MetriCon 2.0: Second Workshop on Security Metrics

Boston, MA
August 7, 2007
Summarized by Dan Geer

MetriCon 2.0 was held on August 7, 2007, as a single all-day, limited-attendance workshop, in conjunction with the USENIX Association's Security Symposium in Boston, Massachusetts. MetriCon 2.0 was the second meeting with this name and topic, the first having been held a year before in Vancouver. The self-selected organizing committee was co-chaired by Betsy Nichols (PlexLogic) and Gunnar Peterson (Artec Group). Also on that committee were Fred Cohen (Fred Cohen & Associates), Jeremy Epstein (Software AG), Dan Geer (Geer Risk Services), Andrew Jaquith (Yankee Group), and Russell Cameron Thomas (Meritology). Dan Geer is the principal author of these notes and assumes full responsibility for any inadvertent reporting errors. The agenda and presentation slides can be seen at http://www.securitymetrics.org/content/Wiki.jsp?page=Metricon2.0.

Seventy-three people attended (compared to forty-four at MetriCon 1.0), predominantly representing industry (62) rather than academia (5) or government (6) (comparable numbers for MetriCon 1.0 were 30, 10, and 4). The meeting lasted from 08:30 until something after 21:00, with meals taken in-room, so as to maximize output—as may be reflected below.

This second such event could perhaps have benefited from more meeting time, but it is likely there will be another and, in comparing this one to the last, the amount of progress is best gauged by the sharp change from "I plan to . . ." toward "I tried this and it turned out that . . ."—which you are invited to consider a metric on MetriCon.

KEYNOTE "DEBATE"—DO METRICS MATTER?

This was not so much a debate as a point-counterpoint from two keen observers.

METRICS DO MATTER

Andrew Jaquith (Yankee Group), describing himself as Dudley Doright, simply went straight to a list of "ten" reasons why metrics matter:

1. Metrics quantify the otherwise unquantifiable.
2. Metrics can show trends and trends matter more than measurements do.
3. Metrics can show if we are doing a good job.
4. Metrics can show if we are doing a bad job.
5. Metrics can show if you have no idea where you are.
6. Metrics build bridges to managers.
7. Metrics allow cross-sectional comparisons.
8. Metrics establish where "You are here" really is.
9. Metrics set targets.
10. Metrics benchmark yourself against the opposition.
11. Metrics create curiosity.

METRICS DO NOT MATTER

Not to be outdone, Mike Rothman (SecurityIncite) started by reminding us all that it is (way) too easy to count things for no purpose other than to count them. He wanted us all to "Stop thinking like a security person, or all this metrics stuff will be a waste; you cannot measure security, so stop trying." This means that you measure, if you measure at all, not just to measure for the purpose of satisfying the counting instinct, but to make a difference. Rothman's own list of what matters includes:

1. Maintenance of availability
2. Preservation of wealth
3. Limitation on corporate liability
4. Compliance
5. Shepherding the corporate brand

Rothman went on to say, "Who cares what Jaquith's (separately published but widely quoted) 'five characteristics of a good metric' are when we already know that Rothman's own list is what really matters?"

With that, Betsy Nichols (PlexLogic) exercised her role as moderator by calling on the audience to ask questions.

DISCUSSION

First up was a suggestion that there are, in fact, metrics that speak to what Rothman was talking about, such as Apdex. Rothman answered with a question of sorts: If you don't have time to burn, then shouldn't you actually be careful what it is you are measuring? Once made, using the results of measurement takes time, but measurement for no purpose is way too easy, making useless work for

yourself and others. Plus, once you start measuring something and incorporate it into the culture of a firm, you will find it harder to stop measuring whatever it is than to have started measuring it in the first place.

Another questioner asked whether to start large or small and whether to risk too much ambition or too little. Rothman took that one as well and reminded us that unless you, the measurer, are seen as a colleague you will be seen as a crank, something he characterized as "making a deposit in the incredibility bank."

Another questioner asked, "Is it not true that metrics only matter when something can be said to be under at least a modicum of control?" Put differently, what good are metrics in a hurricane? Rothman put a new spin on the aphorism that the sum of beauty plus brains is a constant by suggesting that if all a metric does is make you look good, then it has already contributed all the value it ever will.

