MetriCon 4.0 Digest

Dan Geer

1. Background

MetriCon 4.0, entitled The Importance of Context, was held on August 11, 2009, as a single day, limited attendance workshop co-located with the USENIX Association's Security Symposium in Montreal, Quebec. The name MetriCon 4.0 reflects that this was the fourth meeting with this name, topic, and format; previous meetings were in Vancouver, Boston, and San Jose. The organizing committee was self-selected, and was chaired by Jennifer Bayuk (Independent Consultant) with Warren Axelrod (Financial Services Technology Consortium), Fred Cohen (Fred Cohen & Associates & California Sciences Institute), Lloyd Ellam (SigmaRisks), Dan Geer (In-Q-Tel), Andrew Jaquith (Forrester Research), Wayne Jansen (National Institute of Standards and Technology), Gene Kim (Tripwire), Gunnar Peterson (Arctec Group), and Chris Walsh (SurePayroll). Dan Geer is the principal author of these notes.

Forty people registered for MetriCon, representing industry (28), the non-profit sector (5), academia (3), independent consultants (3), and government (1). 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.

2. Baseline Scoring Methods

Reproducible Measurement as a Foundation for Security Assessment Metrics	—	Nye
Orbitz SCAP Metrics	—	Bellis

2.1. Reproducible Measurement as a Foundation for Security Assessment Metrics

John Nye (Independent Consultant)

Nye's central point is that despite any inherent fuzziness, security measurement, *per se*, must be(come) reproducible. Just as we are prepared to grade figure skating or humanities papers, and then make ordinal-scale decisions based on those grades, in security measurement "subjectivity should be baked into the measurement standard, not interpreted upon application." In other words, measuring everything is inessential and measuring risk is a distraction; measuring Control Quality is, however, both essential and core. It is the grading process that needs to be routinized to the point of reproducibility, keeping in mind that the grades are not findings.

Nye then proposed a three-part process of grading Control Quality (think Capability Maturity Models), adjusting the Control Quality grades for complexity of the environment from which they were derived (complexity being the enemy of security), and further qualifying the Adjusted Control Quality based on the Veracity of the measurement process (self-assessment at the low end increasing in Veracity with layers of competence and review). Nye invited help in the form of a working group to the weights & measures involved. Judging by the discussion, attendees were receptive to the ideas and some were perhaps willing to advance the work with Nye.

2.2. Orbitz SCAP Metrics

Ed Bellis (CISO, Orbitz)

Bellis began by acknowledging that Orbitz is a *prima facie* example of complexity, which is to say that a security program there does face the tendency to insecurity that comes from complexity. His work is a work in progress centered on SCAP, the Security Content Automation Protocol (U.S. National Institute of Standards & Technology SP 800-126), where SCAP is itself a synthesis of several other standards. Accordingly, Bellis described how he is gluing together

CPE: Common Platform Enumeration

CVE: Common Vulnerability Enumeration

CVSS: Common Vulnerability Scoring System

WASC-TC: Web Application Security Consortium Threat Class

CCE: Common Configuration Enumeration

XCCDF: Extensible Configuration Checklist Description Format

by walking through a workflow example. His central point is that in the face of complexity, which Orbitz abundantly has, automation is flatly essential and that that automation has as its purpose action plans and their management.

Bellis noted one aspect of rising complexity: that less and less of the environment is subject to an effective snapshot. A questioner asked if the standards involved are ever found wanting and Bellis said that, for example, that he has to fill in where CPE has gaps, and that that is to be expected with a constantly moving target. In response to a different questioner, Bellis said that work done to date represents the investment of a year of his development team.

3. Measuring Impact

The Ugly, The Bad, and The Good	—	Ellam
Metrics for Detecting Compromised Systems	—	Tenginakai

3.1. The Ugly, The Bad, and The Good

Lloyd Ellam (V.P., Risk Science and Innovation, Sigma Risk Management)

Ellam began by declaring the mission of his work, *viz.*, to Identify, Quantify, Mitigate, and Transfer risk. Within that focus on risk, Ellam reiterated that Good Enough beats Perfect and that the quantitative challenge (and hence the effort to be expended) is in determining what is good enough. He further defined Security as a state and that state as an asset, an intangible asset to be sure, but nevertheless as an asset it can be valued.

Ellam reminded us that if threats to security are dynamic (and they are), then countermeasures must be dynamic. By the same logic, if one's countermeasure are static, then they will not mitigate one's security threats. One might almost say that the state of security is when surprises are rare and bounded.

It is the external environment, however, that is in command of surprise and what grade of bounds are required to avoid secondary surprises. On the one hand, regulators increasingly want proof that "X cannot happen" while the capital markets are ever depressing what level of risk can, in fact, be transferred (thus increasing. The former raises costs for all firms, risk-engaged or not, and the latter forces greater risk retention.

