For Good Measure

Five Years of Listening to Security Operations

DAN GEER

Dan Geer is the CISO for In-Q-Tel and a security researcher with a quantitative bent. He has a long history with the USENIX Association, including officer positions, program committees, etc. dan@geer.org

Mukul Pareek, a colleague at a market maker bank, and I have run the Index of Cyber Security for five years [1]. This article is a kind of compendium of what the Index has shown over those five years, but before I get to that I will discuss how we got to where we are.

The only purpose that makes security metrics worthy of pursuit is that of decision support, where the question being studied is more one of trajectory than exactly measured position. None of the indices I'll discuss are attempts at science, although those that are in science (or philosophy) will also want measurement of some sort to backstop their theorizing. We are in this because the scale of the task compared to the scale of our tools demands force multiplication—no game play improves without a way to keep score.

Early in the present author's career, a meeting was held inside a major bank. The CISO, a recent unwilling promotion from Internal Audit, was caustic even by the standards of NYC finance. He began his comments precisely thus:

> Are you security people so #$%&* stupid that you can't tell me:
>
> - How secure am I?
> - Am I better off than I was this time last year?
> - Am I spending the right amount of money?
> - How do I compare to my peers?
> - What risk transfer options do I have?

Twenty-five years later, those questions remain germane. The first, "How secure am I?" is unanswerable; the second, "Am I better off than I was this time last year?" is straightforward given diligence and stable definitions of terms; the third, "Am I spending the right amount of money?" is evaluable in a cost-effectiveness regime, although not in a cost-benefit regime; the fourth, "How do I compare to my peers?" can only be known directly via open information or indirectly via consultants; and the fifth, "What risk transfer options do I have?" is about to get very interesting as clouds take on more risk and re-insurers begin pricing exercises in earnest.

The argument for an index is that when measurement is hard, process consistency is your friend. If we can find one or a few measures that can be tracked over time, those measures, those base numbers do not have to be guaranteed correct—so long as any one series is wrong with some sort of consistency, its wrongness doesn't change the inferences drawn from it. In our kind of work, it is the shape of the trendline that matters. Decisions are supported when we know what direction something is going.

As an example, for some years the National Vulnerability Database has published a daily number called the "Workload Index" [2], which is a weighted sum of current vulnerabilities in the NVD. To quote from NIST:

> [The Workload Index] calculates the number of important vulnerabilities that information technology security operations staff are required to address each day. The higher the number, the greater the workload and the greater the general risk represented by the vulnerabilities. The NVD workload index is calculated using the following equation:

> {
>
> (number of high severity vulnerabilities published within the last 30 days)
>
> +
>
> (number of medium severity vulnerabilities published within the last 30 days)/5
>
> +
>
> (number of low severity vulnerabilities published within the last 30 days)/20
>
> }
>
> / 30
>
> [In other words, t]he index equation counts five medium severity vulnerabilities as being equal in weight with 1 high severity vulnerability. It also counts 20 low severity vulnerabilities as being equal in weight with 1 high severity vulnerability.

Ten years of the NVD Workload Index is shown in Figure 1.

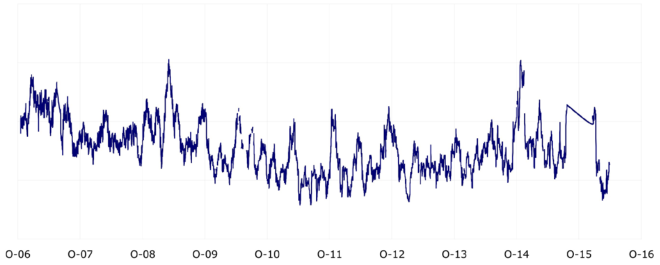


Figure 1: Ten years of the NVD Workload Index

The NVD Workload Index encourages a particular inference: that the arrival rate of new vulnerabilities approximates a random process. Graphing the Workload Index in the aggregate and comparing that to a Gaussian bell curve shows a fair congruence with some right-skew and a bit of kurtosis, as seen in Figure 2.