TRACK 1—GUNNAR PETERSON, TRACK CHAIR

- ***Security Meta Metrics—Measuring Agility, Learning, and Unintended Consequence***

  Russell Cameron Thomas, Meritology

Thomas began by reminding us of the great difficulty of our field: the mutation rate, which, of course, translates into a challenge to continuously learn. That challenge leads to his thesis that meta-metrics, the measurement of whether we are rightly measuring the right thing, is essential, as the learning demand will not recede. More fully, it is learning, agility, and unintended consequences upon which he wants to focus. Thomas distinguished single-loop learning, a control structure with a defined outcome, from double-loop learning, which adjusts the single loop's outcome. This has direct connection to the balanced scorecard idea as found in management schools.

In distinguishing puzzles, problems that have a solution, from mysteries, problems that may have no solution, Thomas suggested that meta-metrics studies focus on the latter through coverage metrics, decision effectiveness metrics, and investment return metrics. Agility meta-metrics (e.g., "Are we learning fast enough?") is richly studied in other fields, but it can be summarized here as meta-metrics for speed (such as the time between "sense" and "respond"), cost, error, and maximum response capability. Rounding out the suite is meta-metrics for discovering and mitigating unintended consequences, including familiar items such as blame shifting and excessive risk aversion, detecting the existence of these unintended consequences, measuring their significance and cost, and scoring their perversity. Thomas's bottom line is that unless your enterprise is small, simple, and static, you need at least one metric for each of learning, agility, and unintended consequences.

A questioner raised the possibility of studying latency in the agility domain with Fourier analysis. As to "Who is doing this learning?" Thomas suggested that it be the enterprise risk team, not individual employees. As to the problem of indirect costs, Thomas referred the questioner to the "total cost" section of Thomas's Web site. The idea of "malicious compliance" came up (e.g., Accounting saying, "Security is important but costs must decline"). Thomas suggested that the most common finding for the root cause of a disaster is that of a "failure of imagination."

- ***Security Metrics in Practice: Development of a Security Metric System to Rate Enterprise Software***

  Fredrick DeQuan Lee and Brian Chess, Fortify

Lee described the "Java Open Review" during which the Fortify team examined 130+ open-source projects for both quality and security defects. Given that many of these projects overlap to some degree in function, this examination naturally led to the question of which project is better.

That question is, even given this work, unsolved, as the downstream risk is dependent on deployment context as well as the existence of defects. In their estimation, risk assessments need an enumeration of either threats, vulnerabilities, and controls or event probability and asset value and, given that static analysis only uncovers vulnerabilities, it cannot yield a risk metric.

Static analysis can, however, measure defects in source code and benchmark software components, use objective and repeatable measures to improve software over time, and feed into any existing risk management system. The Fortify SCA product used in this work can provide most of the base information for a CVSS score, as well as code volume, cyclomatic complexity (per function), and defect densities along several axes.

Fortify's customers, as do perhaps all metrics end consumers, want condensed thumbs-up/thumbs-down views, and Fortify chose to copy the mutual fund star system (much as did the OWASP group). Those stars are:

* No remote/SETUID vulnerabilities
** No obvious reliability issues
*** Follows best practices
**** Documented secure development process
***** Passed independent security review

Lee points out that this rating system is not without flaws: it is harsh, there is some subjectivity, and the introduction of a tiering forces some compromises because of inexact ordering. Nevertheless, such a scheme can be directly used as screening criteria (e.g., "Show me 2-star, mid-size, shopping cart software"), a comparator (e.g., "How does this set of 1-star components compare?"), or, as described earlier, as an input among many to an existing risk management model, if any. Going forward, it will be important to validate this method against the closed-source world and to compare this method's hard numbers to (the accumulation over time of) security auditors' reports.

The direct question of "Are there any open source projects with a nonzero number of stars?" revealed a few (e.g., Tomcat). The similarly expected question, "How do you handle false positives?" was that people remain essential to this. One observer noted that as new attacks appear old ratings lose meaning, which Lee said had no solution other than to say that as of such-and-such a date the rating was X and to retain in a public fashion the rule set that was in use as of that date. Some questions on consistency and rigor were raised, but the truthful answer is that they were early, though Lee did point out that reranking old work with successive new rule sets would shed some light on the consistency questions (over time).

■ *A Software Security Risk Classification System*

Eric Dalci and Robert Hines, Cigital

Dalci described the purpose of the Risk Classification System (RCS) as estimating an application's potential risk with respect to other systems in the portfolio and determining what SLDC actions to require for given risk levels. This would yield, as RCS outcomes, the ability to prioritize (impose an ordinal scale) and an indication of where midcourse corrections in ongoing development should go. As with all efforts to summarize risk, there are separate foci on business risk and technical risk.