After explaining some visuals (see slides) on process, Ellam described three recent engagements which, per the title of his talk were "The Ugly, The Bad, and The Good." A questioner asked whether, as a practical matter, "Is it harder to estimate the cost of reconstruction or the current asset value?" which Ellam said does not have a context-free answer. Another questioner asked the simpler question, "How much do you spend on prevention?" which Ellam said is approximately the difference between the reconstruction cost and the original cost of acquisition. As might be expected for an insurance industry firm, Ellam was unable to provide access to his firm's detailed actuarial data.

3.2. Metrics for Detecting Compromised Systems

Shivaraj Tenginakai (Sarithi)

Tenginakai wants to build security in using unsecured systems of unknown ownership operating on unsecured networks. The conventional model achieves high control at high cost — a fort with defense in depth. He substitutes a collective for a fort; a collective does not *defend* against attacks but rather *detects and drops* compromised members of itself. This is a restatement of the Byzantine fault tolerance problem. Put differently, as a distributed e-commerce system becomes more complex, it eventually becomes impractical to defend each component therefore the only workable strategy is one of isolating compromised nodes from the system, *i.e.*, not trying to fix them or to otherwise defend against compromise.

Collective members are either sound or compromised. When compromised, it is one of: structural compromise (failure to contact other legitimate members), temporal compromise (falsifying timelines), and data compromise (manipulated data).

As an illustration, Tenginakai uses a Lamport Clock as the metric for detecting compromised hosts on the argument that, at least in e-commerce, there is sufficient determinism in the order of processing that a violation of a Lamport Clock inequality will mark a node as compromised and thus the collective will then discard that node from further processing of the e-commerce work flow. Note that this means it is the security metric that determines run-time architecture of the collective e-commerce system. While a work-in-progress, Tenginakai believes he has shown completeness in simulation, and his slides provide a cost analysis that is favorable on several axes.

4. Enterprise Security Management

Security Metrics in Governance, Risk and Compliance	—	Liu
Using Security Metrics to Motivate a Response to A Critical Vulnerability	—	Cowie
Foundational Practices that Optimize Security and Operations	—	Kim

4.1. Security Metrics in Governance, Risk and Compliance

Li Liu (Security Team, eBay, and now student)

Liu began by noting that nothing gets done unless someone has to pay for it if it breaks. As do some others, her view of security is that it is a subset of reliability. She also believes that security metrics must be collected to a central database that is searchable, which is to say that they have to be recast when useful to do so. The justification for such a data base in organization terms is, however, the readily recognizable Governance, Risk & Compliance (GRC) with which so many firms have come to organize themselves around. All of this is impossible to do without substantial automation.

A questioner asked if the eBay problem set was similar to the Orbitz problem set and Liu confirmed that it was — for both, perhaps the most critical vulnerabilities are those which break multiple brands. Another questioner asked about how an entity such as eBay may not itself be attacked (and hence require defense) as much as it may be used by some eBay customers to attack other eBay customers. Liu said that they are very aware of this and that it is important. Another questioner noted the interesting evolution from organizational silos to semi-overlapping Venn diagrams with gaps and double-counting.

4.2. Using Security Metrics to Motivate a Response to A Critical Vulnerability

Jim Cowie (CTO, Renesys)

Cowie got right to the point: nothing motivates like a metric that can generate fear and/or shame. In his view, metrics are a refined method of social engineering because they boil down the complex, put things in some kind of order, force action, and calibrate change.

Cowie took as his working example the Border Gateway Protocol (BGP) which, on the one hand, is the only way Internet routing works and, on the other hand, is both broken and unfixable. As befits something that is inherently insecure, it is complex enough that few understand it in full. Given that fragility, all network “cooks” with their hand in the pot have to keep on their toes, and rely on everyone else to do so, too. In Cowie’s view, this is just the situation for metrics of the sort he mentioned at the outset, and they are: Compliance, Availability, and Diversity.

When an ISP chooses to do so, it can register its routes with a third party and Cowie’s company is (in part) in the business of comparing registered routing with observed routing, generating the Compliance metric. The Availability metric is as it sounds, and is measured by observation. The Diversity metric is also measured by observation, and low scores are for those exposed to single points of failure in their routing design. He closed with displays of interesting graphs that compared various countries and organizations along each of these three metrics, again reinforcing the notion that trusting BGP is both necessary and technically unwarranted hence setting up a situation where motivational metrics have a direct value.