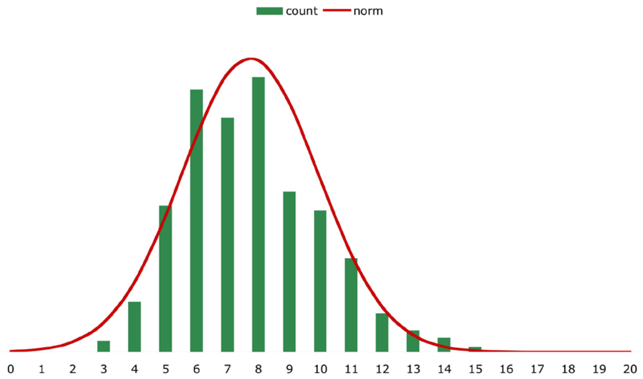


Figure 2: Daily workload number by prevalence

Looking at other methods of binning the Index values, Figure 3 shows some strong variation year over year,

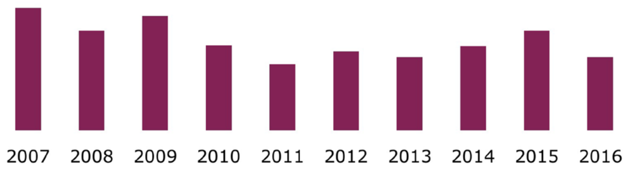


Figure 3: NVD workload year by year

Figure 4 shows seeming seasonality,

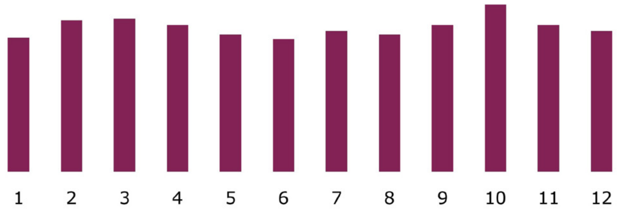


Figure 4: NVD workload month by month

and Figure 5 shows a pretty clear implication of work week.

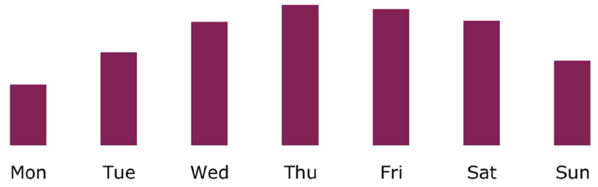


Figure 5: NVD workload day by day

In short, the NVD workload is a straightforward example of an index. I would argue that to be a useful index there has to be something to measure that, once measured, might help one to make some decisions. I would also argue that to be believable, there has to be some transparency as to methods—especially regarding the parameters of sampling—and a believable willingness to carry out a relatively unexciting routine indefinitely. Thank you, NIST, for your long-term diligence in this and so many other things.

Security Pressure Index

Before Pareek and I began the Index of Cyber Security, I had tried various indices before. A different colleague, Dan Conway, and I put together what we called the “Security Pressure Index,” meaning an estimate of the time rate of change in the pressure on security professionals. With indices, seeking generality usually means that you want more than one input. We settled on four: we got a measure of phishing from the Anti-Phishing

Working Group, a measure of spam from Commtouch, a measure of data loss from the Dataloss Database, and that measure of workload from NIST. Together, these four yielded the Security Pressure Index as shown in Figure 6.

