CYBER SECURITY METRICS AND MEASURES

Paul E. Black, Karen Scarfone and Murugiah Souppaya
National Institute of Standards and Technology, Gaithersburg, Maryland

Abstract: Metrics are tools to facilitate decision making and improve performance and accountability. Measures are quantifiable, observable, and objective data supporting metrics. Operators can use metrics to apply corrective actions and improve performance. Regulatory, financial, and organizational factors drive the requirement to measure IT security performance. Potential security metrics cover a broad range of measurable features, from security audit logs of individual systems to the number of systems within an organization that were tested over the course of a year. Effective security metrics should be used to identify weaknesses, determine trends to better utilize security resources, and judge the success or failure of implemented security solutions.

Keywords: cyber security; metrics; measures; software; computer systems; IT

Cyber security metrics and measures can help organizations (i) verify that their security controls are in compliance with a policy, process, or procedure; (ii) identify their security strengths and weaknesses; and (iii) identify security trends, both within and outside the organization's control. Studying trends allows an organization to monitor its security performance over time and to identify changes that necessitate adjustments in the organization's security posture. At a higher level, these benefits can be combined to help an organization achieve its mission by (i) evaluating its compliance with legislation and regulations, (ii) improving the performance of its implemented security controls, and (iii) answering high-level business questions regarding security, which facilitate strategic decision making by the organization's highest levels of management. This article defines

Wiley Handbook of Science and Technology for Homeland Security, Edited by John G. Voeller

some terms, and then discusses the current state of security metrics, focusing on the measurement of operational security using existing data collected at the information system level. This article explains the importance of selecting measures that support particular metrics and then examines several problems with current practices related to the accuracy, selection, and use of measures and metrics. The article also presents an overview of a security metrics research effort, to illustrate the current state of metrics research, and suggests additional research topics.

1 CONTRASTING METRICS AND MEASURES

The term *metric* is often used to refer to the measurement of performance, but it is clearer to define metrics and measures separately. A *measure* is a concrete, objective attribute, such as the percentage of systems within an organization that are fully patched, the length of time between the release of a patch and its installation on a system, or the level of access to a system that a vulnerability in the system could provide. A *metric* is an abstract, somewhat subjective attribute, such as how well an organization's systems are secured against external threats or how effective the organization's incident response team is. An analyst can approximate the value of a metric by collecting and analyzing groups of measures, as is explained later.

Historically, many metrics efforts have focused on collecting individual measures, and given little or no thought as to how those measures could be combined into metrics. Ideally, organizations should first select their metrics, and then determine what measures they can perform that support those metrics. An organization should also have multiple levels of metrics, each geared toward a particular type of audience. For example, technical security staff might be interested in lower-level metrics related to the effectiveness of particular types of security controls, such as malicious code detection capabilities. Security management might be interested in higher-level metrics regarding the organization's security posture, such as the overall effectiveness of the organization's incident prevention and handling capabilities. Lower-level metrics facilitate making more tactical decisions, whereas higher-level metrics are well suited for making more strategic decisions. The lower-level metrics are often used as input to the higher-level metrics.

Organizations can use measures and metrics to set goals, also known as *benchmarks*, and determine success or failure against the benchmarks. For example, suppose that an organization determines that 68% of its systems are in compliance with a particular policy. The organization could set a benchmark of 80%, implement changes in its practices to increase compliance, and then measure compliance again in six months to see if the benchmark has been achieved. Benchmarks are organization specific and are typically based on baselines from an operational environment.

2 SELECTING MEASURES TO SUPPORT METRICS

Once an organization has identified its metrics, it then needs to determine what measures can feasibly be collected to support those metrics. Organizations should favor measures that can be collected via automated means because they are more likely to be accurate than manual collection (e.g. self-evaluation surveys) and can also be collected as often as needed. Organizations should also seek opportunities to use existing data sources and automated collection mechanisms because of the cost of implementing and maintaining new systems and software simply for data collection purposes.

As measures are collected, organizations need a way to analyze them and generate reports for the metrics they support. Organizations can analyze the measures and metrics in many ways, such as grouping them by geographic location, logical division within the organization, system type, system criticality, and so on. Some organizations use products that roll up measures into metrics and present the metrics in a security dashboard format, with the measures underlying each metric available through drill-down. This allows a dashboard user to see the values of the presented metrics and changes in those metrics over time, as well as to examine the metrics and measures comprising those metrics.