A questioner asked whether the three metrics Cowie demonstrated are not, in fact, correlated. Cowie said that they were not as, for example, one can have high stability as a side effect of zero diversity, at least until the day that the inherent single point of failure fails. In response to another questioner, Cowie said that work such as this is, in fact, directly apropos to a “tragedy of the commons.” When asked about the finding of Barabasi that no network can be designed to resist both random faults and targeted faults, Cowie acknowledged the point, but suggested that the metrics here are more similar to re-insurance models where diversity of risk holders is a stabilizing influence, and that while network collapses are highly rare they are also highly damaging. He did, in response to a final question, note that ISPs have such thin margins that they effectively cannot care about these sorts of things, at least not naturally.

4.3. Foundational Practices that Optimize Security and Operations

Gene Kim (CTO, Tripwire)

Kim's talk began by asking what makes high performers and answering that rigor & discipline are predictors of security effectiveness, and this comes from a culture of change management, causality, and continual depression of operational variance. Numerous graphs and lists in his slides should be consulted rather than having them enumerated in digest form here. One punchline, however, is clear and arriving at it is a reportable result, namely that the quality of performance is largely explained by a small number of factors, and these factors can be measured. Specifically, 60% of performance can be traced to (whether a firm has)

Standardized configuration strategy
Process discipline
Controlled access to production systems

Kim also found that the ability to recover from faults and a lower need to do so were correlated and differed markedly between high and low performers.

Again, the nature of Kim's findings is best expressed as the graphs that he provided in his slides, a picture being there being worth a thousand words here.

5. Lunch discussion of handouts, including:

- Measuring the future basis of competition among AV products
- Performance Testing the Vulnerability Response Decision Assistance (VRDA) Framework
- PCI DSS Statistics and Metrics
- Techniques for Enterprise Network Security Metrics
- CIS Consensus Project
- SOX Material Weakness and CIO/CEO turnover

There is little way to summarize these discussions as they were one per table and free-form.

6. Software Security

The Building Security In Maturity Model	—	McGraw & Chess
Does Software Quality Matter?	—	Clark & Blaze

6.1. The Building Security In Maturity Model

Gary McGraw (CTO, Cigital) & Brian Chess (Chief Scientist, Fortify)

The work described by McGraw and Chess has a large body of documentation located elsewhere, enough so that they did not provide redistributable materials specifically for MetriCon. They began their description of that work by noting that it was observation-based, that is to say non-interventionary, resulting in value-free comparisons between individual firms and the aggregate of other firms in their study (*i.e.*, a jackknife process with ordinal scale outcomes). Their observations were many, such as the mean size of a security group within a software build team as 1%. While fascinating throughout its findings, it has not yet reached a large enough sample size to be presumably generalizable and, in fact, McGraw characterized a metrics program as “like an organ, and, furthermore, transferring a metrics program from one enterprise to another seems much like organ transplant — chances of rejection by the host are high.”

6.2. Does Software Quality Matter?

Sandy Clark (Ph.D. candidate) & Matt Blaze (Prof., University of Pennsylvania)

Clark reminded all of the curve from Brook’s classic The Mythical Man Month, which shows that the number of flaws in a program declines over time until a point is reached where, subsequently, it begins to rise inexorably, what Brooks said, in words, as

> Program maintenance is an entropy-increasing process, and even its most skillful execution only delays the subsidence of the system into unfixable obsolescence.

and it is this idea of counting vulnerabilities in a body of code as a predictive correlate of the number of successful attacks against that body of code that Clark means to show as misguided.

In particular, Clark says that all software enjoys a honeymoon: The time it takes an attacker to learn a piece of software, find its bugs, write, debug & test an exploit, and release it into the wild. She finds this natural and unstudied; that it is natural because software attacks are a steady state arms race, and that it is unstudied as a simple proof by observation. If needing a slogan, remember “Patch Tuesday” begets “Exploit Wednesday”.

Clark believes it obvious that the intrinsic security properties of software are a poor predictor of when an attack will occur and/or how devastating an attack will be; *ergo*, a focus on intrinsic security properties leaves us defenseless against new, innovative attacks and, as such, spending attention and resources on intrinsic security properties is unprofitable. Rather, the focus needs to be on extending the Honeymoon. Because of the existence of the honeymoon period, new code is better than old code even when it introduces new vulnerabilities.

7. Trends and Stats

Crunching Metrics from Public Data	—	Nichols
Data Loss DB	—	Shettler

7.1. Crunching Metrics from Public Data

Betsy Nichols (CTO, Plexlogic)

What in some contexts would be called “open source intelligence” is, courtesy of Google and the DataLossDB, easy to do for questions such as the ones Nichols posed:

Is Breach Frequency Increasing?
How do Breach Frequency and Total Affected compare for Inside *vs.* Outside sources?
How does a breach affect stock price?

where easy to do means both that data is available and that it is available in large enough sample sizes to reach statistical conclusions. Because Nichols’ results are quantitative, visual, and self-explanatory, the reader is referred to her slides for further flavor. In the composite, they show that there is little-to-no lasting effect of a data breach with the small caveat that measurements of this sort are somewhat sensitive to their sampling interval.