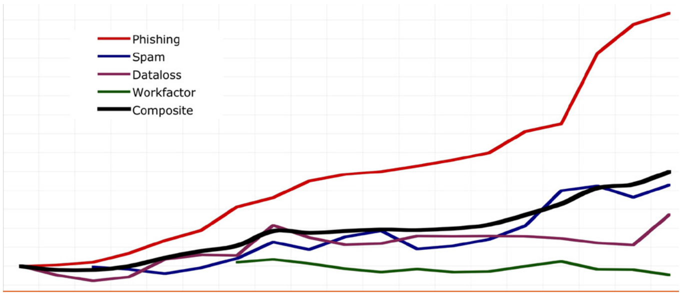
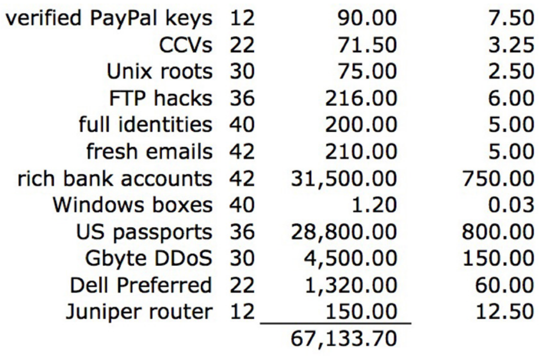


Figure 6: Five years of the Security Pressure Index

To be clear, in each case we were mooching off other people's work, which is rarely polite and leaves you no recourse should, say, one of those sources change its numbering scheme, change its publication schedule, or change anything at all without telling you. We thanked all four sources in print every month, but what we were doing was, in so many words, predictably unsustainable. After five years, we called it quits for the SPI. Close, perhaps, but no cigar.

0wned Price Index

So what next to try? With the aphorism "Good artists copy, great artists steal" in mind, Conway and I ripped off PNC Financial's long-running "Christmas Price Index" [3]. The XPI, as it is called, calculates the price of buying all the gifts described in the song "The Twelve Days of Christmas" such that you might know what your true love's affection is going to cost you. In our case, we put together a price index for stolen data and similar illicit digital goods. To get a little attention, we called it the "0wned Price Index," and after nailing down a variety of stolen goods for which market price information paralleled the XPI, we amalgamated them in a way that could, after a fashion, actually be sung to the tune of "The Twelve Days of Christmas" [4]. We published this for three years (see Figure 7).



verified PayPal keys	12	90.00	7.50
CCVs	22	71.50	3.25
Unix roots	30	75.00	2.50
FTP hacks	36	216.00	6.00
full identities	40	200.00	5.00
fresh emails	42	210.00	5.00
rich bank accounts	42	31,500.00	750.00
Windows boxes	40	1.20	0.03
US passports	36	28,800.00	800.00
Gbyte DDoS	30	4,500.00	150.00
Dell Preferred	22	1,320.00	60.00
Juniper router	12	150.00	12.50
		67,133.70	

Figure 7: 0wned Price Index for Christmas 2009

And then we stopped. The reason we stopped was a kind of progress. The market price data we relied upon came from eavesdropping on so-called carder forums and the like—places where stolen data was sold at auction. But those sources of information dried up once law enforcement infiltrated them and began making arrests. After that, to get auction pricing you had to be a market participant, but Conway and I were not ready to be market participants. Repeating myself, if you rely on data sources you do not control, then what you are doing is inherently temporary.

Index of Cyber Security

Which leads me to the main event for this column. Based on the experience(s) described above and just general knowledge of the field, Pareek and I put together the Index of Cyber Security, which turned five years old in April 2016. The first lesson, that it is better to source your own data if you expect to be in the game for the long haul, means we have to ask our own questions, not just graze in other people's pastures during their growing season.

Another lesson is that, even yet and perhaps forever, as a field we will not be able to agree on precise terminology. Yes, we can all agree that "vulnerability," say, is a term in general use, but as to a fully precise definition, universally held—that's not coming. That, in turn, means that if you ask, "How many vulnerabilities are there?" the answers you get will be biased by the definitions of the individuals answering. This is not completely serious, but terminological confusion substantially interferes with reproducibility of survey results.