3 PROBLEMS WITH THE ACCURACY OF MEASURES

The accuracy of a metric is by definition dependent on the accuracy of the measures that comprise the metric. Organizations currently face several problems related to the accuracy of measures. One problem is that measures are often defined imprecisely. Consider the percentage of systems that are fully patched: does this only include operating system patches or does it also include service and application patches? Does it only mean that the patches have been installed or that subsequent actions necessary to activate the patch (such as rebooting the system or changing configuration settings) have also been performed? Another issue with measure definition is the terminology itself, such as measuring the number of port scans performed. What is the minimum number of ports that must be scanned in a port scan? If an attacker scans ports on 100 hosts, is it one port scan or 100 port scans? If the attacker performed the same scan but only scanned one host each day, is it one port scan or 100 port scans?

Measuring port scans is also a good example of a related common problem—inconsistent measurement methods. Port scans are often identified by intrusion detection systems (IDSs), but each IDS uses its own proprietary algorithms for identifying port scans, so activity identified as a port scan by one IDS may not be identified as such by another. This causes inconsistency in measurement if the organization uses multiple IDSs or if a single product is in use but its sensors have different port scan settings (e.g. the minimum number of ports in a scan or the maximum length of time to track a scan). Another example is system patch status—one operating system might report only on operating system patches, while another operating system might also include some application patches. In such a case, an organization could use multiple measures instead of one, with each measure corresponding to a different measurement method, and then combine the measures into a metric that approximates the collective values of the measures.

Many instances of problems with imprecise measure definitions have been mentioned in the security community, but to date no concerted effort has been made to exhaustively gather information on these problems, document it, and make it available to the security community. It would be much more helpful to identify the factors that organizations should consider when defining their measures than to attempt to provide a single definition for each measure. The best definition for an organization is driven by what the organization is trying to accomplish. For example, in the patching example mentioned above, an organization might be trying to gain insights on general patch distribution and installation practices to verify that all applications deemed critical by the organization have been patched or to verify that the organization’s patch management software is functioning properly (e.g. patches are installed and operate properly based on a predefined schedule).

Another common problem with the accuracy of measures is the use of qualitative measures. As mentioned earlier, data collection methods such as self-evaluation surveys often produce inaccurate or skewed results, depending on the types of questions asked. For example, if users or administrators are asked if their systems comply with the organization's policies, they are very likely to say that they do. It would be more accurate to instead use quantitative measures that assess the systems' compliance. Qualitative measures that do not have well-defined scales or units of measure can be particularly problematic in terms of accuracy. For instance, asking a user to rate the reliability of their computer on a scale of 1–5 where 1 is simply defined as "poor" and 5 as "excellent" is subjective and imprecise. A qualitative measure may be useful if each rating is defined clearly without overlap between the ratings, so that different people, when given the same information, would be likely to assign the same rating. An objective scale might be 5—no crashes or hangs in six months, 4—one crash or hang, 3—two or three instances, 2—four to six instances, and 1—more than six instances. The rating may still be somewhat subjective because it is dependent on the user's recall or because "crash" and "hang" are not defined. Nonetheless, this qualitative measure is more precise than "poor" to "excellent".

Some measures are also considered qualitative because they provide absolute counts without a context, norm, or goal. For example, a measure that indicates that 100 attacks were attempted has no context. What is the period of time? Is 100 a lot or a little? A measure that indicates that 100 attacks were attempted out of 1,000,000 incoming Web server connections adds context.

Context is very important to measures and metrics. Most measures individually have little meaning. Even the example above—the number of attempted attacks per million incoming Web server connections—does not have much meaning by itself. Is the rate of attempted attacks rising, falling, or staying steady? Have any changes been made to the organization's security controls that would change how effectively they can detect attacks or has there truly been a change in the number of attacks? Do changes in the rate of attempted attacks correspond to observations about attack trends reported by other organizations? A single measure may need to be analyzed in context with several other measures, as well as separate events such as security control changes and external trends, to determine its true significance. It would be helpful to organizations to have additional information compiled on the relationships between measures and between measures and separate events, particularly if it includes empirical information based on analyses of real-world operational environments.

Also, because cyber technology is so dynamic, the meaning of measures and metrics changes over time. For example, a measure may have shown an increase in attacks succeeding last year, and the organization determined through other measures and knowledge of external events that this was primarily due to an increase in phishing attacks. This year antiphishing technologies are deployed, but the success rate for attacks continues to increase. Is this due to improved attack techniques, improperly configured antiphishing technologies, inadequately trained users, or other factors? Next year, there may be additional factors that influence the significance of the measure, as well as different relative importances for the existing factors.

