Helping Forensic Analysts to Attribute Cyber-Attacks: An Argumentation-Based Reasoner

Erisa Karafili¹, Linna Wang¹, Antonis C. Kakas², and Emil Lupu¹

¹ Department of Computing, Imperial College London
{e.karafili, linna.wang15, e.c.lupu}@imperial.ac.uk

² Department of Computer Science, University of Cyprus
antonis@ucy.ac.cy

Abstract. Discovering who performed a cyber-attack or from where it originated is essential in order to determine an appropriate response and future risk mitigation measures. In this work, we propose a novel argumentation-based reasoner for analyzing and attributing cyber-attacks that combines both technical and social evidence. Our reasoner helps the digital forensics analyst during the analysis of the forensic evidence by providing to the analyst the possible culprits of the attack, new derived evidence, hints about missing evidence, and insights about other paths of investigation. The proposed reasoner is flexible, deals with conflicting and incomplete evidence, and was tested on real cyber-attacks cases.

1 Introduction

We are currently facing an escalation of cyber-attacks [10] and an aggravation of their effects [3]. The expected exponential increase of the use of IoT and other smart connected devices, together with the growing dependabilities the users have on these devices, make the users more vulnerable and exposed to cyber threats. General preventive and mitigative measures are not sufficient, as there is a need to enforce protective measures that are specific to the attacker (group of attackers) that is performing the attack. Attacker-specific countermeasures would be highly effective and would contain the attack damage. These targeted measures require to discover the entity performing/related to the attack. Identifying who performed the attack would help to bring the culprits of the attack into justice.

Attribution is the process of assigning an action of a cyber-attack to a particular actor. The attribution problem is not trivial as attackers often use deceptive and anti-forensics techniques [4]. Currently, the attribution process is mainly human-based, hence easily biased and error prone, and labour intensive as it involves skilled human resources to analyze enormous amounts of low format data [2]. Digital forensics helps during the attribution process, as it collects, examines, and reports the forensic evidence [9]. Nevertheless, digital forensics techniques suffer from the limitations derived from the big amount of data to

be collected (where there is a need for efficient evidence collection) and analyzed [1,7], and from the fact that it considers only the technical aspects of an attack, without examining the social/geopolitical/economical aspects where the attack took place. A major challenge of digital forensics techniques is that they cannot work with contradictory pieces of evidence or incomplete information.

In this work, we propose an argumentation-based reasoner (*ABR*) to help the forensic analysts during the analysis and attribution of cyber-attacks. Given different pieces of digital forensics and social evidence, our reasoner derives new information, such as, the potential identity of the attacker, together with an explanation of how the reasoner arrived to its conclusion, and proposes to the user new investigation paths. *ABR* was based on preliminary results of a decision process framework introduced in [8], and is able to work with incomplete and conflicting evidence, provided by the user, whom we expect to be the forensic analyst. To the best of our knowledge, this is the first reasoner that combines technical and social evidence during the analysis and attribution process of cyber-attacks.

ABR consists of two main components: a set of reasoning rules, called *core rules*, and the *background knowledge*. The core rules model the reasoning used by the forensic analysts during the analysis and attribution of real cyber-attacks. The background knowledge is formed of common knowledge usually used by the analyst, e.g., countries characteristics, prominent cyber groups, past attacks. The core rules use the pieces of evidence given by the user together with the background knowledge to derive a conclusion. The reasoning behind *ABR* is unbiased and its combination with expert knowledge provides an efficient and accurate analysis and attribution.

We based our reasoner on the Q-Model [13], a social science attribution model, in order to include the social evidence in our reasoning, alongside technical indications. This model describes the procedure of putting together technical and social evidence, followed by the forensic analysts during the attribution of cyber-attacks. The use of the Q-Model enables *ABR* to structure the reasoning rules and the evidence in three different layers: technical, operational, and strategic, and to work with social and technical pieces of evidence.