As a central point, survey research is vulnerable to idiot respondents. If you are looking for generalizability, you administer your survey to as large a population as you can afford and you pick the people replying either by randomization or by selection. If you randomize, you gain some immunity to idiot respondents. The well known Consumer Confidence Index [5] (CCI) is based on 5000 random phone calls a month, thus washing out the idiot fraction, at least so long as that fraction is not growing. The CCI is run consistently, and many financial instruments factor in the new value of it as soon as it is issued. It is a forward-looking indicator.

If generalizability to the public at large is not a goal, then you administer your survey to a vetted population where there is no idiot fraction. But by selecting your respondents, your results are conditional on the methodology of your selection process. The well-known Purchasing Managers' Index (PMI) [6] picks its respondents carefully and has many fewer of them, but because the PMI respondents are selected for what they know, this is a feature not a bug. The PMI is a weighted sum of five variables, in this case production level, new orders, supplier deliveries, inventories, and employment level. Like the CCI, the PMI is run consistently, and many financial instruments factor in the new value of it as soon as it is issued. It is a forward-looking indicator.

So Pareek and I looked at both the Consumer Confidence Index and the Purchasing Managers' Index for inspiration. Both of them ask subjective questions about the opinion of the respondent. The CCI wants opinions that are representative of the population at large, so they take the randomization route. The PMI wants opinion to be knowledge-based, so they take the vetted respondents route. Pareek and I decided that we would follow the PMI approach, that is, to have as respondents people who actually know something.

But what is it they are supposed to know? We decided that if the Index of Cyber Security was to be a forward-looking indicator, then we had to have as respondents people who are on the front lines, people with operational responsibility for cybersecurity. We do not want people whose knowledge of current cybersecurity is academic, or based on police power, or the result of having memoranda passed up the management chain to them. We wanted people who were doing cybersecurity, not people who had knowledge that didn't come from actual daily practice.

This means that we rely on a certain kind of expert, and the Index of Cyber Security is an amalgamated subjective opinion on the state of play as understood by people who are actually in the game, per se. When I say "subjective" it is because we do not have solid, unarguable measures of security. In fact, that we don't is precisely why we are doing the Index—when you don't have unarguable measures, the next best thing is the collected wisdom of experts. And note that I said "experts"—we neither know nor care what a respondent's official position is in some organizational entity; we care about experts wherever we find them. So it may be that some handful of experts work for the same employer, and some employers will have no experts present at all. So be it; we are not collecting insights into the Fortune 500—we are collecting experts.

Because every term we might use has, as I mentioned before, some degree of ambiguity as a term, we cannot just ask, "How bad is malware?" Asking "How bad is malware?" requires a precise, shared definition of "malware" and a malware thermometer that reads "78" or the like. So what then do you do?

What we do is ask a series of 25 questions, and the questions are the same every month. All of the questions read like this:

> Since a month ago, the threat of insider attack has
>
> - Gotten Better
> - Gotten Worse
> - Gotten a Lot Better
> - Gotten a Lot Worse
> - Remained Unchanged

We ask 25 questions like that.

There are two things to note at this point: one, you may recognize the response set as a Likert scale. Likert scales are standard practice in survey-based research. They are always symmetric with an odd number of options so that the central option is considered neutral. The score for a question is a weight assigned to each of the alternatives.

The main point here is that each question is of the form "Since a month ago," meaning that what we are looking for is change, not valuation. That is far easier to estimate reproducibly than estimating a number in an absolute range. The rest of the question, "the threat of insider attack has gotten," does not require everyone to agree on what insider or attack means. We do not have to train our respondents to use this or that word precisely in one way that might differ from how they usually use it. All we need is for the individual respondent to have a mental definition of the word or phrase that is reasonably stable. If your definition of, say, "malware" and mine are subtly different, we can still say whether the pressure from it has gotten better or worse.

