MetriCon 1.0

Vancouver, B.C., Canada

August 1, 2006

Summary by Dan Geer

[This material was excerpted from the digest found at www.securitymetrics.org.]

MetriCon 1.0 was held on August 1, 2006, as a single-day, limited-attendance workshop in conjunction with the USENIX Association's Security Symposium in Vancouver, British Columbia. The idea had been first discussed on the securitymetrics.org mailing list and subsequently an organizing committee was convened out of a lunch at the RSA show in February 2006. There was neither formal refereeing of papers nor proceedings, but there is both a digest of the meeting and a complete set of presentation materials available at the www.securitymetrics.org Web site. Andrew Jaquith (Yankee Group) was chair of the organizing committee, the members of which were Betsy Nichols (ClearPoint Metrics), Gunnar Peterson (Artec Group), Adam Shostack (Microsoft), Pete Lindstrom (Spire Security), and Dan Geer (Geer Risk Services).

KEYNOTE ADDRESS

- *Resolved: Metrics Are Nifty*
  Andrew Jaquith, Yankee Group
- *Resolved: Metrics Are Too Hard*
  Steve Bellovin, Columbia University

Andrew Jaquith opened MetriCon 1.0 by pointing out that other fields have their bodies of managerial technique and control, but digital security does not, and that has to change. He presented his list of what features are included in a good metric: It must (1) be consistently measured, (2) be cheap to gather, (3) contain units of measure, (4) be expressed as a number, and (5) be contextually specific. Jaquith argued that this all breaks down to modelers versus measurers. Modelers think about how and why; measurers think about what. He was quick to admit that measurement without models will not ultimately be enough, but "let's get started measuring *something*, for Heaven's sake."

Steve Bellovin countered with the brittleness of software and thus the infeasibility of security metrics. Beginning with Lord Kelvin's dictum on how, without measurement, your knowledge is "of a meagre and unsatisfactory kind," Bellovin said that the reason we have not had much progress in measuring security is that it is in fact infeasible to measure anything in the world as we now have it. We cannot answer "How strong is it?" in the same style as a municipal building code unless we change how we do software. Because defense in the digital world requires perfection and the attacker's effort is linear in relation to the number of defensive layers, this brittleness will persist until we can write self-healing code. So his challenge: Show me the metrics that help this.

Lindstrom argued that Bellovin's reasoning did not show that metrics are impossible but rather that they are necessary. Another attendee asked, "So what if I agree on bugs being universal and it only takes one to fail a system? The issue is: How do we make decisions?" Butler agreed, saying that if it's that hopeless, then why do security at all? Epstein reminded all that formally evaluated systems still have bugs, too. Another attendee suggested that we can borrow some ideas from the physical world, especially relative measurements such as "this is safer than that." An attendee said that there are certainly things you can measure, if for no other reason than to avoid stupid things. Another attendee cared about data, not software that handled data: "I want to know about changes in data state."

SOFTWARE SECURITY METRICS

Gunnar Peterson, track chair

■ *A Metric for Evaluating Static Analysis Tools*

Brian Chess and Katrina Tsipenyuk, Fortify Software

Peterson began with a call for rethinking what granularity we need if metrics are to be meaningful, not with regard to "system" or "security" but, rather, to C/I/A (confidentiality, integrity, and authentication).

Chess proposed a weighted composite score reflecting the orthogonal interests of the tool vendor, the auditor, and the developer and showed some preliminary results of applying this to a mix of tools and applications. He displayed real data from real work.

■ *An Attack Surface Metric*

Pratyusa Manadhata and Jeannette Wing, Carnegie Mellon University

Manadhata posited a formal framework intended to find an answer to "Is the attack surface of A more serious than that of B?" Using the ratio of damage potential to attacker effort, he displayed several examples in each of which he manually annotated the source code and analyzed the call graph of the application using off-the-shelf tools. The question on the table is whether the number of vulnerabilities is or is not correlated with this attack surface metric.

■ *"Good Enough" Metrics*

Jeremy Epstein, WebMethods

Rather than argue about which numbers it makes sense to collect, Epstein suggests gathering as many as you can and only then decide which make sense. Some numbers have only a distant relationship to vulnerabilities, some are merely retrospective, and some tend to too many false positives. Epstein suggested that ratios of Cowan's "relative vulnerability" sort are valuable, though Epstein's true desire is the security equivalent of "leading economic indicators."