We decided to use argumentation reasoning for our reasoner as it allows to capture the knowledge in the same natural and direct form as the human forensic analysts. The use of preference-based argumentation [6,5], permits *ABR* to deal with conflicting pieces of information, by introducing preferences between rules. We use Gorgias [5], a preference-based argumentation reasoning tool that uses abduction, to construct *ABR*. The ability to work with incomplete information makes *abduction* suitable to use in the attribution process, as it fills the *knowledge gaps* in our reasoning, and gives to the analyst potential leads to follow and carry out further investigations.

Prior to our work argumentation has been applied to attribution in [14,11], where a probabilistic model is introduced. However, the framework introduced in [14,11] does not use any social evidence or background knowledge that are useful in detecting motives, capabilities, and potential culprits.

The goal of our reasoner is *not to* substitute the forensic analyst but rather to provide a supporting tool to help with the analysis and attribution of cyber-attacks. Therefore, when *ABR* gives an output to the user, e.g., who is the culprit, or new insights, it provides also an explanation of the given result, allowing the user to adjust and refine the reasoning used by *ABR*.

In Section 2 we introduce the argumentation framework used by *ABR*. We present *ABR* and its main components in Section 3. We conclude in Section 4 and propose some future research directions.

2 Preference-Based Argumentation Framework

The proposed reasoner is based on preference-based argumentation [12,5], as we believe it is best suited for interacting with an analyst during the analysis and attribution of cyber-attacks. In particular, we use the framework proposed in [6,5] that represents multi-agent application problems via argumentation reasoning. This framework permits us to work with conflicting pieces of evidence, or pieces of evidence that derive conflicting conclusions by introducing preference rules. Preference-based argumentation allows us to handle *non-monotonic reasoning* in attribution, where the introduction of new evidence might change the result of the attribution (due to conflicting arguments). For example, given the argument pair (T, P) ³ extracted from the core rules of *ABR*:

$$\begin{aligned} T = \{ & r_1 : \text{attackOrig}(X, \text{Attack}) \leftarrow \text{ipGeoloc}(X, IP), \text{attackSourceIP}(IP, \text{Attack}), \\ & r_2 : \neg \text{attackOrig}(X, \text{Attack}) \leftarrow \text{ipGeoloc}(X, IP), \text{attackSourceIP}(IP, \text{Attack}), \\ & \text{spoofedIP}(IP) \} \\ P = \{ & \text{pref}_1 : r_2 > r_1 \} \end{aligned}$$

where rule r_1 states that *Attack* originates from X , ($\text{attackOrig}(X, \text{Attack})$), if the source IP of the attack is geolocated in X , while r_2 states that *Attack* does not originate from X , if the source IP of the attack is geolocated in X and the IP is spoofed. If we have the following pieces of evidence:

$$E = \{\text{attackSourceIP}(ip1, \text{attack1}), \text{ipGeoloc}(\text{countryC}, ip1)\}$$

we get $\text{attackOrig}(\text{countryC}, \text{attack1})$ as the conclusion of the argument. However, if we have the evidence:

$$E = \{\text{attackSourceIP}(ip2, \text{attack2}), \text{ipGeoloc}(\text{countryC}, ip2), \text{spoofedIP}(ip2)\}$$

then the conclusion is $\neg \text{attackOrig}(\text{countryC}, \text{attack2})$, thus *attack2* was not originated from *countryC*, as according to the preference rule pref_1 , we prefer rule r_2 over rule r_1 .

3 Argumentation-Based Reasoner

Our argumentation based reasoner, *ABR*, is based on two main components, the *core rules* and the *background knowledge*, see Figure 1. The core rules use the

³ T represents the argument rules, while P represents the preference rules.

