Journal of Information Security and Applications 37 (2017) 101–111

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

On inter-Rater reliability of information security experts

Abdulhadi Shoufan*, Ernesto Damiani

Information Security Center, Khalifa University, Abu Dhabi, UAE

ARTICLE INFO

Article history:
Available online 7 November 2017

Keywords:
CIA triad
Security qualitative assessment
Inter-rater reliability
Fleiss' kappa
Rater bias.

ABSTRACT

The *Confidentiality-Availability-Integrity (CIA) triad* is a time-honored warhorse of security analysis. Qualitative assessment of security requirements based on the CIA triad is an important step in many standard procedures for selecting and deploying security controls. However, little attention has been devoted to monitoring how the CIA triad is used in practice, and how reliable are experts' assessment that make use of it. In this paper, a panel of 20 security experts was asked to use the CIA triad in 45 practical security scenarios involving UAV-to-ground transmission of control and information data. The experts' responses were analyzed using *Fleiss' kappa*, a specific statistics test for inter-rater reliability. Results show agreement to be low (from 13.8% to 20.1% depending on the scenario), but higher on scenarios where the experts' majority estimates tight security to be needed. Low number of polled experts is found to affect inter-rater reliability negatively, however, increasing this number beyond ten does not provide additional reliability. A bias to give a specific rate could be identified with 14 out of the 20 experts. The six unbiased experts showed a higher inter-rater agreement. These findings suggest that (i) there is no guaranteed "safety in numbers" for recruiting security expert panels and (ii) expert selection for security rating processes should include verification of agreement level on toy problems for all subsets of the panel to highlight subsets showing high inter-rater agreement.

© 2017 Published by Elsevier Ltd.

1. Introduction

Polling experts' opinion is an important technique in all settings where experts are asked to rate or rank assets, vulnerabilities, threats, risks, attacks, and security objectives according to different criteria. Areas where experts' opinion are routinely collected and processed include threat analysis, which is often based on surveys. In their seminal paper on ICT threats' analysis, Loch et al. identified serious threats to information systems and resident data [1] by preparing a list of threats from the literature and asking a pool of security executives and consultants to rank the top three. In [2], IT executives were asked in an online survey to rank threats to information security, to identify the priority of expenditures to protect against these threats, and to estimate the frequency of attacks. The European Network and Information Security Agency (ENISA) routinely uses experts' panels to identify and classify threats in a number of areas, including smart health services, virtualized ICT infrastructures, and Big Data security. Experts are also polled about security concerns regarding new technology developments; for instance, experts were asked to name major privacy concerns in social network applications [3]. Expert opinion is also used in the initial stages of event or incidents of potential concern to facilitate rapid risk assessment, i.e. give an estimate of risk posed by a threat. Rapid risk assessment is a core part of incident response and thus widely undertaken by security professionals. Additionally, standards for IT service management that describe techniques for establishing IT service strategy like ITIL (formerly an acronym for Information Technology Infrastructure Library) and ISO/IEC 20000 (previously BS 15000) explicitly require use of experts' assessment to develop information security standards, policies, or guidelines.

Expert security ratings, however, are not perfect and some authors questioned their reliability and validity. Halsum et al., for instance, list inconsistent expert judgment among the causes of uncertainty in risk assessment [4]. Along the years, other researchers have tried to improve the outcome of experts' ratings using different methods (see Section 2). However, few quantitative results are available on consistency of assessments among information security experts, and even evaluations of the tools for measuring such consistency are almost entirely missing. The contribution of this paper is twofold: (i) We propose an experimental design for measuring inter-rater reliability in security assessments. Inter-rater reliability is a statistical concept that describes the degree of agreement among raters. Our approach uses Fleiss' kappa statistics according to the interpretation by Landis and Koch [5] to check consistency between expert assessments. (ii) We validate our design

* Corresponding author.
E-mail addresses: abdulhadi.shoufan@kustar.ac.ae (A. Shoufan), ernesto.damiani@kustar.ac.ae (E. Damiani).

https://doi.org/10.1016/j.jisa.2017.10.006
2214-2126/© 2017 Published by Elsevier Ltd.

with an extensive case study on a real security problem and analyze its results.

We focus on security rather than safety because standard systems and processes for safety ratings involve physical device testing instead of (or in addition to) expert opinions [6]. We rely on the classic *Confidentiality-Availability-Integrity (CIA) triad* for expressing security objectives. The CIA triad was introduced in the Nineties as a multi-purpose, standard way to express security requirements concerning information assets. Later, it was often extended to include additional security properties or to address specific domains [7]. The triad's use for standardizing security experts' responses is common practice in the field and well-documented in the literature. For instance, experts assign weights to the triad's components in order to instantiate a fuzzy model for risk assessment [8]. In the US, using the CIA triad in ratings is one of the standard practices listed by the National Institute of Standards and Technology (NIST) for rating ICT systems used in the federal public administration [9]. Use of CIA in security rating system is also advised by professional groups and associations. According to the ISACA Rating General Description,[1] "A rating system is based on the typical five levels (from *A* to *E*), which are assigned to three dimensions of security for each service rated: Confidentiality, Integrity and Availability".