In other words, the Likert scale's symmetry avoids biasing the respondent in one direction or the other. Additionally, by asking about the trend of a characteristic rather than the value of some measurement of that characteristic, the respondent is relieved of having to conform to either some official definition or to a scaling mechanism they did not invent. Instead, they can use their own definition and don't need numbers.

Because each question is of the same form, the Index of Cyber Security is then calculated by counting how many "Gotten Better" answers, how many "Gotten Worse" answers, etc.,—one count for each Likert category. Those counts are combined in a weighted sum:

Much Better	Better	Unchanged	Worse	Much Worse
-20%	-7.50%	0	7.50%	20%

Being a measure of risk, the ICS is bounded on the low side but not on the high side, hence the directionality of the weightings. In other words, the ICS rises as perceived risk rises. An example:

Much Better	Better	Unchanged	Worse	Much Worse
6	58	614	150	15
-20%	-7.50%	0	7.50%	20%

Multiplying it out and dividing by the sum of the above, we get 0.010235. Exponentiating that gets a multiplier to apply to last month's ICS to get this month's, i.e., 1.010287, or an increase in the ICS of a tiny bit over 1%.

We do this calculation not only in the aggregate so as to derive the Index of Cyber Security value, but also on a question-by-question basis so as to watch trends in specific risks. These trendlines by question we refer to as sub-indices, and they are part of a detailed monthly report that only respondents get.

And, yes, we occasionally replace one question with a new one. To maintain continuity of the ICS as a whole, we apply a correction factor done in precisely the same way that any financial index such as the Dow Jones Industrial Average does when it replaces one stock with another.

Perhaps you did not need to know all that, but our point is that the way the ICS is calculated is 100% conventional and entirely boring. We want "boring" because whatever our results, we want them to never be thought of as an artifact of some new method we cooked up on the spot. Much as amateurs should rarely create their own crypto algorithms, amateurs should rarely create their own analytic regimes.

So this is the scheme—a largely fixed set of Likert-valued questions, a vetted respondent base, a trade of data for data, and a commitment to a long run. This is information sharing at its best.

What we have learned so far: our respondents believe that risk in the aggregate is and has been rising almost inexorably, but which of the 25 components of the ICS is changing the most each month varies over time—a lot—as seen in Figure 8,

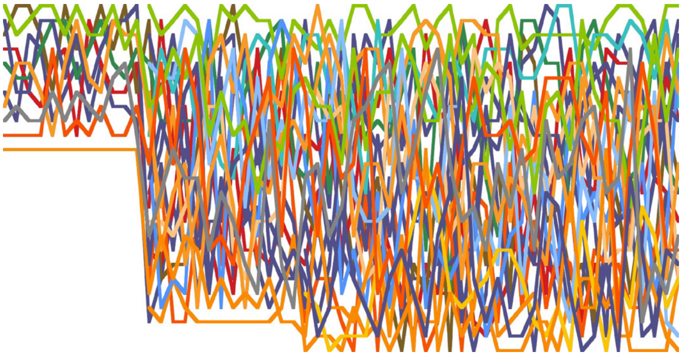


Figure 8: Rank order % change across sub-indices month by month

which can also be seen looking at the trailing four-month volatility of the sub-indices in Figure 9.

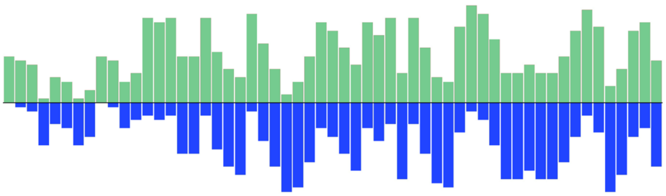


Figure 9: Trailing four-month volatility up versus down

Another way of looking at dispersion of risk across questions is that for 14 of the 58 months seen in Figure 10, at least one question reached its lowest value, and in 14 of those months at least one question reached its highest value.