In producing the RCS, Dalci and Cigital dropped cyclomatic complexity (because it was not clear how to correct for language differences), process-related metrics (since organizations rarely are internally consistent in how they apply security processes), and generally any factors that contribute expensive or squirrelly answers. Roughly speaking, their strategy involves weighted aggregation of various measurable characteristics and then use of the weighted sum as a score for portfolio segregation. Dalci listed the systems that tended to have a high score as:

- Independent security review systems
- Web-facing systems
- Large code-size applications
- Complex applications
- New applications

and those with a low score as:

- Low user count and/or internal applications
- Low corollary (downstream) impacts
- Small code-size applications

His slides displayed the weights used and the correlation achieved with aggregate scores.

In response to a question, Dalci clarified that no dependent downstream applications would be scored as low, while more than four such downstream applications would be judged as high. Another audience member suggested that adapting data gathering to the measurement system sounded consistent with Thomas's double-loop learning construct. Dalci confirmed that the aim of this effort was that the method be fast and light. He also described the correlation figures as essentially a measure of cascade failure.

Another questioner suggested that the business and technical risk views would be good to summarize as a 2x2 table. In response to a question as to where revenue factors in here, Dalci said that that is a subject for future work. Another respondent suggested that using Dalci's method to get a probability of failure makes this similar in style to a credit risk score.

TRACK 2—JEREMY EPSTEIN, TRACK CHAIR

■ *Web Application Security Metrics*

Jeremiah Grossman, WhiteHat Security

Grossman stated his bias with respect to security metrics, namely that bad things are generally unmeasurable. As of today, there are 128 million Web sites and these sites are accessible to 1 billion people. We will all acknowledge that a percentage of these can be hacked and that when hacked there are consequences. Grossman's study looked at the composite outcome of 20 months of weekly remote black-box assessment of hundreds of the largest and most popular Web sites (in all sectors), all of which are custom Web applications without well-known issues. The threat classification from the Web Application Security Consortium (WASC) was used as the baseline. His results are that 7 of 10 Web sites have "serious" vulnerabilities, and he assessed the likelihood that a Web site has a vulnerability of a given severity.

Grossman went on to say that, putting aside infrastructural matters such as PHP, cross-site forgery remains very difficult to scan for, and new ways to evade XSS filters keep showing up. HTTP response splitting is, he believes, the coming thing and must be watched carefully. He provided a number of looks at what his data shows, such as cross-tabulating the kinds of flaws found with their severity, ranking the filename extensions most involved, and showing that the kinds of flaws present do vary by industry vertically.

Perhaps more hopefully, the custom Web applications that are more secure come from development environments where the security configs are actually turned on, have a software development life cycle that does include security in a formal way, and prioritize the remediation of vulnerabilities in a rational fashion. Looking ahead, Grossman particularly wants to continue comparisons across verticals and technology and examine the rate at which problems reappear.

A questioner asked whether one can include Web site complexity or size in the vulnerability rankings; Grossman does not believe that complexity is related to security: Security comes from the code being beat on. Another thought Grossman's SQL injection numbers were low, and Grossman confirmed that they could be hiding issues in that space. Grossman did not yet have prevalence by platform data but it is coming, and he will also be introducing trending. A hard problem is in environments where part-

ner Web sites function as an apparent whole; much work needs to be done on how to characterize the risk in such settings.

- ***Operational Security Risk Metrics: Definitions, Calculations, and Visualizations***

  Brian Laing, Mike Lloyd, and Alain Mayer, Redseal Systems

Mayer's work includes many graphics aimed at making objective operational security metrics and visualizing them in ways that make for real communication. One part of his visuals shows tracing a network path through a set of servers from the outside (Internet) to a DMZ to an internal host that becomes compromised, thus leading to a general compromise. With that as a lead-in, Mayer stated the goals and nongoals for a metrics program. The main idea is that hierarchies are natural, that cascade failure is their downside feature, and thus that drill-down for root cause analysis is a high-value capability. Mayer suggests treemaps as a well-matched tool for this. To illustrate this point, the reader will have to consult his materials, as they are visually rich. Mayer's main point, and one on which he was questioned as well, is that treemaps are effective in conjunction with more traditional topologic visualization, but that some people take to treemaps immediately and some do not.

Mayer called his metrics "opinion-based math" and wondered about the absence of user-side pushback—do the users get it, or do they not? Nevertheless, mapping of cascade failure to the hierarchies in which they occur with a drill-down-friendly visual summarization does seem to be an advance.

