Hard Data Is Good to Find

In much of our "official" life, change is recorded as an index where by "index" we mean the representation of a time series relative to some known starting point. As relative measures, indices contribute to decision support without having to achieve arbitrary precision. Some are all but iconic — the Lehman Brothers' Bond Index, for one. Security (in our sense) could well learn something from securities (in their sense) as to how to do data fusion and meta-analysis — how to communicate via the vehicle of an index. Although brevity might be the soul of wit, an index shows that wits are the soul of brevity.

Daniel E. Geer Jr.
In-Q-Tel

Daniel G. Conway
Augustana College

Quoting directly from Lehman's (now Barclay's) description of how such a construction comes to matter in the large:

- Unbiased, rules-based methodology used to determine index constituents
- Comprehensive databases and accurate pricing sources
- Timely and reliable data-delivery platform

Would it not be more-than-nice if we had something equally cumulative for our world? Of course it would — which is probably why you're even reading this column. That is not to say it won't be hard since in our world the rate of change is always working against us — and trimming off the time-range of comparability for nearly any measure.

In fall 2003, the Computing Research Association sponsored a charrette for the US Congress on four Grand Challenges in information security to be met by 2013:

- An end to epidemics
- Certification that is trustworthy
- Minimization of the skill required to be safe
- Quantitative risk management on par with financial risk management

We're here to look at the creation of indices as a bedrock of meeting the fourth Grand Challenge, both for us and for officials.

Attentive readers of the January/February issue will recall our second annual "Øwned Price Index," or ØPI. Beginning with the next installment of For Good Measure, we'll publish the ØPI with every column. We hope to introduce other indices as the months progress, but because we insist on both rigor and practicality, we begin today with four relatively long-running sources of data with which to compose an index. God bless each of them for the work they've done in creating, publishing, and being consistent. Our work here today is the sincerest form of flattery.

The Anti-Phishing Working Group collects phishing data; taking April 2005 as the base (100) point, we plot together their four measures, the numbers of phishing reports, phishing sites, malware variants, and malware sites (see Figure 1a).

This shows that the opposition's creative output is steadily rising (the variants) but that the mechanisms they use appear to have fads (the sites). We nevertheless propose a Phishing Index that rolls all four together (see Figure 1b).

Commtouch is a company that provides antispam and virus outbreak protection. The world they measure is very noisy (has large variances), and there is a background of steady increase. A proposed Spam Index, using November 2005 as the basis point, might then look like Figure 2.

The National Vulnerability Database publishes a daily number that is a priority-weighted sum of the important vulnerabilities that information technology security operations staff must work to address — that is, smaller is better. Since October 2006, that number has had a slight downward trend, heavily masked by a large amount of day-to-day variability, with spikes not just on second Tuesdays. Figure 3 shows the Daily Workfactor Index, marked for its high and low values so far.

And for the fourth of four, the Open Security Foundation's dataloss database has been collecting for some time, but early records are spotty enough that we chose to begin with January 2006. All we're using is the number of breaches per unit time and the number of persons exposed per unit time. Because the latter has a big dynamic range — from a few records lost to TJX's 94,000,000 — the data has to be smoothed with a moving average. We chose to combine breach count and person count, using them together to arrive at an overall Dataloss Index (see Figure 4)

Combining the four indexes, at least provisionally for this month's

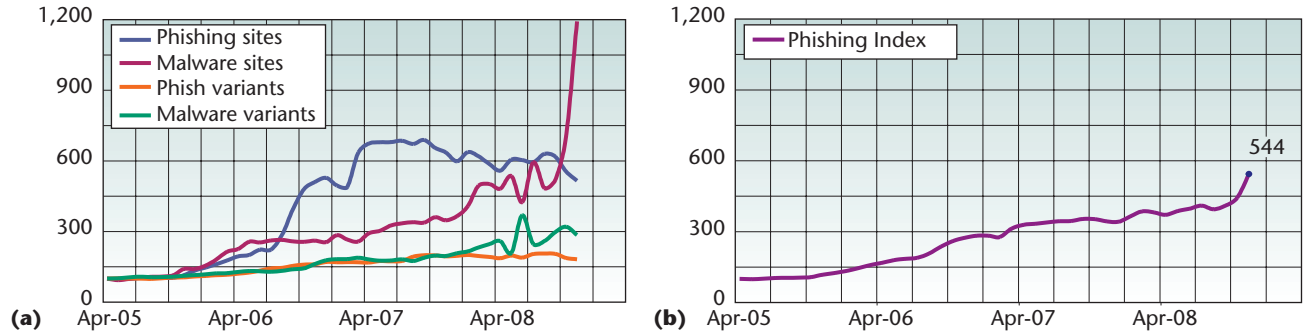

Figure 1. Phishing data. (a) We plot the numbers of phishing variants, phishing sites, malware variants, and malware sites. (b) The Phishing Index rolls all four measurements together.

