MetriCon 1.0 Digest

1. Background

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 of this year. Specifically decisions included that this was to be a workshop rather than a formally refereed academic event and that it was to have a cap on attendance. Andrew Jaquith (Yankee Group) became chair of the organizing committee. Betsy Nichols (ClearPoint Metrics), Gunnar Peterson (Artec Group), Adam Shostack (Microsoft), Pete Lindstrom (Spire Security), and Dan Geer (Geer Risk Services) also served. Dan Geer arranged that the MetriCon 1.0 event be co-located with the USENIX Security Symposium and is the principal author of these notes.

Forty-four people attended, predominantly representing industry (30) rather than academia (10) or government (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.

2. Opening "debate"

Well, it wasn't a debate, but there were two positions from diagonal if not opposite corners: "Metrics are nifty" *v.* "Metrics are infeasible."

2.1. Metrics are nifty

Andrew Jaquith opened MetriCon 1.0 with a discussion of why bother with metrics and what metrics he favors. Other fields have their bodies of managerial technique and control, but digital

security does not. This has to change. Security metrics are for the purpose of making judgment calls, and a good metric should be (1) consistently measured, (2) be cheap to gather, (3) contain units of measure, (4) be expressed as a number, and (5) be contextually specific. A good metric should pass the smell test ("This metric helps me how?")

Jaquith argued that this all breaks down to modelers and/versus measurers. Modelers think about how and why; measurers think about what. He offered a comparative table

modelers	measurers
Risk equations	Empirical data
Loss expectancy	Correlation
Economic incentives	Data sharing
Why	Causality

though he was quick to admit that measurement without models will not ultimately be enough, more like let's get started measuring something for Heaven's sake.

2.2. Metrics are infeasible

Steve Bellovin spoke in counterpart on the brittleness of software and the infeasibility of security metrics. He, like all of us, begins with Lord Kelvin's dictum on how without measurement your knowledge is of a meagre and unsatisfactory kind. However, Dr. Bellovin concludes that the reason we have not had much progress in measuring is that it is in fact infeasible to measure anything in the world as we now have it. He would want to answer "How strong is it?" in the same style that fire resistance is referenced in the building code, but maintains that we cannot do this unless we change how we do software.

The reason is that defense in the digital world requires perfection, and he talked through several well-known illustrations of the point (Witty, Kerberized telnet, and SSH buffer overflows). His main point was that the attacker's effort is linear in the number of defensive layers, not supra-linear (much

less exponential). Put differently, layering as done today means that strength is linear in the number of layers, though layers can even interact malignantly, *e.g.*, Java applet versus FTP proxy.

Brittleness will persist until we can write self-healing code or we have composition rules that have greater than linear increase in layering strength. So a Challenge: Show me the metrics that help this.

2.3. Discussion

Lindstrom disagreed with the conclusion and instead argued that Bellovin's reasoning did not show that metrics are impossible but rather that they are necessary (as in "I will die, but when?"). Another attendee asked "But then what about diamonds? Diamonds are non-composable, diamonds can be fractured along well known fault lines, *etc.*"

Yet another attended 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.

An attendee suggested that we can borrow some ideas from the physical world, especially relative measurements like "this is safer than that." Bellovin replied that in the physical world there is little equivalent to the "one undetected bug" problem, and that is why our job is so hard.

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: "I want to know about changes in data state." Bellovin, in reply, asked whether Internet Explorer 7.0 will be better or worse than today's version?

Opacki objected by saying "Come on; simplifying assumptions are built into everything. So what?" Jaquith, satisfied that everyone is fully engaged, then cut off conversation in the name of schedule.

3. Software Security Metrics — Gunnar Peterson, track chair

A Metric for Evaluating Static Analysis Tools	-	Chess & Tsipenyuk
An Attack Surface Metric	-	Manadhata & Wing
"Good enough" Metrics	-	Epstein
Software Security Patterns and Risk	-	Heyman & Huygens
Code Metrics	-	Chandra

Peterson began with a call for rethinking what granularity we need if metrics are to be meaningful, such as to be specific about goals like integrity, confidentiality, or availability, but not security which is not a goal for which metrics are meaningful. He reminded the attendees of the first rule of engineering: fast, cheap, reliable — choose two, and argued that at present reliability is chosen too often while fast and cheap are underutilized. He felt that the same is true with metrics, that it is not helpful to focus on any one area even if somehow we think that we know more about that area; rather, it is time to just throw lots of measurements at the wall and see what sticks. In Peterson's view, "system" and "security" are not helpful — that we need to be more granular: CIA, authentication, *etc.*, noting that we have pretty good availability metrics already.