A questioner asked whether using absolute-risk or delta-risk is better. Mayer said that delta-risk might be more informative. Another questioner asked whether this might be aggregated over industrial verticals, which Mayer acknowledged but thought to be too early. Mayer responded to "Where does the source data come from anyhow?" by suggesting that firewall configuration files and scan data suffice.

- ***Metrics for Network Security Using Attack Graphs: A Position Paper***

  Anoop Singha, NIST; Lingyu Wang and Sushil Jajodia, Center for Secure Information Systems, George Mason University

Singhal described his group's motivation by contrasting the typical qualitative questions about a database's security (e.g., "Is that server secure from intruders?") with the quantitative questions that are actually needed (e.g., "How secure is that server?"). He sees the challenge as one of composing a variety of measures into one metric.

He focused on attack graphs, annotated with both point probabilities (of exploit of a given flaw) and cascade probabilities (of reaching through this flaw to the next host, beginning with an attacker at node 0). The point is that such graphs can make clear the value returned in hardening (blocking the exploit of) any given node in such a graph. Singhal suggests that such mechanisms of analysis are common-sensical and can be generalized, which he proposes to do as further research. Questioners asked about the level of effort required to set the probabilities in such graphs, whether vulnerabilities were statistically independent, whether this was scalable, and how it meshed with business needs.

TRACK 3—ADAM SHOSTACK, TRACK CHAIR

- ***Software Security Weakness Scoring***

  Chris Wysopal, Veracode

The purpose of Wysopal's work is to develop a standardized set of software security analysis techniques addressing inter-rater and test-retest reliability, and with actionable outcomes. Wysopal's method builds on what is available at the outset, the Common Weakness Enumeration (CWE) and the Common Vulnerability Scoring System (CVSS), noting that all current techniques have serious levels of false positives and false negatives.

Wysopal's method is layered and should be looked at in the original, with the logical outcome of being able to rank weaknesses in the sense of "How likely is it that bad things will come from this weakness?" The ranking is thus a contributor to security decision-making, and the metric proposed is thus well worth further effort.

Wysopal suggests that the CVSS Environmental Score can be used unchanged, although, of course, this implies foreknowledge of the deployment environment into which software will go. He further suggests some plausible goals:

- Standardized false positive rate testing
- Possible use of data and control flow between taint source and weakness
- Addition of false negative rates, moving from "badness" score to "goodness"
- Empirical field testing

Questioners asked which version of CVSS Wysopal was using (version 2) and whether the appearance of new attacks would change the risk scores he computed. Wysopal thought that the appearance of new attack methods was likely a research-grade problem at this time.

- ***Developing Secure Applications with Metrics in Mind***

  Thomas Heyman, Christophe Huygens, and Wouter Joosen, K.U. Leuven

Building on their work presented at Metricon 1.0, Heyman et al. set out to answer, "How secure is my application?" In their prior work, a "pattern" is the observable connection between the core of one's computing environment and the ecosystem in which it lives, leading to ratio scores such as the number of firewall invocations versus the number of service invocations, or the number of guards versus the number of access points for each component. With this new work, they are trying to use patterns to piggy-back security metrics into applications.

In this case, domain-specific security requirements are assigned domain-independent security objectives, and design involves composing systems from primitives, such as Accountability through Authentication plus either Auditing or Non-Repudiation, and, in turn, Auditing through both an operational interceptor and a secure logging facility. Just as the building blocks are composed into the final system, the measurements that come with each building block are rolled up into a final metric. As in the aphorism "A chain is only as strong as its weakest link," this roll-up process will propagate minimum values upward, such as if Auditing decomposes into both an operational interceptor and a secure logging facility; whichever of those two is the least reliable will determine the reliability of the Auditing function.

Heyman expects a proof of concept where sensitivity analysis can be done on the dependency graph and, perhaps, to automate the integration of metrics into the code base of the building blocks. Questions went right to the hard parts, such as "Where might the numbers come from?" Heyman said they are assigned heuristically and can be thought of as relative capabilities. Confidence scores seem eventually possible, as would sensitivity analysis. Although multidimensional methods are not in place now, they may be necessary if risk is taken into account.

- ***Correlating Automated Static Analysis Alert Density to Reported Vulnerabilities in Sendmail***

  Michael Gegick and Laurie Williams, North Carolina State University

The security metrics arena has many parallels to the field of reliability, such as the similarities between fault-prone components and vulnerability-prone components and between failure-prone components and attack-prone components, making borrowing from the latter field useful. The research objective of Gegick and Williams's work is to predict vulnerability and attack-prone components from static analyzer alerts.