4 PROBLEMS WITH THE SELECTION OF MEASURES

Most organizations have many existing sources of security measures, automatically generated by enterprise security controls such as antivirus and antispyware software, intrusion detection systems, firewalls, patch management systems, and vulnerability scanners.

There may be accuracy issues with some of these measures, but this can still leave an organization with many existing measures from which to choose. Organizations could also create additional measures such as utilities to extract information from security logs, but there may be considerable cost in creating and deploying software and, in some cases, entire systems to collect such measures.

Some organizations collect many measures under the assumption that it is better to have more information than less information, or because it is easier to collect a lot of measures than it is to create a set of metrics and then determine which measures support those metrics. Collecting measures without evaluating their usefulness and having a plan for how to use them has several disadvantages. Firstly, it can waste considerable time and resources to collect, analyze, and report measures: only the measures that support the organization-selected metrics are generally needed. Secondly, if the measures are not selected and organized so that the dependencies between the measures are clear and accurately represented in the corresponding metrics, analysis of the measures and related metrics is likely to generate misleading results. Thirdly, it often causes people involved in the measure collection process to feel that the effort is a waste of time, because it is unclear what value there is in collecting so many measures. Another reason is that if people are allowed to choose which measures they will collect and share with others, they are more likely to collect measures that demonstrate positive results (e.g. 100% of desktop computers have antivirus software installed) than measures that demonstrate negative results (e.g. 15% of antivirus software installations are up to date).

Currently, there are many suggestions in the security community for what measures organizations should collect. However, little work has been done to determine the value of these measures in real-world operational environments, including which measures are most supportive of particular metrics. For example, suppose that an organization wants a metric for how effectively its security controls detect and stop attacks. Dozens, if not hundreds, of measures that could support this metric have been suggested by the security community, but little research has been done as to which of those measures are most closely correlated with the metric. If the characteristics of real-world operations were studied and analyzed, it may become apparent which measures are most indicative of the overall security posture and which measures are of little or no value. It might also be possible to approximate a metric by using just a few carefully selected measures.

5 PROBLEMS WITH THE USE OF MEASURES

In addition to issues with measure accuracy and selection, many organizations also face challenges involving the use of measures. Some of these challenges, such as ensuring that the selected measures support the determination of the chosen metrics, have been discussed earlier. Another common challenge is determining how to combine the values of the measures into a metric. The measures may use different units of measurement, have different scales, and have varying precision; these issues can be addressed through careful creation of equations to combine the values. Also, some of the measures may be more important than others in the scope of the metric; however, it is often difficult to quantify what weight each measure should be given. Empirical research in this area could provide organizations with a factual basis for weighting measures instead of either guessing or weighting each measure equally.

Organizations need to recognize that over time, they will need to alter their measures and metrics. Although high-level metrics may stay the same, low-level metrics need to change over time as the security posture of the organization changes. For example, an organization with relatively immature security practices may need to initially focus on measures and metrics involving its most basic security controls, such as what percentage of computers are protected by antivirus software. As the organization's security controls mature, these metrics may become less useful, and the organization may want to answer new questions, such as how effective its security controls are at stopping malware. This may require the development of new metrics and corresponding measures, and the collection of the old metrics and measures can be stopped if the organization no longer finds them to be of value.

6 COMMON VULNERABILITY SCORING SYSTEM (CVSS)

To better illustrate the current state of research on security metrics, we will examine an ongoing research effort for metrics that indicate the significance of vulnerabilities in systems. The Common Vulnerability Scoring System (CVSS) is a standard for assessing the severity of flaws in operating system and application software [1]. CVSS is composed of three sets of measures: base measures that are constant over time, temporal measures that change over time but are the same for all environments, and environmental measures that may be different for each environment. There is a different equation for each set of measures, and the result of each equation is a score—a base, temporal, or environmental score—that is in essence a metric. The measures are for particular characteristics of each vulnerability, such as whether it can be exploited remotely (over a network) and to what degree the confidentiality, integrity, and availability of a target could be impacted. The score metric is intended to give a general indication of the relative severity of the vulnerability.

The initial version of the CVSS measures, equations, and metrics were released in 2005. On the basis of feedback from their real-world use, particularly examination of empirical measure and metric data by security experts, deficiencies in the CVSS standard were identified. The measures, equations, and metrics, as well as the corresponding documentation, have all been revised to make the measures more consistent and to improve the accuracy of the metric scores. Version 2 of the CVSS standard was released in mid-2007.