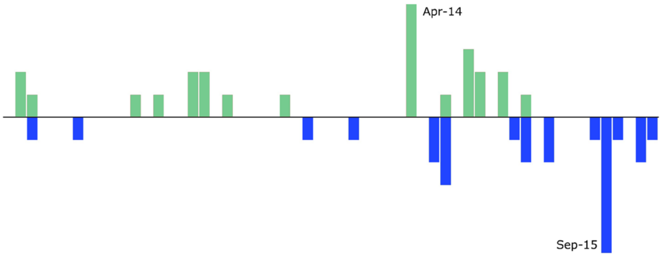


Figure 10: Trailing four-month volatility up versus down, highest and lowest values

In five of those months, both one question's highest and another question's lowest values were set, and in 32 of those months, no question reached its most extreme value. In April 2014, 20% of the questions returned their highest values ever for rate of change. In September 2015, 25% of the questions returned their lowest values ever for rate of change. "Why?" is hard to guess.

Let me be clear that we are not trying to do science here. If your purpose in building a model is to come to a definitive conclusion about causality, about how nature works, then you are saying that the inputs to your model and the coefficients that calibrate their influence within your model are what matters. Parsimony in the sense of Occam's Razor is your judge, or, as Saint-Exupéry put it, "You know you have achieved perfection in design, not when you have nothing more to add, but when you have nothing more to take away."

By contrast, when your purpose in building a model is to enable control of some process or other, then you will not mind if your input variables are correlated or redundant—their correlation and their redundancy are not an issue if your goal is to direct action rather than to explain causality. A goal of understanding causality in its full elegance leads to $F = ma$ or $E = mc^2$. A goal of control leads to econometric models with thousands of input variables, each of whose individual contribution is neither clear nor relevant.

That said, if you look month by month you see that some questions are perceived to indicate more risk than others. Ranking the magnitude of individual risks over a 58-month period gives us:

Risk	Number of times in the top three for the month
Counterparty	52
Media & public perception	40
Hacktivist/Activist	30

If you rank not by risk score but by which risk had the biggest jump (volatility) that month, then you find

Risk	Number of times in the top three for the month
Media & public perception	42
Phishing/Social engineering	35
Counterparty	16

With the usual caveats about correlating too many things at once, if you put all 25 current questions into a correlation matrix, then some do appear to be in lock step.

Correlation	Risk Pairs
0.971	Effect desired: Data theft Weapons: Phishing/Social engineering
0.958	Effect desired: Data theft Attackers: Criminals
0.946	Overall: Media & public perception Weapons: Phishing/Social engineering
0.940	Weapons: Phishing/Social engineering Attackers: Criminals
0.931	Overall: Media & public perception Effect desired: Data theft

Given that array, one could argue that there is really only one risk between all of those: the risk of data theft by criminals using social engineering so that you look stupid in the newspaper.

Of course, one thing that we wish we had done from the get-go was to record the dates of important security events, whether that is in the newspaper, the laboratory, or the underworld. We didn't, and we're not going to start now. But when you look at all the variation, we do often want to say, "Where did that come from?" We can't answer that, so we won't make believe we can; to do so given our methods would be pure speculation [7].

We also compute for each risk and overall a diffusion index and do it the same way as diffusion indices are done in finance. Diffusion indices are a symmetric construct; they are just the sum of all the indicators in a basket of indicators that are going in one direction plus half of those that are static. As the ICS is a risk index, we report what percentage of responses are either "Worse" or "A Lot Worse" plus half the responses that are "Neutral." For January of this year, the top three were

Risk	Diffusion Index
Phishing/Social engineering	69%
Criminals	63%
Customization to target	62%

Of the 25 risks, five of them had diffusion indices of 50% or less. The other 20 were above 50%.