■ *Software Security Patterns and Risk*

Thomas Heyman and Christophe Huygens, University of Leuven

Huygens argued that we should attach metrics to security patterns, where a "pattern" is the observable connection between the core of one's computing environment and the ecosystem in which it lives. He is interested in ratio scores such as the number of firewall invocations vs. the number of service invocations, or the number of guards vs. the number of access points for each component. Preliminary results indicate that this approach is feasible; the aim is to craft indicators to use in the system design space.

■ *Code Metrics*

Pravir Chandra, Secure Software

Chandra focused on remediation metrics—metrics that help (and assess) getting better and better. His main tool is a 4x4 matrix crossing severity (Critical, Error, Warning, Informational) with review state (Unknown, Known, Accepted, Mitigated). For each review state, he plans to use capture-recapture or capture-for-removal metrics to estimate flaw count and then look at changes in market share by severity to track progress. Chandra proposed correlating this with software complexity metrics (McCabe Cyclomatic, System, and Information Flow Complexity).

ENTERPRISE AND CASE STUDIES A

Adam Shostack, track chair

Adam Shostack led off with a discussion of "Enterprise Case Studies: Substitute for Ongoing Data," followed by Butler on "What Are the Business Security Metrics?"

■ *Data Breaches: Measurement Efforts and Issues*

Chris Walsh

Walsh began with the dates of (U.S.) adoption of data-breach laws and asked whether online breaches are a significant source of ID theft. Most studies focus on firm-level impact on income or stock price. To Walsh, such studies lack enough reach to establish causality or even to establish whether the public is simply becoming inured to breaches. More than anything else, Walsh was outlining what it is that we don't know and we will need to know, in other words, a research agenda.

■ *The Human Side of Security Metrics*

Dennis Opacki, Covestic

To Opacki, the point of any metric is to change behavior, but behavior change is prone to pitfalls. Opacki tied together findings in evolutionary psychology, social psychology, and behavioral economics, noting that human intuition has low energy cost and runs in parallel, whereas human reason has high energy cost and runs single-threaded. Focus on scales that people gauge intuitively, keep the number of metrics small, do not neglect entertainment value, and give bad news first. Measure the impact of your delivery before and after, express everything in dollars where you can, and use plain language. Remember, it is better to be vaguely right than precisely wrong.

■ *No Substitute for Ongoing Data, Quantification, Visualization, and Story-Telling*

John Quarterman and Gretchen Phillips, InternetPerils

Quarterman demonstrated how his firm handles phishing attacks, making an argument for data aggregation. He used animation to illustrate a time series of complex interconnection paths. Quarterman is squarely in the observation (measurement) camp, not the simulation (modeling) camp, and he suggests collecting data when you do not yet need it so that when you do need it baselines are already in hand, such as on the other side of your firewall. Some discussion followed on what measured level of misbehavior an ISP needs before it can break its contract to support a phishing site.

■ *What Are the Business Security Metrics?*

Shawn Butler, MSB Associates

For Butler, business decisions are what security metrics are about. What we must have are frequency and impact if we are to get at true cost, and cost is the basis of business decision-making. Without impact, there is no importance to management. Irrationality is involved everywhere, driven, ironically, by standards of due care and the perception thereof. Butler does endorse the idea of decision support, but she notes that although requests are not coming down from on high, massive amounts of data are moving up and, worse, data represents lots of information about frequency (number of probes, viruses, unauthorized this or that) but nearly no information about impact (cost). She feels that impact is the hardest question to answer and that not assessing impact means there is no feedback between effectiveness and investment.

ENTERPRISE AND CASE STUDIES B

Betsy Nichols, track chair

Nichols began the session by showing that maturity of security metrics deployment and market capitalization are uncorrelated. As session lead, she set out the core question: Why are metrics so hard? Her answer focused on three issues: vast and unclean data, a lack of consensus on indicators and models, and difficulty in packaging results.

■ *Leading Indicators in Information Security*

John Nye, Symantec

With his easy access to work at the Symantec Attack Center, Nye undertook to show what leading security indicators might look like. Beginning with the results of 449 remote penetration tests, he calculated a "vulnerability score" and "vulnerability saturation." Dividing his data set into quartiles by saturation, Nye showed an expectably sharp rise in vulnerability saturation from quartile to quartile. From that, he was able to identify specific vulnerabilities that might serve as leading indicators.

