity and, respectively, for the communities that rely on them are high; and de-escalation measures are therefore strongly needed. Even if only established unilaterally, a state's transparency, credibility, and self-restriction of its own capabilities to conduct offensive measures in cyberspace could, as the analysis shows, be a valuable signal to overcome distrust (CSCE 1975, p. 3) between potential conflict parties. This does not reduce the offensive cyberspace capabilities of other states in question, but would—with the restriction of the discussed limitations—prevent the hidden, undetected use of a cyberattack in foreign IT systems as every connection would be logged. The proposed approach should help to foster the important task of preventing conflict escalations.

6.3 Outlook

Given the aspect that this paper presents a concept and outline but not an actual implementation proposal, a next step could provide such a proof-of-concept system. An actual implementation would have to be based on an analysis of the current military and intelligence service IT networks, their structures and technical characteristics. Such an analysis could therefore be used to evaluate the exact dimension of the technical parameters mentioned, such as storage duration, necessary additional storage space, etc. This would provide a basis to investigate the dependencies of the technical parameters, the necessary detailedness of the provided information as well as their impact on the intended plausibility presentation. In addition, future work should also perform simulations and calculations of how different network sizes and topologies, network traffic capacities and storage duration influence the necessary storage capacities and, thus, the necessary new infrastructures.

Following a proposal for international accountability in cyberspace (Davis II et al. 2017), the collected information of this measure could also directly help to strengthen the international credibility of the attribution processes carried out by a supranational institution under the UN regime (Droz and Stauffacher 2018; Davis II et al. 2017). Tamper-proof information that is collected in a standardized procedure could provide relevant contributions to this task. With a long-term perspective of arms control and arms regulation for cyberspace, the approach might also be implemented in safeguard agreements. Such treaties of international security could obligate member states to collect and share credible information among themselves that can be used for mutual compliance control if certain restricted cyber weapons have not been used against

each other or used at all. Besides the proposed implementation as a de-escalation measure, it can also be evaluated how the measure can be implemented between allied states in order to jointly prevent false-flag operations among this group of states and their IT systems and networks.

Further research could analyze whether and how traces of malware samples or logging information collected during cyberattacks from third parties could be used and compared against logging information that has been collected by the proposed measure and provided by a state to detect compliance when implemented as a safeguard measure. Such a comparison could offer additional tools to verify that an attack has been allegedly performed by an actor over the detected, accused third party and to further reduce the possibilities for *hidden attacks.* Additionally, the proposed approach could be extended to actors whose IT systems have been verifiably used for cyberattacks to plausibly argue that these had been performed by external hackers who misused the IT systems. This could provide a relevant forensic approach to bypass the current third-party-based hacking methods that are commonly used. Such third-party attacks often use public IT systems, this includes the challenge of how these IT systems could be integrated into the proposed de-escalation measures while preserving the integral principle of data protection, privacy and civil rights. Another issue could address the minimization of the proposed data storage, either in terms of reducing necessary resources or—more importantly—in terms of secrecy. This can be performed, for instance, by differentiating the storage of data connections into separate lists of addressed networks and connection metadata like application types. These lists could enable an accused party to provide precise data for specific incidents and prevent the handover of excessive, irrelevant, or potentially secret information. Finally, the measures could also be used to strengthen the development of digital trust and confidence-building measures as well as verification regimes that monitor and control the compliance of actors, e.g., towards a hypothetical non-use agreement of cyberweapons or an agreement to not attack specific critical infrastructures. In this context, it would be necessary to develop and establish practical control measures such as on-site or live inspections of gateway servers by neutral third parties in the sense of the safeguard agreements performed by the International Atomic Energy Agency (IAEA) to control the nuclear program of Iran or the verification regimes performed by the Organization for the Prohibition of Chemical Weapons (OPCW) under the Chemical Weapons Convention (CWC)—with the difference that inspections would affect military or intelligence service facilities in this case. Even the unilateral implementation of the proposed system, as a means of self-restraint from conducting offensive covert military or intelligence operations and, notably, as a means to signal transparency regarding one's own cyberspace activities, could establish a strong political signal of trustworthiness and political willingness. As cyberspace is increasingly becoming a domain of military power play, such signals are urgently needed.