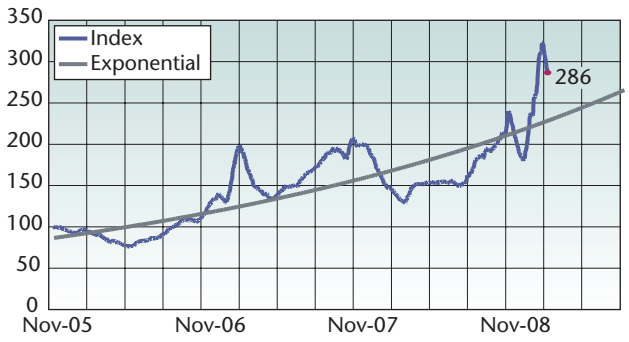

Figure 2. The Spam Index.

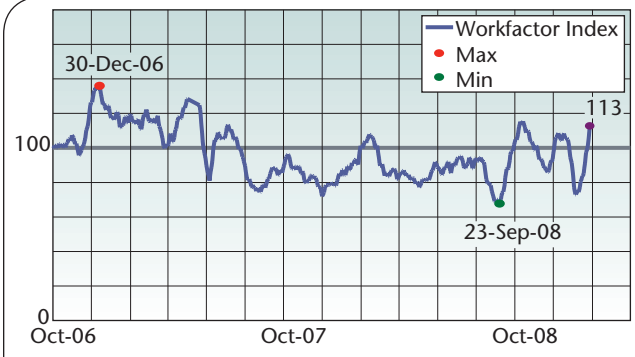

Figure 3. Daily Workfactor Index.

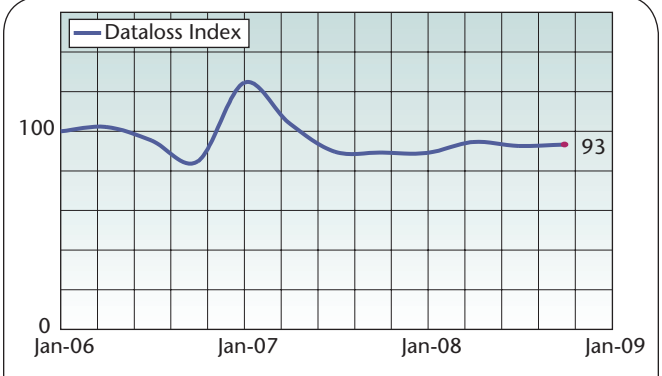

Figure 4. Dataloss Index.

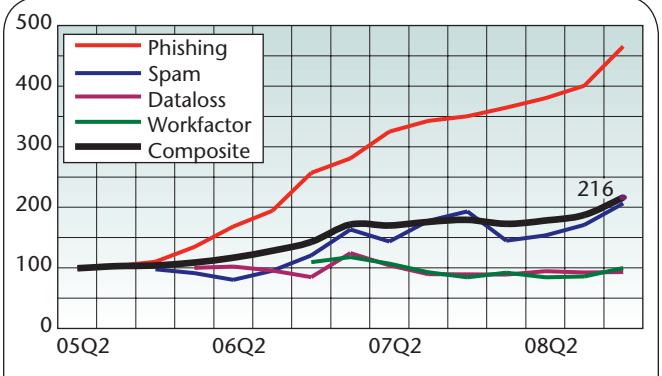

Figure 5. All four indexes combined.

column, results in the graph in Figure 5.

We're going to explore this a bit more, but starting next issue, we'll begin offering up the current value of these indices or ones like them, probably including them in a single paragraph as sparklines plus the current number. Once we start, we won't stop unless our data sources go away. We're out of space this issue, so see you next time—and keep those cards and letters coming. □

Daniel E. Geer Jr. *is the chief information security officer for In-Q-Tel. He was formerly vice president and chief scientist at Verdasys, and is a past president of the Usenix Association. Contact him at dan@geer.org.*

Daniel G. Conway *is an associate professor of business administration at Augustana College. He previously served as associate professor at the University of Notre Dame and Indiana University. Contact him at danielconway@augustana.edu.*