■ *Top Network Vulnerabilities Over Time*

Vik Solem

Solem used a similar data set limited to Nessus scans in a contiguous timespan, to identify the top 10 vulnerabilities over the study interval. Questions from the attendees mainly concerned details of data sources and methods. Using the Symantec Threat Report, Solem found no correlation between attacks and Nessus plug-in IDs, but there is correlation between attacks and what is in the Qualys "Laws of Vulns" report, although, as Opacki remarked, any tool including Nessus could be scanning for the wrong things.

■ *IAM Metrics Case Study*

Andrew Sudbury, ClearPoint Metrics

Sudbury confirmed that real work is hard; you must start with real goals and, within that, identify what is it that you do not know. His measures are designed to determine whether you are in control of your controls, and he confirmed that business value comes from fusing multiple data sources. Discussion was brisk: Kirkwood suggested that Sudbury add targets to his graphs of trend data. Jansen asked how one would confirm that a help desk clearance score is actually clearance and not just somebody skipping work. Blakley asked, Which of the following should be considered true? (1) Management is dumber than technical staff; (2) management and tech staff want to see different things; (3) you cannot give management bad news. Daguio said such tools let managers decide whether to ever give a particular team a project again.

■ *Assessment of IT Security in Networked Information Systems*

Jonas Hallberg and Amund Hunstad, Swedish Defence Research Agency

Hallberg made the insightful observation that, although system properties control the security level and the security level controls consequences, the security level is not measurable, whereas system properties and consequences are. Ergo, something that bridges the gap between system properties and likely or potential consequences has to be crafted. The Swedish Armed Forces uses five high-level security properties: access control, security logging, protection against intrusions, intrusion detection, and protection against malware. Saaty's "Analytic Hierarchy Process" was then used to differentially weight 20 low-level properties related to access control. The relationship between the properties was interesting and useful and was close to Bellovin's comments made at the beginning of the day.

GOVERNANCE

Dan Geer, track chair

Geer set the tone for this session by simply declaring that the only metrics that matter are those for decision support in risk management.

■ *Model Concepts for Consideration and Discussion*

Bryan Ware, Digital Sandbox

Ware described how his firm calculates the U.S. Department of Homeland Security's allocation of grant dollars to municipalities. Before Ware's involvement, the criterion of per-capita dollars was used, which is fair but useless. But after Ware's involvement, risk-centric dollar allocation was used, which is easier said than done. The first step was to require management plans from states and cities that respond to the risk measures. Because 2x2 tables show the decisions you are making, Ware's firm used 17 sets of 6 experts each to work out criteria that set thresholds between high and low effectiveness (of proposed dollars spent) and high and low risk. Ware demonstrated how this was done, and subsequently how dollars were allocated—first to quadrants, then within individual quadrants. In the low/low quadrant, a minimum amount of money is used. In the high-risk but low-effectiveness quadrant, the two obvious East Coast U.S. cities were the most problematic. Choosing between low risk/high effectiveness and high risk/low effectiveness was hardest. Most money went to high effectiveness, which got Ware's firm raked over the coals. Discussion was brisk.

■ *Mission and Metrics from Different Views: Firm/Agency, Industry, and Profession*

Kawika Daguio, Northeastern University

Daguio reminded all to "Do no harm" as we introduce new metrics, that accountability matters, and that separating risk and compliance is essential. Compliance is more important than security's C/I/A requirements. A lot of what banks do is imposed on them, and the change from a compliance model to a risk model is a breath of fresh air. Daguio says that we should use nominal and ordinal measures to avoid bad effects and that we should not do interval or ratio scales because those invite comparison and hence organizational interference. Getting information sharing will require competitive, policy, technical, and political reasons for doing so, or it simply won't fly. Daguio was clear; although we are about metrics, these metrics do not exist in a vacuum nor are the recipients of the metrics necessarily going to be good-hearted and forthright. Discussion was again brisk.