Acknowledgements The paper is based on a previous (Reinhold 2018) work that has already been presented. The former publication (a) was intended as a policy brief for political decision-makers, (b) written and structured with a high degree of explaining technical terms, functions, and procedures and not based on scientific criteria (c) with a focus on presenting the technical concept for a practical application within the presented context of inter-state relations. This paper has been restructured and rewritten and extended

to (d) present, follow, and discuss an explicit research question, (e) present and argue along a chosen research methodology, (f) provide the current research perspective and related work for the technical as well as the political science perspective in order to underline the relevance of the research questions and its argumentation, (g) include new IT security developments, and within the conclusion to (h) highlight possible research topics as well as (i) related and possible fields of application of the proposed measure.

Funding This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE as well as the Deutsche Forschungsgemeinschaft (DFG)—SFB 1119—236615297.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Baezner, M. 2018. *Regional rivalry between India-Pakistan: tit-for-tat in cyberspace*. Zurich: Center for Security Studies (CSS), ETH Zürich.

Baezner, M. 2019. *Hotspot Analysis: Iranian cyber-activities in the context of regional rivalries and international tensions*. Zurich: Center for Security Studies (CSS), ETH Zürich.

Bargar, A., S. Pitts, J. Butkevics, and I. McCulloh. 2019. Challenges and opportunities to counter information operations through social network analysis and theory. In *2019 11th International Conference on Cyber Conflict (CyCon)*, 1–18. IEEE. https://doi.org/10.23919/CYCON.2019.8756832.

Bendiek, A., and M. Schulze. 2021. *Attribution als Herausforderung für EU-Cybersanktionen: Eine Analyse von WannaCry, NotPetya, Cloud Hopper, Bundestag-Hack, OVCW*. SWP-Studie 17. Berlin: Stiftung Wissenschaft und Politik.

Biermann, K. 2015. Vorratsdaten kosten mindestens 260 Mio. €. Zeit Online. https://www.zeit.de/digital/datenschutz/2015-10/vds-vorratsdatenspeicherung-millionen-kosten. Accessed 20 April 2023.

Blagden, D. 2020. Deterring cyber coercion: the exaggerated problem of attribution. *Survival: Global Politics and Strategy* 62(1):131–148. https://doi.org/10.1080/00396338.2020.1715072.

Broeders, D., E. De Busser, and P. Pawlak. 2020. Three tales of attribution in cyberspace: Criminal law, international law an policy debates. *EU CYBER DIRECT*. https://eucyberdirect.eu/research/three-tales-of-attribution-in-cyberspace. Accessed 7 Jan 2022.

Brown, J., and T. Fazal. 2021. #SorryNotSorry: Why states neither confirm nor deny responsibility for cyber operations. *European Journal of International Security* 6(4):401–417. https://doi.org/10.1017/eis.2021.18.

Bundesnetzagentur. 2015. Gesetzentwurf zur Speicherpflicht – Kostenschätzung. https://gruen-digital.de/wp-content/uploads/2015/10/BNetzA.pdf. Accessed 20 April 2023.

Carvalho, M., and R. Ford. 2014. Moving-target defenses for computer networks. *IEEE Security and Privacy* 12(2):73–76.

Clark, D.D., and S. Landau. 2010. The problem isn't attribution: it's multi-stage attacks. In *Proceedings of the re-architecting the Internet workshop*. ReARCH' 10., 1–6. New York: Association for Computing Machinery.

Coughlin, Tom. 2020. HDD market history and projections. Forbes. https://www.forbes.com/sites/tomcoughlin/2020/05/29/hdd-market-history-and-projections/?sh=7f2ddb986682. Accessed 20 April 2023.

CSCE. 1975. Conference on security and cooperation in Europe: final act. https://www.osce.org/helsinki-final-act. Accessed 7 Jan 2022.

Davis, J.S., II, B. Boudreaux, J.W. Welburn, J. Aguirre, C. Ogletree, G. McGovern, and M.S. Chase. 2017. Stateless attribution: toward international accountability in Cyberspace. RAND Corp Arlington VA United States. http://www.rand.org/pubs/research_reports/RR2081.html. Accessed 7 Jan 2022.

Dishington, C., D.P. Sharma, D.S. Kim, J.-H. Jin-Hee, T.J. Moore, and F.F. Nelson. 2019. Security and performance assessment of IP multiplexing moving target defence in software defined networks. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, 288–295. IEEE. https://doi.org/10.1109/TrustCom/BigDataSE.2019.00046.