3.1. A Metric for Evaluating Static Analysis Tools — Katrina Tsipenyuk & Brian Chess, Fortify Software

Chess began by noting that metrics for evaluating security analysis tools depend on the audience: the tool vendor wants to track the improvement of the tool itself, an auditor wants to avoid the risk of false negatives, and a developer wants to avoid the noise of false positives. Chess broke

out the developer's versus the auditor's point of view, reminding attendees that the value of a set of metrics is based on who is consuming them.

Chess thus proposes a weighted composite score based on those three interests, *i.e.*, where the true positive rate (TP), the false positive rate (FP), and the false negative rate (FN) are combined but are weighted depending on the audience for the composite score. Chess' Proposal is to compute 100*TP/(TP+FP+FN) times a series of weights best examined in the handout itself.

Chess then showed some preliminary results of applying this effort to three applications, four versions of Fortify's tool, and for two consumers of the resulting metric. Chess stressed that the purpose of their effort at Fortify was to guarantee improvement in their product itself, and that that may limit applicability. As with any tentative scoring mechanism, more data is needed to verify and, then, calibrate the model. [*Ed*: Receiver Operating Characteristic (ROC) analysis is often applicable here; attendee Cardenas is working on this and has papers to read.]

3.2. An Attack Surface Metric — Pratyusa Manadhata & Jeannette Wing, Carnegie-Mellon

Manadhata wants to find an answer to an ordinal question: Is A>B? (Read "Is A>B?" as "Is the attack surface of A greater than that of B?") He posits a formal framework composed of a set, M, of Methods including entry and exit points, a set, C, of Channels, and a set, I, of untrusted data Items. For each of these resources, he estimates the ratio of damage potential to effort (called "der" and where a high ratio is naturally a sign of notable danger). His measure, then, is based on system design and not on behavior of the attacker, and the attack surface is the triple

$$\left(\sum_{m \in M} der_m , \sum_{c \in C} der_c , \sum_{i \in I} der_i \right).$$

He then displayed several examples. For each, he manually annotated the source code and analyzed the call graph of the application using off-the-shelf tools to identify the set M and used run-time monitoring for C and I (at a labor cost of 5-6 days for 40K lines of code). Some validation was done, but in the spirit of a workshop this is very much a work in progress. The question on the table is whether the number of vulns is or is not correlated with the attack surface metric, including whether it can be found to be correlated with honeypot data in the field. Note that this particular metric decouples channels, methods, and data when looking at the attack surface, and as such is much more granular than "system security."

3.3. "Good enough" Metrics — Jeremy Epstein, WebMethods

Epstein has a basic point: Rather than argue about which numbers it makes sense to collect, gather as many as you can and 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 called attention to Cowan's "relative vulnerability" as an example of good work. In that, Cowan suggests taking the ratio of exploitable vulnerabilities with and without an intrusion prevention system (IPS) present. Epstein suggested that ratios of this sort and the orderings that they induce on various applications are valuable, but what he really wants are analogs to the well-known "leading economic indicators" which are naturally a set of leading security indicators.

Epstein's interest was thus on prediction rather than on keeping score retrospectively. In fact, he considers retrospective scores to be, at best, useful only to early adopters and then only for assessing the reputation of the vendor when a product is new or the transaction is a merge or acquisition. The metric then might be some $f(insecurity * popularity * ubiquity)$ which would grow the more juicy, the more hated, and the more widely accessible you are.

Note that "Good enough" does not mean "just any" and in that vein Epstein describes why he does not much believe in the CSI/FBI annual survey. To get somewhere, we need more data and Epstein proposes using the `securitymetrics.org` website to post release of data such as:

> Number of (unfiltered) static or dynamic analysis hits
> Number of Bugtraq or CVE entries / time
> Average education/experience per developer
> # of LOC/developer/time
> % of code that's reused from other products/projects
> % of code that's third party (e.g., libraries)
> Leading security indicators adherence

3.4. Software Security Patterns and Risk — Thomas Heyman & Christophe Huygens, Univ. 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 primarily interested in ratio scores like the number firewall invocations *vs.* the number of service invocations, or the number of guards *vs.* the number access points for each component. Huygens suggested that the patterns they proposed may actually represent the middle ground (and perhaps even reality) between the modelers and the measurers as patterns arrange things in sequence like models, but are specific enough for precise measurements.