Fig. 1. *ABR* Overview

pieces of evidence of an observed cyber-attack, which are given as input by the user, together with the background knowledge and give as output the answer as to who/where the observed attack can be attributed to. As the main goal of *ABR* is to help the analyst to gain new insights, it is crucial that *ABR* provides explainable and transparent results to the user. Therefore, *ABR* outputs the result as well as its associated score and the used derivations together with their supporting and/or conflicting rules, and the corresponding graphical representations. The score represents the confidence *ABR* has in the result, and is used to order multiple results. Furthermore, *ABR* provides hints about other investigation paths, by giving a list of missing pieces of evidence that, if provided, can bring into new conclusions. The suggested hints permit the analyst to put in act an efficient prioritized evidence collection. *ABR* models the iterative and incremental nature of the attribution process, where it provides an answer with new information to the analyst, and can integrate additional input from the analyst, which makes *ABR* suitable to be used in conjunction with human experts.

Let us now introduce an example, where, for the sake of understandability, we simplify the representation of rules and pieces of evidence.

Example 1. Suppose an attack a has occurred, and the analyst is able to recover the IP's from where the attack came from, geolocate them, and check if they can be used or not for the analysis process by checking if they are spoofed or not. An analysis of the attack states that it was performed using malware m and

that m shares code with another known malware $mal2$. We represent below the pieces of evidence that the analyst provides to ABR , denoted by $caseE_i$.

$$\begin{aligned} caseE_1 &: attackSourceIP(ip1, a), & caseE_2 &: attackSourceIP(ip2, a), \\ caseE_3 &: ipGeoloc(ip1, countryX), & caseE_4 &: ipGeoloc(ip2, countryY), \\ caseE_5 &: spoofedIP(ip2), & caseE_6 &: hasMotive(countryY, a), \\ caseE_7 &: malwareUsed(m, a), & caseE_8 &: sharedCode(m, mal2). \end{aligned}$$

The first five focus on technical evidence related to the IP's from where the attack originates, $ip1$, $ip2$, and their geolocation, correspondingly in $countryX$ and $countryY$. $ip2$ is found to be spoofed. Evidence $caseE_6$ provides ABR with some social information related to the attack, stating that $countryY$ has motives to perform a . The last two pieces of evidence, deal with the type of attack, stating that malware m was used in a , and it shares code with $mal2$.

We introduce below some of the background knowledge of ABR that might be useful for the given example.

$$\begin{aligned} bg_1 &: malwareLinkedTo(mal2, aGroup), & bg_2 &: notFromBlackMarket(mal2), \\ bg_3 &: country(countryX), & bg_4 &: country(countryY). \end{aligned}$$

The first two pieces of information are part of the domain specific knowledge. bg_1 states that $mal2$ is linked to a known group of attackers $aGroup$, and bg_2 states that $mal2$ is not from the black market. bg_3 and bg_4 are general knowledge, where ABR recognises $countryX$ and $countryY$ as countries. $\square$

3.1 Division of the Core Rules

The core rules are composed of the reasoning rules used to derive the various conclusions. We extracted the core rules from past real cyber-attacks' attributions, where the rules represent the reasoning process followed by the forensic analysts. To deal with the social evidence, we base our reasoner on a social model for attribution, called *Q-Model* [13]. Following the Q-Model, the pieces of evidence (given and derived), and the reasoning rules are divided into three layers: *technical*, *operational*, and *strategic*.

In our reasoner, the rules of the *technical layer* deal with technical evidence of the attack, e.g., IP from where the attack was originated, time of attack, logs. The rules of the *operational layer* deal with other social aspects of the attack, e.g., the motives of the attack, the needed capabilities to perform it. The rules of the *strategic layer* deal with who might have done the attack, or who is taking advantage from it. Depending on the layer a rule/evidence is part of, we call it a technical, operational, or strategic rule/evidence.