Droz, S., and D. Stauffacher. 2018. *Trust and attribution in cyberspace: a proposal for an independent network of organisations engaging in attribution peer-review*. Cyber Security Policy Process Brief. Geneva: ICT4Peace Foundation.

Dwyer, A., and J. Silomon. 2019. Dangerous Gaming: Cyber-Attacks, Air-Strikes and Twitter. *E-International Relations*. https://www.e-ir.info/2019/09/23/dangerous-gaming-cyber-attacks-air-strikes-and-twitter/. Accessed 7 Jan 2022.

Egloff, F.J., and A. Wenger. 2019. Public attribution of cyber incidents. *CSS Analyses in Security Policy* 244:1–4.

EGNYTE. 2021. Data Retention 101: Policies and Best Practices. EGNYTE Blog. https://www.egnyte.com/guides/governance/data-retention. Accessed 20 April 2023.

Ekran Systems. 2022. 15 Cybersecurity best practices to prevent cyber attacks in 2022. https://www.ekransystem.com/en/blog/best-cyber-security-practices. Accessed 20 April 2023.

EUFRA. 2017. *Data retention across the EU*. European Agency for Fundamental Rights. https://fra.europa.eu/en/publication/2017/data-retention-across-eu#publication-tab-0.

Felton, Don. 2019. What is a Trusted Execution Environment (TEE)? Trustonic Blog. https://www.trustonic.com/technical-articles/what-is-a-trusted-execution-environment-tee/. Accessed 20 April 2023.

Giraldo, J., D. Urbina, A.A. Cardenas, and N.O. Tippenhauer. 2019. Hide and seek: an architecture for improving attack-visibility in industrial control systems. In *Applied cryptography and network security*. 17th International Conference, ACNS 2019, Proceedings., 175–195. https://doi.org/10.1007/978-3-030-21568-2_9.

Gür, G., Ş. Bahtiyar, and F. Alagöz. 2015. Security analysis of computer networks: key concepts and methodologies. In *Modeling and simulation of computer networks and systems*, Eds. Mohammad S. Obaidat, Petros Nicopolitidis, and Faouzi Zarai, 861–898. Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-800887-4.00030-4.

Hess, G.D. 2021. The impact of a regional nuclear conflict between India and Pakistan: two views. *Journal for Peace and Nuclear Disarmament* 4(S1):163–175. https://doi.org/10.1080/25751654.2021.1882772.

Hoon, K.S., K.C. Yeo, S. Azam, B. Shunmugam, and F. De Boer. 2018. Critical review of machine learning approaches to apply big data analytics in DDoS forensics. In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, 1–5. IEEE. https://doi.org/10.1109/ICCCI.2018.8441286.

IBM. 2019. Informix Servers: 12.10: Estimate the size and number of log files. https://www.ibm.com/docs/en/informix-servers/12.10?topic=files-estimate-size-number-log. Accessed 7 Jan 2022.

Juniper. 2022. Junos ® OS interfaces user guide for security devices. https://www.juniper.net/documentation/us/en/software/junos/interfaces-security-devices/interfaces-security-devices.pdf. Accessed 25 Aug 2022.

Karaganov, S.A., and D.V. Suslov. 2019. *The new understanding and ways to strengthen multilateral strategic stability. Report*. Moscow: Higher School of Economics, National Research University.

Koch, R., and M. Golling. 2019. Silent battles: towards unmasking hidden cyber attack. In *2019 11th International Conference on Cyber Conflict (CyCon)*, 1–20. IEEE. https://doi.org/10.23919/CYCON.2019.8757146.

Kostyuk, N., and Y.M. Zhukov. 2019. Invisible digital front: can cyber attacks shape battlefield events? *Journal of Conflict Resolution* 63(2):317–347. https://doi.org/10.1177/0022002717737138.

Kumar, S., and K.M. Carley. 2016. Understanding DdoS cyber-attacks using social media analytics. In *IEEE international conference on intelligence and security informatics: cybersecurity and big data, ISI 2016*, 231–236. Institute of Electrical and Electronics Engineers.

Kumar, Siva R.S., A. Wicker, and M. Swann. 2017. Practical machine learning for cloud intrusion detection challenges and the way forward. In *AISec 2017—Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2017*, 81–90. Association for Computing Machinery.

Lemay, A., J. Calvet, F. Menet, and J.M. Fernandez. 2018. Survey of publicly available reports on advanced persistent threat actors. *Computers & Security* 72:26–59. https://doi.org/10.1016/j.cose.2017.08.005.