Preliminary results indicate that this is indeed possible, and the results are consistent with Jaquith's five rules (*vide supra*). Ultimately, the aim is to aggregate multiple metrics into indicators and to bring these indicators into the design space (and not have security as a bolt-on at the end) though that will require changes to how design is done. It also appears that the computing fabric itself has to be more resistant, such as micro-IDS scattered throughout.

Numbers gathered in this approach may be applied to both structural and behavioral elements in a system. As almost all "security" numbers one sees are structural, this may be a critical advantage if, in fact, the security of software in the field is more behavioral.

3.5. Code Metrics — Pravir Chandra, Secure Software

Chandra's focus is on remediation metrics, specifically those metrics that help (and assess) getting better and better. As befitting a workshop, this is quite preliminary work, but his main tool is a 4x4 matrix as follows:

	Review State			
Severity	Unknown	Known	Accepted	Mitigated
Crit				
Error				
Warn				
Info				

which he proposes to display in reports as a series of sparklines (ref: Edward Tufte). For each Review state, he will use capture-recapture or capture-for-removal metrics for to estimate flaw count and then look at changes in market share by severity to track progress. [*Ed*: "Measuring Security" tutorial has explanatory notes on these techniques.]

Chandra also proposes to look at software complexity, specifically McCabe, System Complexity (SYSC), and Information Flow Complexity (IRC) metrics for which he gives references. To that, Chess pointed out the interesting example (due to Bill Pew, U Maryland) of a function that calls itself and questioned whether this is a problem or not. Specifically, Chess maintains that complexity metrics don't capture what is going on in such a case, and in general he feels that complexity has little relationship to insecurity..

4. Enterprise & Case Studies A — Adam Shostack, track chair

Data Breaches: Measurement Efforts and Issues	-	Walsh
The Human Side of Security Metrics	-	Opacki
No Substitute for Ongoing Data, &c	-	Quarterman & Phillips
What are the Business Security Metrics?	-	Butler

4.1. Data Breaches: Measurement Efforts and Issues — Chris Walsh

Walsh began with a time-line slide on the dates of adoption of (California) SB1386-like laws post ChoicePoint, and asked a question: Are on-line breaches a significant source of ID-theft? He, and most studies, focused on firm-level impact on income or stock price. As Walsh pointed out, such studies are too few in number and too few in coverage, and they lack enough reach to establish causality, say, 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, a research agenda in other words.

4.2. The Human Side of Security Metrics — Dennis Opacki, Covestic

Opacki began with a straightforward question: "How can you manage if you don't have gauges to read?" In his view, the point of any metric is to change behaviour, but behaviour change can have pitfalls such as means becoming ends in and of themselves, unintended consequences, and, of course, passive or active resistance. Opacki presented recent findings in evolutionary psychology, such as that we humans are hard-wired not for logic but for detecting injustice, and he followed with a number of observations in social psychology and behavioural economics. One of these is that intuition is low energy cost and runs parallel while reason is high energy and single-threaded. Opacki's advice is to focus on scales that people gauge intuitively, and keep the number of metrics

small, to not neglect entertainment value, and to give bad news first. Measure the impact of your delivery before and after, express everything in dollars where you can, and use plain language.

Overall, his point is basically that "data = power" and recognizing that the political level of discourse does have a lot of steering power, Opacki noted that you cannot deliver just bad news, that you should lead with bad news so that good news is last and remembered, and you want to rely on your audience's mental prototypes/analogy. In short, it is better to be vaguely right than precisely wrong.

4.3. No Substitute for Ongoing Data, Quantification, Visualization, and Story-Telling — John S. Quarterman & Gretchen K. Phillips, InternetPerils

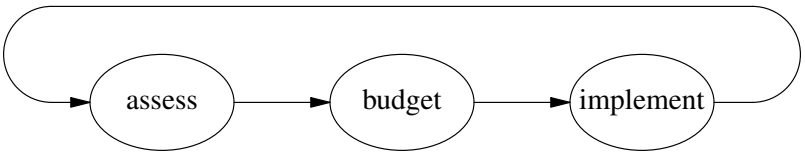
Quarterman demonstrated how InternetPerils handles phishing attacks, making an argument for data aggregation. He used animation to illustrate time-series of complex interconnection paths. Quarterman feels that only so much slicing-and-dicing can be done for public consumption; that only after getting the receiver's full attention can you add complexity of the sort that slicing and dicing requires. He is squarely in the observation (measure) camp, not the simulation (model) camp, and suggests collecting data when you do not yet need it, *viz.*, getting and retaining solid baseline data. Quarterman also suggested that metrics help you understand — and react to — what goes on beyond your span of control such as on the other side of your firewall.

