Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

An expert-based investigation of the Common Vulnerability Scoring System

Hannes Holm [a,*], Khalid Khan Afridi [b]

[a] Swedish Defence Research Agency (FOI), 583 30 Linköping, Sweden
[b] Stockholm University, 164 40 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form
30 March 2015
Accepted 25 April 2015
Available online 22 May 2015

Keywords:
Cyber security
Vulnerabilities
Security metrics
Expert judgment
Common Vulnerability Scoring System

ABSTRACT

The Common Vulnerability Scoring System (CVSS) is the most widely used standard for quantifying the severity of security vulnerabilities. For instance, all vulnerabilities in the US National Vulnerability Database are scored according to this system. Unfortunately, it is largely unexplored whether or not its scores are accurate. This paper studies this property through a survey with opinions by 384 experts, covering more than 3000 vulnerabilities. The results show that the mean disagreement between the judgments of the experts and the CVSS Base Score is −0.38, with a variance of 4.46 (on a scale from 0 to 10). The direction of this difference depends on the type of vulnerability that is concerned. The experts then suggest a number of possible revisions to the CVSS that could explain this difference.

1. Introduction

The importance of computer and network security is steadily increasing in relation to the development of wide spread global infrastructure technologies and progressively more complex enterprise IT environments. Organizations are forced to spend a lot of effort on cyber security issues as intrusions can cause significant losses of confidentiality, integrity and availability. Sound cyber security decision models can enable decision makers to conduct more well-informed choices on this matter and thus reduce the risk and impact of security incidents.

One of the most widely used cyber security models is the Common Vulnerability Scoring System (CVSS) (Mell et al., 2007), which is used to quantify the severity of security vulnerabilities. For instance, all vulnerabilities in the US National Vulnerability Database (NVD) (NIST, 2013) are scored according to this system (more than 61,600 to date). The CVSS quantifies the severity of vulnerabilities according to three metrics, of which the main is aptly named "Base Score". The Base Score uses an interval scale of 0–10 for measuring severity of vulnerabilities. This scale corresponds to three discrete states: Low severity, Medium severity and High severity (some also employ a fourth state, Critical severity, however, this is not defined in the CVSS standard). This metric is furthermore an aggregation of two other submetrics, "Exploitability", the ease of utilizing a vulnerability, and "Impact", the damage measured in confidentiality, integrity or availability that exploitation of the vulnerability can lead to. The

* Corresponding author.
E-mail address: hannes.holm@foi.se (H. Holm).
http://dx.doi.org/10.1016/j.cose.2015.04.012

current revision of the CVSS is v2. Whenever the CVSS is mentioned in the remainder of this paper, otherwise stated, it refers to the CVSS v2.

The CVSS developers suggests that security professionals should conduct the scoring of Base Score and Temporal Score, and that users (e.g., system administrators) should provide the Environmental Score (Mell et al., 2007):

> "Generally, the base and temporal metrics are specified by vulnerability bulletin analysts, security product vendors, or application vendors because they typically have better information about the characteristics of a vulnerability than do users. The environmental metrics, however, are specified by users because they are best able to assess the potential impact of a vulnerability within their own environments."

While analysts generally are capable of describing the Temporal Score, this is rarely made in practice. In fact, Temporal Score does not even have any search fields in the NVD. Thus, from the perspective of many users', the Base Score *is* the CVSS.

While the CVSS is a great effort for enabling comparisons of different vulnerabilities, it is relatively unknown whether its estimates are valid. In other words, is its scoring algorithm accurate? If it is inaccurate, is it due to errors in the mathematical formula or due to lack of completeness (i.e., whether significant attributes and states are missing)? Here we define accuracy as the difference between the *actual* severity of a vulnerability and its severity as indicated by its corresponding CVSS score.

The actual severity of a vulnerability is however difficult to capture in a valid manner as there is no standard method available for this purpose. As a consequence, previous studies of the CVSS have examined its validity from a wide variety of angles, e.g., by measuring the time required to compromise computers (Holm et al., 2012), by observing what vulnerabilities that actually are exploited in the wild (Allodi and Massacci, 2012; Allodi et al., 2013), or by studying the distribution of vulnerability severity levels in the NVD (Liu and Zhang, 2011). Unfortunately, these studies do not attempt to identify the actual reasons behind why their measurements differ from what is expected according to the CVSS. For example, it is not analyzed in depth why some computers require more time to compromise than others' (Holm et al., 2012) and why some vulnerabilities are exploited more frequently than others' (Allodi et al., 2013).

A cause of these delimitations is that it is difficult to capture such data during experiments, and even more so for systems in operation. Given this difficulty, a reasonable alternative is to employ expert judgment. Expert judgment is frequently used to estimate phenomena that otherwise are difficult to measure (Cooke, 1991). In the domain of cyber security, it has been employed to measure a range of different aspects, from very abstract measures such as overall risk (Ryan et al., 2012) to more concrete measures such as the effort required to discover novel web application vulnerabilities (Holm et al., 2013a). Expert judgment was used to create the CVSS itself and is used to assign states to CVSS Base Score submetrics for new vulnerabilities.

This paper presents a questionnaire-based study that asked experts to judge the severity of security vulnerabilities and provide motivations behind their judgments. This method was chosen due to that analysts categorize vulnerabilities according to the CVSS, and thus effectively provide the CVSS scores, yet are unable to express their own perceived severity and the actual reasons behind their opinions. This investigation is summarized by the research question below.

- *RQ1*: How accurate are the values produced by the CVSS?

Accuracy is operationalized as the disagreement between vulnerability severity estimates by the CVSS and vulnerability severity estimates by cyber security experts. With CVSS, this research refers to CVSS v2 and Base Score, the metric that is described for vulnerabilities in the NVD, as this is the viewpoint that is available to practitioners. The experts were provided typical vulnerability information (descriptions and Base Score states, see Section 4.3) and told to judge the severity of each corresponding vulnerability based on their own experiences. This means that the experts were allowed to consider variables out of the scope of the CVSS. The overall motivations behind this are (i) that the validity of the Base Score is unclear and (ii) that each expert is bound to have experiences related to vulnerability severity that might not be not completely captured by the CVSS. This research question is more formally expressed by hypothesis H1 below.

- *H1*: There is a significant difference between vulnerability severity estimates by the CVSS Base Score and vulnerability severity estimates by cyber security experts.

It is a difficult task to examine whether the constructs of the CVSS are appropriate from a theoretical perspective. That is, the constructs are all clearly related to vulnerability severity, but also very abstract. For instance, loss of integrity is related to vulnerability impact – yet it can be difficult to deduce what "Partial" integrity impact actually means. Unfortunately, there is not yet any known "truth" or "golden standard" available in literature on these matters. An exploratory approach involving the previously mentioned security experts (who have first-hand experiences on the matter) is thus a reasonable method for investigating the validity of the CVSS constructs. As the experts were not constrained by the CVSS metrics, their opinion is able to provide a more comprehensive explanation of vulnerability severity than the constructs of the Base Score, Temporal Score and Environmental Score. The overall research question regarding this topic is described below.

- *RQ2*: Are the current constructs of the CVSS Base Score appropriate?

The remainder of the paper is structured as follows: Section 2 presents a brief overview of the CVSS. Section 3 presents related works. Section 4 presents the methodology of the study. Section 5 describes the results of the study. Section 6 critically examines the results from the study. Finally, Section 7 concludes the paper.