■ *Measuring Information Security Risk*

Bob Blakley, Burton Group

Blakley began with a formal definition intended to disambiguate a measurement from a metric, and to look at metrics with an eye to finding "normal limits" and thus to act when you are outside them. In short, a measurement is something you take; a metric is something you give. He argues that we are not measuring risk, which is probability times impact. Instead of probability, we have to use game theory, and instead of measuring the probability of bad things, we have to measure consequence(s) of those bad things. Further, you use game theory to measure your opponent's goals as well as your own, which is a key point. Blakley illustrates this with a 2x3 matrix aimed at decision-making: high/low impact versus whether to mitigate, mitigate and recover, or recover alone. Blakley also pointed out that, for decision-making, correlates of risk are just as good as direct measures of risk, using as his example that while blood pressure, temperature, and pulse rate may not make you ill it is hard to make you ill without changing one or more of those three measures. He suggested we should find and be happy with such correlates in our sphere. There was then some discussion of frequentist versus Bayesian approaches and whether a bimodal probability distribution (1 x 10^6 vs. 10^6 x 1) doesn't make any probabilist approach impossible. Quarterman asked about risk aggregation, and Butler reminded all that decision analysis is not about the "right" decision but about the "informed" decision, a meaningful difference. Geer and Blakley agreed that there is no probability distribution for a sentient opponent, so pure probability cannot be the answer.

■ *Information Assurance Metrics Taxonomy*

Wayne Jansen, NIST

Jansen showed one slide summarizing the taxonomy work of Vaughn et al. Jansen described himself as a novice in the metrics area and asked the audience to consider a number of questions drawn from the taxonomy. Does there exist somewhere a set of well-established metrics and measures on which a new organization should be focusing its initial efforts? The apparent discontinuity between strategic efforts and tactical ones leads one to ask whether there is a way to bridge the gap. What kinds of things need to be done to advance the state of the art? Do we even know where we want to go? Ware answered that two of the most fascinating are the FICO (Fair Isaac) score, which made it possible to have instant credit decisions, and the KMV-Merton model to predict likelihood of default for corporations.

Daguio, as a banker at that time, asked that we all please not do something that wrenching again. He said that he had exhausted all the mechanisms he has for scoring security or something; the corporate end result is always to find a way to kill projects. An attendee asked whether it is a two-player game, or is game theory just intrinsically easier. Blakley answered that games are just as challenging and that what is going on now is, at least, a two-player game as illustrated by Microsoft's first Tuesday security drill and its monthly sequelae. Second, infosec is an economic game and not just a technical game. More discussion followed.

DINNER/RUMP SESSION

Three unscheduled presentations rounded out the day: Leversage on "The Security Incident Database," Ozment and Schecter on "Does Software Security Improve with Age?" and Lindstrom on "Security Metrics."

Leversage observed that target of choice losses vastly exceed target of chance losses, that good old wiretapping is on the rise, that infected laptops as a transmission mechanism are very much on the rise, that human intelligence (HUMINT) is still the main source of information, and all in all his world is very much like the intelligence community. There is a growing demand from potential consumers, and it is private in every way.

Ozment described a fine-detail, time-series look at the history of OpenBSD. As this was a full conference paper with overlapping relevance to MetriCon, this summary is brief: Software does improve with age and is thus, as the title asks, more like wine than milk. As an inspired use of security metrics, this is a quotation from the paper's summary:

"We found statistically significant evidence that the rate of foundational vulnerability reports decreased during the study period. We utilized a reliability growth model to estimate that 67.6% of the vulnerabilities in the foundation version had been found. The model's estimate of the expected number of foundational vulnerabilities reported per day decreased from 0.051 at the start of the study to 0.024."

Lindstrom made a number of points about risk, using a number of Venn diagram examples of how to calculate varieties of risk. In Lindstrom's view, risk fluctuates the way a financial index like the S&P 500 fluctuates; as such, quantifying risk necessarily requires an actuarial tail (i.e., you calculate risk by looking at incidence and/or prevalence of activities in the past). That said, his examples are worth examining closely.

In summary, 44 people attended, predominantly representing industry (30) rather than academia (10) or government (4). Altogether the meeting lasted about 12 hours and ended on that note of happy exhaustion that marks a successful event. Not bad as a first try, and if you believe that imitation is the sincerest form of flattery, then MetriCon is already being flattered in more ways than one. If you want to be involved in this area, visit www.securitymetrics.org. Thanks go to USENIX for continuing its tradition of putting its trust in experiments.