Opacki countered with an InterNAP experience where 100M probes went out and 1K complaints came back — a result that is statistically zero but it overloaded the network operations center anyway. Quarterman agreed, noting that well designed probing does not require the risk of overload. Several commentors discussed what a reasonable standard of performance for an ISP to take a known phishing site off the air really is, noting that it's a hard problem as masquerading attacks

from one ISP to another are feasible.

4.4. What are the Business Security Metrics? — Shawn Butler, MSB Associates

Butler says that business decisions are what security metrics are about. Given that this is so, what we must have is frequency and impact if we are to get at true cost, and cost is the basis of the 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 notes that while requests are not coming down from on high, massive amounts of data are moving up and, worse, that 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 that there is no feedback between effectiveness and investment. Put differently, we break this feedback cycle:



Butler posits a sample goal: Reducing confidentiality risks to sensitive data. To do that, you need questions answered, questions like: What are my greatest risks to this type of data? Who has access to it? What is the likely impact? What metrics answer these questions?

5. Enterprise & Case Studies B — Betsy Nichols, track chair

Leading Indicators in Information Security - Nye

Top Network Vulnerabilities Over Time	-	Solem
IAM Metrics Case Study	-	Sudbury
Assessment of IT Security	-	Hallberg & Hunstad

Nichols began the session by displaying some interesting survey data showing that maturity of security metrics deployment and market capitalization are uncorrelated. As session lead, she set out the core question, *viz.*, Why are metrics so hard? Her proposal is this: Vast and unclean data (scattered, uncorrelated, incomplete, and inconsistent), a lack of consensus on indicators and models, and difficulty in packaging results.

5.1. Leading Indicators in Information Security — John Nye, Symantec

Nye wants us to have leading security indicators. As he has easy access to work at the Symantec Attack Center (fed from many sources), he undertook to show what those indicators might look like. Beginning with the results of 449 remote penetration tests (scans) of Symantec clients, he calculated a "vulnerability score" (the average severity of vulnerabilities per host within each target firm) and, in parallel, a "vulnerability saturation" (the average number of vulnerabilities per host within each target firm), though the remainder of the presentation looked only at saturation.

Dividing the data set into quartiles by saturation, and sadly discarding the high-order quartile due to last minute questions about data quality, Nye nevertheless showed an expectably sharp rise in vulnerability saturation from quartile 1 to 2 and on to 3. From that, he was able to find specific vulnerabilities that might serve as leading indicators. Work is ongoing.

5.2. Top Network Vulnerabilities Over Time — Vik Solem

Solem used a similar a data set limited to only Nessus scans (over a contiguous eight-month period). In particular, Solem showed stacked line graphs of the top 10 vulnerabilities (by frequency) over the eight-month study interval, of the 23 vulnerabilities that occurred in all eight months however little, and the sum of the two. As with Nye, work is ongoing.

Questions from the attendees included were mainly about data sources and methods, and Nye/Solem took away a sense of how to improve their work. While some attendees wanted to see the data broken out by industrial sector, this was not possible due to confidentiality agreements. Solem did look at the Symantec Threat Report and happily found that there's no correlation between attacks and Nessus plug-in IDs. This makes some analysis difficult, but there is correlation between attacks and what is in the Qualys "Laws of Vulns" report. Opacki reminded all that any tool, e.g., Nessus, could be scanning for the wrong things.

5.3. IAM Metrics Case Study — Andrew Sudbury, ClearPoint Metrics

Sudbury began by saying that real work, and this recounts real work but fabricated replacement data, is hard; you must start with real goals and, within that, what is it that you do not know. His measures are to see if you are in control of your controls, and he has found that business value combines explicitly from combining data from multiple sources. The rest of his talk was to show this kind of consolidated security metrics as his firm does them.

Kirkwood suggested that Sudbury add targets to his graphs, not just trend data. Jansen asked how does one know, for example, that a help desk clearance score is actually clearance and not just somebody skipping work (thus making clearance look better than it is)? Sudbury's answer was to be careful what you are measuring, of course.