One final thing; each month, in addition to the standing set of 25 questions, we ask a question of the month. Once in a while these are suggested. Most of the time Pareek and I think them up. In 2015, Questions of the Month covered encryption, safe harbor, ransomware, IPv6, affordability, change management, CEO involvement, regulation, worst case scenarios, security metrics, and offensive dominance.

Sometimes we will repeat a question. For example, in September of 2012 we asked, "What percentage of the security products you are running now would you still run if you were starting from scratch?" In January of this year we asked that question again. Compiling the answers, we found in September of 2012 that 35.5% of the products then installed would not be reinstalled should the respondents be in a position to start fresh. Call that buyers' remorse. In January of this year, we found that buyers' remorse had swelled from 35.5% to 51.9%. I don't have figures for the number of cybersecurity products available for sale month by month, but it is surely greater now than it was three and a half years ago. I can tell you from where I work that the number of cybersecurity startups has never been greater; a spokesman for Kleiner-Perkins says that they are tracking over 1100 cybersecurity startups now in some part of the funding game. Is a rising level of buyers' remorse a sign that better tools are on offer or that unmitigable risk is getting worse? It's a puzzle.

Conclusion

This seems a good place to stop insofar as it is surely possible to just keep doing exploratory data analysis for pages more. But that isn't actually what I have been doing. What I've been doing is talking about a different kind of information sharing, bottom up, as it were. All the talk about information sharing always seems to mean something top down, something where those with more power or better eyes or an enforceable structural advantage share a portion of their information trove with the worthy below them. I am not making fun of that; it is a time-proven technique and it is policy across the board. It comes out of the idea of "need to know," and need to know is a protective mechanism in so many things. Yet it seems to us that once upon a time any one of us could start from nothing and, by diligence, come to know just all that was necessary for cybersecurity. That is clearly less true than it once was. The technical knowledge base has both deepened and broadened, deepened in that sense of an accumulating welter of obscure interdependencies, and broadened in that sense of cybersecurity becoming an issue wherever networks go.

That affects need to know in ways we have only barely acknowledged. Sure, the Federal government, or any Western government, can grant security clearances to the CISOs of every market maker bank, or any other institution that matters to them, so as to share classified information.

But, for Pareek and myself, the argument for official channels is unsatisfactory and insufficient. We don't mind them, but cybersecurity in its complexity just doesn't seem to us to be headed for some sort of denouement when all will become clear at taxpayer expense. We are doing the Index of Cyber Security the way we are on the grounds that (1) you can't know what's going on unless you are on the playing field yourself, and (2) that there is no way to tell if the risks you are seeing are specific to you without comparing your risks to those of other people in your position elsewhere.

In the fullness of time, we may add other things to our repertoire, but we are expecting to keep doing the ICS for the indefinite future. We invite you to participate. The respondent's workload is insignificant, the shared data cannot be gotten elsewhere, and we are doing everything we know to do to make it possible for respondents to be frank without concern to being quoted in any way. To take part in this project, see the Contact page under reference [1].

References

[1] Index of Cyber Security: cybersecurityindex.org.

[2] NIST Workload Index: nvd.nist.gov/Home/Workload-Index.cfm.

[3] PNC Christmas Price Index: www.pncchristmaspriceindex.com/cpi.

[4] D. Geer and D. Conway, "What We Got for Christmas," *IEEE Security & Privacy*, January 2008: geer.tinho.net/ieee/ieee.sp.geer.0801.pdf.

[5] Conference Board Consumer Confidence Index, issued at 10 a.m., Eastern Time, on the last Tuesday of every month: www.conference-board.org/data/consumerconfidence.cfm.

[6] Institute for Supply Management, Purchasing Managers' Index, issued at 10 a.m., Eastern Time, on the first business day of every month: www.instituteforsupplymanagement.org/ismreport/mfgrob.cfm.

[7] If you want to read the best talk ever given on speculation, the late Michael Crichton nailed it in 2002 with "Why Speculate?" archived at geer.tinho.net/crichton.why.speculate.txt.