Let us show through an example, how these layers are connected to each other and what type of evidence and rules goes to which layer. One of the rules of the strategic layer is the following⁴:

$$s_1 : isCulprit(X, A) \leftarrow hasMotive(X, A), hasCapability(X, A).$$

⁴ The rule's name represents the layer of the rule, i.e., the rules' names of the technical, operational, and strategic layer start correspondingly with t , o , and s .

saying that entity X is the culprit of attack A , if it has the motivations to perform the attack $hasMotive(X, A)$ and it has the capability to conduct it $hasCapability(X, A)$. The evidence used in the above rule is given by the user or is derived using rules of the technical and operational layer. In particular, the evidence $hasCapability$ depends on the type of attack.

$$o_1 : hasCapability(X, A) \leftarrow requireHighResource(A), hasResource(X).$$

The above rule is an operational layer rule and states that X has the capability to perform attack A , in case A requires high resources $requireHighResource(A)$ and X has these resources $hasResource(X)$. The rule below, which is part of the technical layer, derives the $requireHighResource(A)$ evidence from the skill level needed by the attacker to perform A .

$$t_1 : requireHighResource(A) \leftarrow highLevelSkill(A).$$

The $highLevelSkill(A)$ evidence is derived using the technical layer rule below:

$$t_2 : highLevelSkill(A) \leftarrow malwareUsed(M, A), usesZeroDayVuln(M).$$

where the used pieces of evidence are technical ones. The above rule states that A is a high level skill attack, if the used malware, $malwareUsed(M, A)$, exploits zero day vulnerabilities, $usesZeroDayVuln(M)$.

Example 2. Following from Example 1, reasoning rules can be used to derive the results. For example, we show below a strategic reasoning rule used by ABR to derive that $aGroup$ might be the culprit of the attack.

$$s_2 : isCulprit(X, A) \leftarrow malwareUsed(M1, A), similar(M1, M2), \\ notFromBlackMarket(M1), \\ notFromBlackMarket(M2), \\ malwareLinkedTo(M2, X).$$

The rule states that a certain entity X is the attacker, if the malware used $M1$ is similar to malware $M2$ that is linked to entity X . The $similar(M1, M2)$ evidence can be derived using a technical rule, such as:

$$t_3 : similar(M1, M2) \leftarrow sharedCode(M1, M2).$$

which considers that two malwares are similar if they share a significant part of their code, $sharedCode(M1, M2)$. $\square$

3.2 ABR execution

Given the pieces of evidence to ABR , the user can execute ABR in two modes. In *standard mode*, it gives as result the answers to the queries, e.g., whether a particular entity performed the attack. In *verbose mode*, it gives as result hints of what other pieces of evidence are required to reach a specific conclusion, given the evidence found so far. We tested ABR with real cyber-attacks cases, taken from public reports, and artifactual examples. ABR was able to correctly identify the culprits given the appropriate pieces of evidence. In verbose mode, ABR 's results include the missing evidence (when some pieces of evidence were omitted) or suggest appropriate investigation paths to be followed by the analyst.

Standard Mode The standard execution is called after providing to *ABR* the pieces of evidence and making a query. This mode gives as result the answer of the query, e.g., if an entity is the culprit or not of the attack, if there are other attacks similar to the exiting one, from where the attack originated, or all the entities that might be the culprits of the examined cyber-attack. For every given answer, *ABR* provides the derivations rules and pieces of evidence used to arrive at the conclusion, together with a graphical representation of the derivation, and a score of the result, based on a scoring system. It also provides an argumentation tree of the used reasoning rules that shows how the different conflicting arguments supporting/attacking the various possibilities of attribution are considered.

Example 3. Going back to Example 1, when *ABR* is given the pieces of evidence and is asked “who is the culprit” it gives the following output:

aGroup, countryX, countryY.

ABR gives three different answers as it is able to prove for each of them that they might be the culprit. Thus, it provides all the above answers together with their explanations. $\square$

Verbose Mode The verbose mode is usually used when there are not enough pieces of evidence to support a conclusion, or when the user wants to know other paths of investigation. Given the pieces of evidence *ABR*’s verbose execution identifies the *missing pieces of evidence* necessary to reach other or more precise conclusions. The list suggests to the user other pieces of evidence to be collected, and hints to other investigation paths.