Blakley asked which of (1) management is dumber than technical staff, (2) management and tech staff want to see different things, or (3) you cannot give management bad news. Sudbury said "all of the above," but due to (1). Daguio said such tools are a good way to decide whether to ever give a particular team a project again.

5.4. Assessment of IT Security in Networked Information Systems — Jonas Hallberg & Amund Hunstad, Swedish Defence Research Agency

Beginning with a call for metrics as necessary and useful, Hallberg made the insightful observation that while system properties control the security level and the security level controls consequences, the security level is not measurable whereas the other two (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, working top down, came up with five high-level security properties: access control, security logging, protection against intrusions, intrusion detection, and protection against malware. In total, the five high-level security properties include 77 system properties.

Security property	n(Low-level properties)
access control	20
security logging	12
protection against intrusions	17
intrusion detection	12
protection against malware	16

Saaty's "Analytic Hierarchy Process" is a decision making technique relying on empirical weights between and amongst criteria within a class, *i.e.*, weights within a class must sum to 1 so, here, the 20 low-level properties related to access control are weighted to reflect their relative importance to each

other and to sum to 1. Hallberg also described a system modeling approach call MASS.

The relationship between the properties was likely the most interesting and useful part of this. Endpoint metrics are valuable, but how do they relate to properties? Hallberg's discussion was perhaps the closest thing to what Bellovin was after that we saw all day.

6. Governance — Dan Geer, track chair

This session run more as a discussion.

Metrics that matter are for decision support	-	Geer
DHS funding allocation via risk metrics	-	Ware
American Bankers Association and self policing	-	Daguio
Framework for risk from rare events	-	Blakley
What is needed for metrics	-	Jansen

6.1. The only metrics that matter are for decision support — Dan Geer, Verdasys

Geer's title says/said it all, though there was some argument but mostly over what is "future" and did it mean that you could be amnesiac, plus whether modifying behavior through education versus decision-centered metrics was better. Argument was deferred to looking at some real examples

6.2. Model Concepts for Consideration and Discussion — Bryan Ware, Digital Sandbox

Ware has been figuring out for the (US) Department of Homeland Security how to allocate grant dollars. The past (V1.0) method: per-capita $$ for terrorism which sounds fair but is nutcake. Now (V2.0): conversion to risk-centric $$ allocation, all of which is easier said than done. In particular, there are Givens: For example, what is the Risk Level is dictated, not derived. The problem, however

is that the risk level, as calculated, has exponential decay but that would lead to all $$ going to the very few not the many, and this is not acceptable, largely though not just politically. Problems with the marginal rate of return (being low) also arise. So the first step was to require management plans from states and cities that respond to the risk measures/estimates.

Ware is a fan of 2x2 tables because they show the decisions you are making, especially given the clientele. For this exercise, Ware's firm used 17 sets of 6 experts each in an HP (hierarchial) mechanism, and so forth. The central problem though is setting the thresholds that divide the rows and columns into two.

RISK			
	High	Discount for low effectiveness	Best investments
	Low	Apply minimum funding	Incentivize high effectiveness
		Low	High

EFFECTIVENESS

On the Y (RISK) axis, it's exponential so Ware used an inflection point in the curve (exponentials have no obvious place to cut). On X (EFFECTIVENESS), the median was a good enough cut point. Quadrant allocation of $$ comes first, then within quadrant. In the low/low quadrant, a minimum amount of money is used. In the high risk but low effectiveness quadrant, New York City (give us all the money) and Washington (incompetent management team) were the most problematic. Choosing between low risk / high effectiveness and high risk / low effectiveness was hardest. Most $$ went to high effectiveness, which got Ware's firm raked over the coals but institutional behavior change has resulted.

Discussion was brisk. Kirkwood asked if any decisions were made outside the matrix. Ware replied that though some deserved none, every entity did get the low/low minimum. To accomplish this they ended up making NYC and DC special cases by removing them from the matrix altogether and setting the $$ allocation by law, not by formula. One could argue that this was an outlier removal problem.

Geer asked about risk aversion? Ware more or less said that expectations (of the recipients) do matter.