We applied our experimental design by performing a study on the security of communications between Unmanned Aerial Vehicles (UAVs, often called *drones*) and ground stations. We argue that this domain's strong link to classic transmission security ensure some knowledgeability and confidence on the part of security experts, while its relative novelty prevents agreement by "conventional wisdom".

We applied our experimental design by performing a study on the security of communications between Unmanned Aerial Vehicles (UAVs, often called *drones*) and ground stations. We argue that this domain's strong link to classic transmission security ensure some knowledgeability and confidence on the part of security experts, while its relative novelty prevents agreement by "conventional wisdom".

We selected 45 use cases of civil drones and asked a panel of 20 experts to use the CIA triad to rate for each case the security objectives regarding control and information data exchanged between drones and ground stations. The experts' responses were analyzed using *Fleiss' kappa*, a specific statistics test for inter-rater reliability [5]. The experts' overall agreement is low enough to raise serious concerns (from 13.8% to 20.1% depending on the scenario). However, closer analysis shows a trend to "agree on the extremes": expert agreement is substantially higher when the general perception of the security level of the use case is especially low or especially high.

A further analysis was performed to check experts' bias and how such bias would affect the inter-rater reliability. We found out that 14 out of the 20 experts show a permanent tendency to give a low, medium, or high rate, regardless of the case they are assessing. The 20 experts were then clustered according to their rating bias and the inter-rater reliability in each cluster was analyzed separately. Interestingly, we found out that unbiased raters show better inter-rater reliability. This result hints at using bias control on toy problems as a technique for expert selection.

Also, the impact of the number of the experts on the inter-rater reliability was investigated. We found out that Fleiss' kappa increases with the number of experts as long as the latter is below 10. Increasing the number of raters beyond 10, however, does not affect the agreement level. We claim that analyzing the influence of inter-rater agreement on panel-based security ratings can provide some operational suggestions to ensure that these ratings help rather than harm businesses' security decision-making. According to the principles endorsed by the US Chamber of Commerce for security ratings,[2] reporting expert opinions should "include a coordinated process for adjudicating errors or inaccuracies". Our results suggest that a posteriori analysis of inter-rater agreement should become a key part of such a coordinated process, as well as of other risk assessment procedures that make use of expert ratings.

The remainder of the paper is structured as follows. Section 2 reviews the related work on security assessments. Our methodology is presented in detail in Section 3. Our experiment and its results are described in Section 4 and discussed in Section 5. Section 6 describes the implications and limitations of this research. Section 7 draws our conclusions and provides an outlook.

2. Related work

In this section we provide a literature review, which is structured as follows. To show the relevance of our study we first review some related work that has highlighted some issues associated with qualitative approaches in information security [10–13]. Then, we describe concerns raised about experts' judgment in this field [14–16]. Following, we argue for the use case-based approach employed in our study [17,18]. Finally, we review some related work on drone security [19–25].

Uncertainty associated with qualitative methods in information security is a well-known issue. Some authors relate it to using imprecise natural language for communication [10]. While rating risks using ordered categorical labels such as "low", "medium", and "high" can simplify risk assessment, some researchers believe that this approach does not necessarily improve decisions [11]. To mitigate the impact of uncertainty in qualitative methods, some researchers proposed applying the classic Delphi method to security analysis [12,13]. This method relies on a number of rounds of experts' rating. The outcome of each round is fed back to the experts who revise their ratings in an iterative manner. It is believed that the Delphi method helps the experts' ratings to converge towards a "correct" answer [12].

Miller et al. attribute the uncertainties of designing secure software systems to missing data on uncommon attacks, difficulty of security cost estimation, and continuous change in technology and tools [14,15]. The authors claim that uncertainty has an impact on the experts' perceptions of security risks, which in turn leads to wide variations in their assessments of potential attacks' probability and severity. Their approach is based on Spearman's Rho statistics, which measures the statistical dependence of two sets of rankings. It takes values between -1 and $+1$, whereas -1 and $+1$ indicate perfect negative or positive correlation, respectively, and 0 indicates no correlation. The authors found rankings of the same attacks across multiple scenarios to be weakly or un-correlated. Interestingly, an unpublished paper by the same authors reports on an experiment where experts ranked the attack vectors in a single scenario [16]. The authors studied the agreement between the rankings using a different statistics (known as Kendall's W) and identified a relatively high agreement. Increase in the agreement may be attributed to the fact that the experts rated a single scenario which was accurately specified by the researchers.

In many cases, however, accurate specification of attack scenarios is missing. This is especially true for emerging technologies. When technologies are still in their infancy, risk assessors

[1] https://www.isaca.org/groups/professional-english/cloud-computing/groupdocuments/rating_general_descriptionv1.0.pdf.

[2] https://www.uschamber.com/above-the-fold/why-we-need-fair-and-accurate-cybersecurity-ratings.

Download English Version:

https://daneshyari.com/en/article/6884631

Download Persian Version:

https://daneshyari.com/article/6884631

Daneshyari.com