Lilly, B., L. Ablon, Q.E. Hodgson, and A.S. Moore. 2019. Applying indications and warning frameworks to cyber incidents. In *2019 11th international conference on cyber conflict (Cycon)*, 1–21. IEEE. https://doi.org/10.23919/CYCON.2019.8756949.

Lindsay, J.R. 2015. Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack. *Journal of Cybersecurity* 1(1):53–67. https://doi.org/10.1093/cybsec/tyv003.

Lunden, K., D. Kapellmann Zafra, and N. Brubaker. 2021. Crimes of opportunity: increasing frequency of low sophistication operational technology compromises. *Mandiant*. https://www.mandiant.com/resources/increasing-low-sophistication-operational-technology-compromises. Accessed 7 Jan 2022.

Mandiant. 2016. M-Trends 2016. https://www.fireeye.com/content/dam/fireeye-www/current-threats/pdfs/rpt-mtrends-2016.pdf. Accessed 7 Jan 2022.

Mandiant. 2017. M-trends 2017: a view from the front lines. https://www2.fireeye.com/rs/848-DID-242/images/RPT-M-Trends-2017.pdf. Accessed 11 Aug 2021.

Mandiant. 2020. M-Trends 2020. https://content.fireeye.com/m-trends/rpt-m-trends-2020. Accessed 11 Aug 2021.

Marcus, J. 2019. Why Saudi Arabia and Iran are bitter rivals. *British Broadcasting Corporation (BBC)*. https://www.bbc.com/news/world-middle-east-42008809. Accessed 7 Jan 2022.

Ni, X., D. He, S. Chan, and F. Ahmad. 2016. Network anomaly detection using unsupervised feature selection and density peak clustering. In *Applied cryptography and network security*. 14th International Conference, ACNS 2016, Proceedings., 212–227. https://doi.org/10.1007/978-3-319-39555-5_12.

Noguchi, M., and H. Ueda. 2017. An analysis of the actual status of recent cyberattacks on critical infrastructures. *NEC Technical Journal, Special Issue Cybersecurity* 12(2):19–24.

Perlroth, N., and C. Krauss. 2018. A Cyberattack in Saudi Arabia Had a Deadly Goal. Experts Fear Another Try. *The New York Times*. https://www.nytimes.com/2018/03/15/technology/saudi-arabia-hacks-cyberattacks.html. Accessed 7 Jan 2022.

Perry, L., B. Shapira, and R. Puzis. 2019. NO-DOUBT: attack attribution based on threat intelligence reports. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 80–85. IEEE. https://doi.org/10.1109/ISI.2019.8823152.

Tor Project. 2019. Tor: overview. https://2019.www.torproject.org/about/overview.html.en#overview. Accessed 7 Jan 2022.

Putz, B., F. Menges, and G. Pernul. 2019. A secure and auditable logging infrastructure based on a permissioned blockchain. *Computers & Security* 87:101602. https://doi.org/10.1016/j.cose.2019.101602.

Reinhold, T. 2018. Rethinking the attribution problem—A plausible proof of non-involvement as an alternative to attribution. In *Issue Brief 2: Briefing and Memos from the Research Advisory Group of the Global Commission on the Stability of Cyberspace (GCSC)*.

Reuter, C. (ed.). 2019. *Information technology for peace and security: IT applications and infrastructures in conflicts, crises, war, and peace*. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-25652-4.

Reuter, C., J. Altmann, M. Göttsche, and M. Himmel. 2020. Zur naturwissenschaftlich-technischen Friedens- und Konfliktforschung: Aktuelle Herausforderungen und Bewertung der Empfehlungen des Wissenschaftsrats. *Zeitschrift für Friedens- und Konfliktforschung* 9(1):143–154.

Riebe, T., M.-A. Kaufhold, T. Kumar, T. Reinhold, and C. Reuter. 2019. Threat intelligence application for Cyber attribution. In *Science peace security '19—proceedings of the interdisciplinary conference on technical peace and security research*, 56–60. Darmstadt: TUprints.

Rovnyagin, M.M., S.O. Dmitriev, A.S. Hrapov, A.A. Maksutov, and I.A. Turovskiy. 2021. Database storage format for high PerformanceAnalytics of immutable data. In *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, St. Petersburg, Moscow, Russia, 618–622. https://doi.org/10.1109/ElConRus51938.2021.9396453.