6.3. Mission and Metrics from Different Views: Firm/Agency, Industry, and Profession — Kawika Daguio, Northeastern University

Daguio is not used to working in the open. He reminded us all to "Do no harm" as we introduce new metrics. As well, he reminded us that accountability matters, and that separating risk and compliance is essential. Compliance, which comes down to individuals and not corporations, is more important that 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. Remember that metrics programs in banks are a lot about pruning internal parts of the organization (getting someone else's money and power), which does not mean to lie/cheat/steal. The financial services sector is happy with (California) SB1386 as it spreads the banks' pain to other industries. Put differently, it says to get rid of pain or share it.

Daguio says that we should use nominal and ordinal measures to avoid bad effects; that we should not do interval or ratio scales because those invite comparison and hence interference and, then, further problems. To get information sharing will require competitive, policy, technical, and political reasons for doing so, or it simply won't fly. He noted that too much red is bad, but a special

color of red for unknowns is ok.

Daguio's slides, which should be read, remind us all that while we (and this Workshop) are about metrics that 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. Ware pointed out that even an ordinal scale for how to spend dollars brings criticism ("You spent the #1 $$ on >1 risk"). Kirkwood asked whether the Financial Services Information Sharing and Analysis Center (FS/ISAC) a forum on, at least, leading indicators? To that question, Daguio suggested that FS/ISAC's main function is preventing worse with respect to government interference. Geer responded with a diatribe on information sharing and how no General Counsel will ever let it happen as their jot title is really "formally risk averse" to a fault. Ware, referring to the "dread continuum," suggested that FS executives tended to not know about FS/ISAC so FS/ISAC, having kept the regulators at bay, is of enormous value. Daguio replied that no, the most important function of FS/ISAC was to establish trust between people of similar job descriptions and that 9/11 was proof that having your counterpart's cell number in your cell phone was more valuable than anything else.. Geer, Kirkwood, Daguio and others then argues about what industries had similar common interests in safety or security (*e.g.*, finance, oil, airplanes, *etc.*) and whether there was anything to borrow between them.

6.4. Measuring Information Security Risk — Bob Blakley, Burton Group

Blakley began with a formal definition (*ex* Wikipedia) intended to disambiguate a measurement from a metric, and to look at metrics with an eye to finding "normal limits" 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 what we are not doing is 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 the decisions to be made:

high impact	Mitigate	Mitigate & Recover	Recover
low impact	Mitigate	Ignore	Ignore
	common	uncommon	rare

and in his slides further states what needs to be measured in support of these decisions. Blakley also points out that for decision making, correlates of risk are just as good as direct measures of risk, using as his example that it 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 suggests we should find and be happy with such correlates in our sphere.

Opacki asked for an example of "high impact" & "common" to which Daguio volunteered trade matching failures in financial exchanges (seller said he sold 15,000 and buyer said he bought 50,000). Ozment asked why one would estimate probability if you are already calling it "common" and Ware volunteered work by Bob Jacobson of IST in New York City, notably his CORA (cost of risk analysis) product.

There was then some discussion of frequentist versus Bayesian approaches, and whether a bimodal probability distribution ($1x10^6 *vs.* $10^6x1) doesn't make any probabilist approach impossible. Quarterman asked about risk aggregation, to which Blakley answered "I don't want to touch this." 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. Blakley finished by noting (with example) that what medicine calls differential diagnosis may well be the goal of a security metrics program.

6.5. Information Assurance Metrics Taxonomy — Wayne Jansen, NIST

Jansen showed one slide summarizing the work of Vaughn, *et al.*, entitled "Information Assurance Measures and Metrics" from 2002 (see slide notes). 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? No such baseline, or even documented case studies with lessons learned, seems to be available. Could such a baseline be established? If so, should those metrics and measures be organized according to the capability and maturity of an organization? The apparent discontinuity between strategic efforts and tactical ones leads one to ask whether there is a way to bridge the gap? Where are best practices for doing this documented? Getting any metrics established within the organization process seems hard — how can it be done in a systematic way? 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 (forward looking) credit decisions even though it was oddly constructed, and the KMV-Merton model to predict likelihood of default for corporations (look at Bloomberg data for what happened when it was introduced even though it was Heisenberg-like as it changed what it was measuring because it was measuring it).

Daguio, as a banker at that time, asked that we all please don't 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. He sees that that in the Basel Capital Accords. He sees that money that would be spent in Gramm-Leach-Bliley or Sarbanes-Oxley re-directed to other things where nothing gets fixed just reported better. Geer and Daguio spoke more of Basel, with Daguio pointing out that it is a backwards-looking measure. While Daguio says that Basel has had an impact, he has spent eight digits of money and what did he get for that??