CVSS is most commonly used by organizations to prioritize their vulnerability mitigation activities, such as applying patches to systems. However, researchers are investigating other uses for CVSS. For example, work has been done at the National Institute of Standards and Technology (NIST) on using CVSS to determine metrics for security-related software configuration settings [2]. Researchers at Veracode are looking at using CVSS to rate software weaknesses [3]. There is also interest in bringing CVSS scores down from the enterprise level to the individual system level so that CVSS could be used to help assess the overall vulnerability of individual systems. To accomplish this, considerable research and empirical validation are needed for applying CVSS to software configuration settings and weaknesses, as well as a new way of measuring the strength of security controls on individual systems.

7 RESEARCH DIRECTIONS

Literature contains hundreds of measures. Research is needed to validate connections between measures and security, determine correlations, and model effects. Although there is some readily accessible data, for example, National Vulnerability Database [4] and A Chronology of Data Breaches [5], such research and analysis require more and higher-quality data. State laws requiring companies to notify consumers of data breaches have the benefit of supplying data. Additional information can come from developing or articulating motivations for organizations to share information, as is the practice for business case studies or the airline industry.

Beyond measures, a secure nation needs research to understand which metrics lead to higher security, the measures supporting those metrics, and analytical methods to aggregate measures.

REFERENCES

1. Common Vulnerability Scoring System. (2008). *Forum for Incident Response and Security Teams (FIRST)*, http://www.first.org/cvss/.
2. Scarfone, K. and Mell, P. (2008). The Common Configuration Scoring System (CCSS) (Draft), NIST, NIST Interagency Report 7502, http://csrc.nist.gov/publications/PubsNISTIRs.html.
3. (2008). https://securitymetrics.org/content/attach/Metricon2.0/Wysopal-metricon2.0-software-weakness-scoring.ppt.
4. National Institute of Standards and Technology (NIST). (2008). *National Vulnerability Database*, http://nvd.nist.gov/.
5. Privacy Rights Clearinghouse. (2008). *A Chronology of Data Breaches*, http://www.privacyrights.org/ar/ChronDataBreaches.htm.

FURTHER READING

Corporate Information Security Working Group. (2005). *Report of the Best Practices and Metrics Teams*, Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, Government Reform Committee, United States House of Representatives, 17 November 2004, revised 10 January http://www.educause.edu/ir/library/pdf/CSD3661.pdf.

Dorofee, A., Killcrece, G., Ruefle, R., and Zajicek, M. (2007). *Incident Management Capability Metrics*, Version 0.1, Software Engineering Institute/CERT, http://www.cert.org/archive/pdf/07tr008.pdf.

Herrmann, D. S. (2007). *Complete Guide to Security and Privacy Metrics*, Auerbach Publications, Boca Raton, FL.

Jaquith, A., (2007). *Security Metrics: Replacing Fear, Uncertainty, and Doubt*, Addison-Wesley, Upper Saddle River, NJ.

McCurley, J., Zubrow, D., and Dekkers, C. (2007). *Measures and Measurement for Secure Software Development*, Software Engineering Institute, https://buildsecurityin.us-cert.gov/daisy/bsi/articles/best-practices/measurement/227.html.

Safety and Security Measurement Technical Working Group. (2005). *Security Measurement*, Practical Software and Systems Measurement (PSM), http://www.psmsc.com/Downloads/Other/Security%20White%20Paper%202.0.pdf.

(2008). securitymetrics.org, http://www.securitymetrics.org/.

Chew, E., Swanson, M., Stine, K., Bartol, N., Brown, A., and Robinson, W. (2008) *Performance Measurement Guide for Information Security*, NIST, NIST SP 800-55 Revision 1, http://csrc.nist.gov/publications/PubsSPs.html.

Conferences and workshops

(2008). *Third Workshop on Security Metrics (MetriCon 3.0)*, http://www.securitymetrics.org/content/Wiki.jsp.

(2008). Security measurements and metrics. *Fourth International Workshop on Quality of Protection (QoP 2008)*, http://dit.unitn.it/~qop/.

(2005). *IEEE International Symposium on Software Metrics*, http://www.informatik.unitrier.de/~ley/db/conf/metrics/ (latest year was 2005).

CROSS-REFERENCES

Countermeasures: Robustness, Resilience and Security

Attack Detection, Preemption and Response

Integrated, Enterprise-Wide Security Monitoring and Management

Techniques for Quantifying the Vulnerability of Interdependent Physical and Logical Networks

Cyber Defense