This objective leads them to a general linear model with a Poisson distribution for the number of vulnerabilities per component based on the alert density for that component. Although Gegick and Williams scanned (with Fortify's SCA) ten releases of Sendmail totaling 1,000 files, they still had few data points when it came to vulnerabilities per se. With that caveat, they did show a relationship between SCA alert density and the number of vulnerabilities per file but found no relationship between SCA alert density and the number of exploits per file.

A questioner led Gegick to describe how the SAMATE project at NIST is a similar effort to this work and to note how version changes in Sendmail make double counting likely. Gegick is working on other targets besides Sendmail, as the main issue at this stage is getting more data. He hopes that in due course he will be able to publish correlations between vulnerability density and the alert density from Fortify.

PRACTITIONER PANEL—BECKY BACE, TRACK CHAIR AND MODERATOR

Brad Freeman of GE GIS Security Services, Shambla Naidoo of Wellpoint, and Ed Georgia of Booz Allen Hamilton's Information Security Practice described how they use metrics to make better decisions. The panelists opened with a few remarks.

Freeman began with the desirables for a metrics program within a firm the size of GE: simple, flexible, and hierarchical. Their program is roll-up oriented with a homegrown built around the products of ClearPoint Metrics. The basic issues in building any metrics program are:

- What are we measuring?
- Beware of poorly defined metrics and poor measurement systems
- Why are we measuring it?
- The "So what?" factor and tying metrics to business benefits
- How are we measuring?
- Manual vs. automated, actionable reports

In response to a question, Freeman said that comparison across departments is valuable and helps justify a metrics program.

Naidoo also began with a set of basic questions:

- With whom are we communicating?
- What is the message?
- Why is it important to hear?
- What do the numbers mean?

In so many words, she stressed that the top of an organization is populated by people who are overwhelmed by distractions, and thus brevity will be a key factor in getting through. She made several like points:

- Messages must be aligned with corporate priorities.
- Metrics will not get you an audience with the Board of Directors.
- Clarity for risk profiles is essential.
- You must show ROI and/or risk reduction if you are to be heard.

A questioner asked if, in so many words, this was selling, and Naidoo said that at the top everything is about selling. When asked whether that selling is just a matter of FUD, Naidoo reminded us that when fear declines so does funding. A third questioner asked whether the numbers are pushed on the management committee or pulled from Naidoo's team. She suggests that you ask top management, "What are your problems?" and speak only to their answer; that is all they will listen to anyway.

Giorgio stated that measurement perturbs a system and, as such, you must put metrics in the right hands. In government, the reason you measure is to subsequently acquire dollars. In business, the reason you measure is to drive di-

rection. Visualization matters because visualization, such as in dashboards, is what drives tactical decision-making. He challenged the audience to ask themselves, "Whom do we serve?" and, in that light, reminded the audience that metrics do not a compliance program make.

A questioner took this to heart and asked, "So what are the 'go-to' numbers?" Giorgio said that the Board of Directors wants to know, "Am I safe?" with a strong emphasis on the "I." With that in mind, Giorgio pointed out that if all you are doing is counting something, then that is not Board-worthy. He also pointed out that, in government, certification and accreditation only cost money—there is no positive return on the investment in them.

A questioner asked about the government point, whether there was a way to boil down the mix of program dollars, other resources, head count, and so forth. Giorgio said no, and that that is why he (we) are not welcome, and that metrics will only be useful as a backstop in an argument.

In Bace's view, it is time to rethink how we practice. As an industry, we are now into a period of specialization, and only in like specialization can our metrics be meaningful. An unanswered question was raised about how this guides the particularly vexing problem of counterparty risk, where the trading of data with counterparties endangers both sides of the transaction.

- ***Off-Program Comments***

  Adam Shostack

Shostack argued that breaches are great for metrics programs because they create sources of information with very low levels of bias. He referred all to two sites, http://attrition.org/dataloss and http://etiolated.org/.

- ***Debate: Stump the Chumps***

  Russell Thomas, Meritology; Mike Rothman, SecurityIncite; Pete Lindstrom, Spire Security; Andrew Jaquith, Yankee Group

Rather less organized than other interactions, the "chumps" took questions from the audience entirely. The present author regrets that he could not make enough sense of what followed to make a useful addition to this digest.