An attendee asked about game theory, such as 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. Secondly, infosec is an economic game and not just a technical game. Our mental model of the economics is influenced by the underlying mathematical model, so we need to pick that well. As it is adversarial, so our job is to shift the adversary's investment into different spots.

Daguio suggested that Jansen and Geer were dead-on. He recounted how, seven times, he had been brought in to "bring accountability" which always meant to get them killed. Four of those seven times, he saved the individuals by proving what was impossible. Daguio asks whether we can cross off the things that don't work, thus saving people and money all around? A good target, and practicable, is "professional wild-assed guesses" and that is a good enough goal for us.

7. Rump session and extra material

The Security Incident Database - Leversage

Does Software Security Improve with Age? - Ozment
Security Metrics - Lindstrom

7.1. The Industrial Security Incident Database - Eric Byres, Wurldtech Analytics & David Leversage, British Columbia Institute of Technology

Leversage observed that target of choice losses vastly exceed target of chance losses, that good old wire tapping 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.

Jaquith asked how it was that there seem to be no lawyers involved? Are there no agreements? Leversage said two interesting things: first that trade secret rules do apply, and, two, that his HUMINT is based on idealists.

7.2. Milk or Wine: Does Software Security Improve with Age? - Andy Ozment and Stuart Schechter, MIT

Ozment described a fine-detail time-series looking at the history of Open BSD. As this is a full conference paper with overlapping relevance to MetriCon, there are no materials here. As an inspired use of security metrics, this is a quotation from the paper's summary:

> Over a period of 7.5 years and fifteen releases, 62%of the 140 vulnerabilities reported in OpenBSD were foundational: present in the code at the beginning of the study. It took more than two and a half years for the first half of these foundational vulnerabilities to be reported.
>
> We found that 61% of the source code in the final version studied is foundational: it remains unaltered from the initial version released 7.5 years earlier. The rate of reporting of foundational vulnerabilities in OpenBSD is thus likely to continue to greatly influence the overall rate of vulnerability reporting.
>
> We also 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.

7.3. Security Metrics - Pete Lindstrom, Spire Security

Lindstrom, in rapid fire fashion, made a number of points about risk and doing a number of set-theory (Venn diagram) examples of how to calculate varieties of risk. In his 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 prevalance of activities in the past. That said, his examples are worth examining closely.

Summary — This event was an idea whose time had come, so much so that perhaps a longer time together would have been better. It is likely there will be another. All errors in these notes are the responsibility of the note taker, Dan Geer.

Thanks to the USENIX Association for the venue, and to those brave enough to come and those braver still to organize. Attendees were Steve Bellovin (Columbia U), Mike Belton (Berbee), Stu

Berman (Steelcase), Bob Blakely (Burton Group), Adam Bryant (Air Force IT), Shawn Butler (CMU and MSB Assoc), Alvaro Cardenas (U of Maryland), Pravir Chandra (Secure Software), Brian Chess (Fortify), Crispin Cowan (Novell), Kawika Daguio (Northeastern University), Jeremy Epstein (Web Methods), Carrie Gates (CA Labs), Dan Geer (Verdasys), Michael Grimaila (Air Force IT), Jonas Hallberg (Swedish Defence Research Agency), Thomas Heymann (U of Leuven), Sean Houlahan (Liberty Mutual), Chandler Howell (Motorola), Christophe Huygens (U of Leuven), Wayne Jansen (NIST), Andrew Jaquith (Yankee Group), John Kirkwood (AMEX), David Leversage (Wurldtech Analytics), Pete Lindstrom (Spire Security), Pratyusa Manadhata (CMU), Alain Mayer (RedSeal), Mike Murray (nCircle), Elizabeth Nichols (ClearPoint Metrics), John Nye (Symantec), Yeketerina O'Neill (Fortify) Dennis Opacki (Covestic), Andy Ozment (University of Cambridge), Gunnar Peterson (Artec Group), John Quarterman (InternetPerils), Stuart Schecter (MIT Lincoln Laboratory), Adam Shostack (Microsoft), Vik Solem (Symantec), Andrew Sudbury (ClearPoint Metrics), Tine Verhanneman (U of Leuven), Chris Walker (Microsoft), Chris Walsh (self), Bryan Ware (Digital Sandbox), and Jeannette Wing (CMU).