Rowe, N.C. 2015. The attribution of cyber warfare. In *Cyber warfare: a multidisciplinary analysis*, ed. James A. Green, 61–72. London, New York: Routledge.

Rubio, J.E., C. Alcaraz, R. Roman, and J. Lopez. 2019. Current cyber-defense trends in industrial control systems. *Computers & Security* 87:101561. https://doi.org/10.1016/j.cose.2019.06.015.

Saudi Gazette. 2016. Saudi embassy in Tehran attacked by protesters. *Saudi Gazette*. https://web.archive.org/web/20160206202830/http://saudigazette.com.sa/saudi-arabia/saudi-embassy-in-tehran-attacked-by-protesters/. Accessed 7 Jan 2022.

Scaglia, S. 2007. *The embedded Internet: TCP/IP basics, implementation and applications*. Addison-Wesley. https://dl.acm.org/citation.cfm?id=1512809.

Schneier, B., and J. Kelsey. 1998. Cryptographic support for secure logs on untrusted machines. In *Proceedings of the 7th conference on USENIX security symposium*. San Antonio: USENIX Association.

Shute, S., R.K.L. Ko, and S. Chaisiri. 2017. Attribution using keyboard row based Behavioural Biometrics for handedness recognition. In *2017 IEEE Trustcom/bigdataSE/ICESS*, 1131–1138. Sydney: IEEE.

Starski, P. 2015. Right to self-defense, attribution and the non-state actor. *Zeitschrift für ausländisches öffentliches Recht und Völkerrecht (ZaöRV)* 75:455–501.

Steffens, Timo. 2020. *Attribution of advanced persistent threats, attribution of advanced persistent threats*. https://doi.org/10.1007/978-3-662-61313-9.

Tarakanov, D. 2012. Shamoon the wiper: further details (part II). *securelist*. https://securelist.com/shamoon-the-wiper-further-details-part-ii/57784/. Accessed 7 Jan 2022.

The State Council Information Office of the People's Republic of China. 2019. *China's national defense in the new era. Report*. Beijing: Foreign Languages Press. http://www.xinhuanet.com/english/download/whitepaperonnationaldefenseinnewera.doc. Accessed January 7 2022.

The White House. 2021. Remarks by President Biden at the office of the director of national intelligence. https://www.whitehouse.gov/briefing-room/speeches-remarks/2021/07/27/remarks-by-president-biden-at-the-office-of-the-director-of-national-intelligence/. Accessed 7 Jan 2022.

UNIDIR. 2016. Report of the international security cyber issues workshop series. https://unidir.org/publication/report-international-security-cyber-issues-workshop-series. Accessed 7 Jan 2022.

UNIDIR. 2018. Preventing and mitigating ICT-related conflict: cyber stability conference 2018 summary report. https://www.unidir.org/publication/preventing-and-mitigating-ict-related-conflict-cyber-stability-conference-2018-summary. Accessed 7 Jan 2022.

United Nations. 2015. *UN-GGE: report of the group of governmental experts on developments in the field of information and telecommunications in the context of international security.</i> A/70/174*

Warrell, H., and H. Foy. 2019. Russian cyberattack unit 'masqueraded' as Iranian hackers, UK says. *Financial Times*. https://www.ft.com/content/b947b46a-f342-11e9-a79c-bc9acae3b654. Accessed 7 Jan 2022.

Weinberg, Adam. 2021. Analysis of top 11 cyber attacks on critical infrastructure. First Point Blog. https://www.firstpoint-mg.com/blog/analysis-of-top-11-cyber-attackson-critical-infrastructure/. Accessed 20 April 2023.

Wingfield, T., and H. Wingo. 2021. International law for Cyberspace: competition and conflict. In *The Oxford handbook of cyber security, Oxford handbooks*, ed. P. Cornish https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780198800682.001.0 001/oxfordhb-9780198800682-e-37. Accessed August 25 2022.

Wissenschaftsrat. 2019. *Empfehlungen zur Weiterentwicklung der Friedens- und Konfliktforschung. Report*., 7827–7819. Gießen: Wissenschaftsrat.

Zetter, K. 2016. Inside the cunning, unprecedented hack of Ukraine's power grid. *Wired Magazine*. https://www.wired.com/2016/03/inside-cunning-unprecedented-hack-ukraines-power-grid/. Accessed 7 Jan 2022.