Example 4. Some of the results of executing *ABR* in verbose mode given the pieces of evidence of Example 1 are shown below:

hasMotive(countryX, a), target(X, a), hasMotive(countryZ, a).

The first suggestion asks if *countryX* has motive to perform the attack, in order to make the attribution more precise. The second one, *target(X, a)*, asks which were the targets of *a*. The last suggestion, asks if the user can provide evidence (*hasMotive*) for another country, *countryZ*, as this country is linked with a prominent group of attackers that is able to perform this type of attack, and it usually performs attacks for *countryZ*. $\square$

4 Conclusions and Future Works

A correct and swift attribution of cyber-attacks permits to put in place adequate mitigative and preventive measures. Attributing cyber-attacks is a delicate and difficult task, due to the large amount of often conflicting evidence that needs to be analysed and the fact that attackers use deceptive and anti-forensics techniques. In this work we proposed a reasoner (*ABR*), based on a preference-based

argumentation that helps the forensic analyst during the analysis of forensic evidence of an observed cyber-attack and its attribution. *ABR* provides the analyst with an answer as of where the observed attack originated and to who it can be attributed. The proposed reasoner automates the attribution process and is fully flexible, as it allows and supports the analyst to introduce new evidence, rules, and preferences.

As future work, we plan to increase *ABR* reasoning capabilities by adding new reasoning rules, and background knowledge. In order to fully automate the attribution process, we intend to enrich *ABR* with an automatic evidence extraction/collection, by integrating *ABR* with digital forensics tools and data mining techniques.

Acknowledgments

Erisa Karafili was supported by the European Union's H2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 746667.

References

1. Carrier, B.: Defining digital forensic examination and analysis tools using abstraction layers. *International Journal of Digital Evidence* **1**(4), 1–12 (2003)
2. da Cruz Nassif, L.F., Hruschka, E.R.: Document clustering for forensic analysis: an approach for improving computer inspection. *IEEE Trans. Inf. Forensic Secur.* **8**(1), 46–54 (2013)
3. DCMS: Cyber security breaches survey 2018 (2018), <https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2018>
4. Goutam, R.K.: The problem of attribution in cyber security. *International Journal of Computer Applications, Foundation of computer science* **131**(7), 34–36 (2015)
5. Kakas, A., Moraitis, P.: Argumentation based decision making for autonomous agents. In: *AAMAS '03*. pp. 883–890 (2003)
6. Kakas, A.C., Mancarella, P., Dung, P.M.: The acceptability semantics for logic programs. In: *Intern. Conference on Logic Programming*. pp. 504–519 (1994)
7. Karafili, E., Cristani, M., Viganò, L.: A formal approach to analyzing cyber-forensics evidence. In: *ESORICS 2018*. pp. 281–301 (2018)
8. Karafili, E., Kakas, A.C., Spanoudakis, N.I., Lupu, E.C.: Argumentation-based Security for Social Good. In: *AAAI Fall Symposium Series* (2017)
9. Kent, K., Chevalier, S., Grance, T., Dang, H.: SP 800-86. Guide to Integrating Forensic Techniques into Incident Response. Tech. rep., NIST (2006)
10. Newman, L.H.: The Biggest Cybersecurity Disasters of 2017 So Far. <https://www.wired.com/story/2017-biggest-hacks-so-far/> (2017)
11. Nunes, E., Shakarian, P., Simari, G.I.: Toward argumentation-based cyber attribution. In: *AAAI Workshops* (2016)
12. Prakken, H., Sartor, G.: Argument-based extended logic programming with defeasible priorities. *J. of Appl. Non-Class. Log.* **7**(1), 25–75 (1997)
13. Rid, T., Buchanan, B.: Attributing Cyber Attacks. *Journal of Strategic Studies* **38**(1-2), 4–37 (2015)
14. Shakarian, P., Simari, G.I., Moores, G., Parsons, S.: Cyber Attribution: An Argumentation-Based Approach. In: *Cyber Warfare*, pp. 151–171. Springer (2015)