7.2. Data Loss DB

David Shettler (Open Security Foundation)

Shettler described the methods of collection, the resulting data quality, and the challenges of the Open Security Foundation's DataLoss Data Base, formerly hosted by attrition.org. No distribution slides were made available and the external documentation is, of course, well available.

The DataLoss DB is a child of Chris Walsh and Attrition, as is well known, but these days the primary method of data acquisition is the public records request made under the US FOIA (Freedom of Information Act). Needless to say, completeness in the DLDB is difficult and difficult to assess, and in pursuit of completeness the collectors are now gunning for court records, too. Getting this data, as he detailed, is labor intensive and the labor intensity is the rate limiting factor in both quality and completeness of the DLDB. Because there is an ongoing and evidently endless process of retrospectively backfilling gaps in the record, there are no findings, *per se*.

8. Security Manager Panel

Moderator: Jennifer Bayuk (Independent Consultant)
Panelists: Ed Bellis (Orbitz), Chris Walsh (SurePayroll), and Robert Masse (Reitmans)

The panelists began by, in various ways, acknowledging that complexity and change-rates are the principle opponents of a security manager. Questions came quickly.

Nye asked whether they thought that a point scale for control quality is possible, *i.e.*, for their reaction to the thesis of his earlier presentation. Masse felt that Red/Yellow/Green was the best that he could get across to others in the firm. Walsh thought it possible, but was sceptical of the the comparability of the scores. Bellis also thought it possible, but noted the problem of snapshots in time can be misleading in short order.

Ellam asked whether good enough is good enough, or whether aiming for 100% is more rational. Walsh thought that consistent output of a security program is good enough, *i.e.*, that consistency does let one look at trends. Bellis felt, instead, that the important thing is not so much whether good enough is good enough but rather how far from 100% are you at any time. Masse agreed with Ellam's premise.

Kim wanted to know what the panelists' key controls are. Bellis looks for "sanity checkpoints," while Walsh looks to effective change management and availability, tempered by a trust-but-verify scepticism. Masse offered that he is learning to enjoy audits, and hopes to teach others to do so.

Chess asked whether a separate software security group (SSG) is or is not an indicator of the firm's ability to have secure software. Bellis said that it is not essential and, in fact, that the responsibility for software security simply rests wherever the builds get done. Walsh agreed with both Bellis and Masse, that bringing the knowledge to where the work is done is the only practical approach.

Nichols asked what it was that influenced management to support security metrics now that it has been shown that security breaches do not affect the firm's stock price. Bellis offered that compliance is the driver and the only driver. Masse likewise agreed that compliance is everything, but only to the extent that compliance is visible on the bottom line. Walsh, however, felt that compliance for his firm is secondary to the compliance needs of his customers, but also cited the reputational harm that can come from publicly known data breaches. Masse countered that, for him, reputation is not all that big a deal as "we are not the CIA". Bellis agreed that reputation risk is no big deal, but because customers have become jaded to the seemingly daily reports of data breaches.

Arrot asked what metrics are best for defense of a security budget. Masse thought that *n*(*incidents*) first and *value*(*assets*) second, but said that *exists* some definitional matters between availability and security that have to be worked out. Bellis thought that it was the *n*(*incidents*); Walsh agreed with *n*(*incidents*) but added a declining *cost*(*audit*) is also reinforcing.

Jaquith asked what it is that drives behavior change. Bellis reiterated that it is shame, and told an interesting story of a trophy that is awarded and which, in fact, no recipient would want on his or her desk for long if at all. Masse takes a different path, forcing a statement of intent along the lines of "If you don't want to implement this control, then sign off on it here," which tactic he does save for important occasions.

Another questioner asked whether it is possible to make the business owner accept residual risk, by forcing a signature on a list of controls, risks, and recommendations. Bayuk offered that when this is necessary is in a compliance situation. Arrot, asked for clarification on whether forcing acceptance is useful. Masse said that he does not put a gun-to-the-head as he prefers to make awareness his metric rather than percentage of risks which are signed-off. Another questioner asked whether reporting the residual (and accepted) risks upward is a good idea. Masse simply asserted that this does not happen. Bayuk said that it does but only for policy exceptions and then only if firing someone is a potentially rational outcome. Masse countered that that would be like whether or not to ignore the recommendation of one's General Counsel. Bayuk replied that, for instance, the CFO of AIG is going to jail so if there is a law and you don't push it then you go to jail amounts to a teachable moment, as it were.

Closing

Various closing remarks were made, mostly around whether you want to work on this or that to come see such and such a person. Dinner followed with discussions leading well into the evening hours, but not captured here.