Shades of Green

one advisor's views on the direction of CyberGreen statistics

---

Security metrics are desirable when they enable something, when they have a role to perform that has a receiver ready to make use of them. Otherwise they are stamp collecting.

The issue is one of purpose. The only purpose that makes security metrics worthy of pursuit is that of decision support, where the question being studied is one more of trajectory than exactly measured position. We are not in this for reasons of science, though those that are in it for 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 demand 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 market maker bank. The CISO, who was a recent, unwilling promotion from Internal Audit, was caustic even by the standards of NYC finance. He began his comments mildly enough:

> 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 is unanswerable, the second is straightforward given diligence and stable definitions of terms, the third is evaluable in a cost-effectiveness regime though not in a cost-benefit regime, the fourth can only be done directly via open information or indirectly via consultants, and the fifth is about to get very interesting as clouds take on more risk and re-insurers begin pricing exercises in earnest.

It stands repeating that the core Internet protocols were designed against a particular goal state: optimal resistance to random faults in the network fabric. There is no need to elaborate on that here

except to remember that it is impossible to have a network design that is optimally resistant to random faults and optimally resistant to targeted faults. The CyberGreen effort is logically focused on resistance to targeted faults, especially targeted faults that have effect at something approximating global scale.

There are two genus and five species of cyber attacks:[1]

Passive, *i.e.*, pure listening attacks
  Traffic Analysis — who is talking to whom
  Release of Contents — who is saying what to whom
Active, *i.e.*, packet insertion attacks
  Message Stream Modification — change what they said
  Denial of Service — don't let them talk
  Spurious Association Initiation — false claim of identity

Because (attempts at) passive attacks cannot be detected, they must be prevented. Because (attempts at) active attacks cannot be prevented, they must be detected. Ergo, the possible goals for any communications security technology:

Prevention of traffic analysis attacks
Prevention of release of contents attacks
Detection of message stream modification attacks
Detection of denial of service attacks
Detection of spurious association initiation attacks

Of those, CyberGreen has nothing directly to do or say about (the prevention of) passive attacks. CyberGreen also has nothing directly to say about false claims of identity ("spurious association"). CyberGreen's number one goal is to detect the precursors to attacks, especially to denial of service attacks; assuming that the motivation to attack will always be available, our job is measure its opportunity. While it is true that vulnerable endpoints can be used for any bad activity, we cannot measure some kinds of bad activity so must stick to those we can measure and, to the point, measure in a way that is solid decision support.

At the outset of Obama's first term, Hathaway led a "sixty day review"[2] of the U.S. cybersecurity stance. She concluded that the primary targets were members of the Defense Industrial

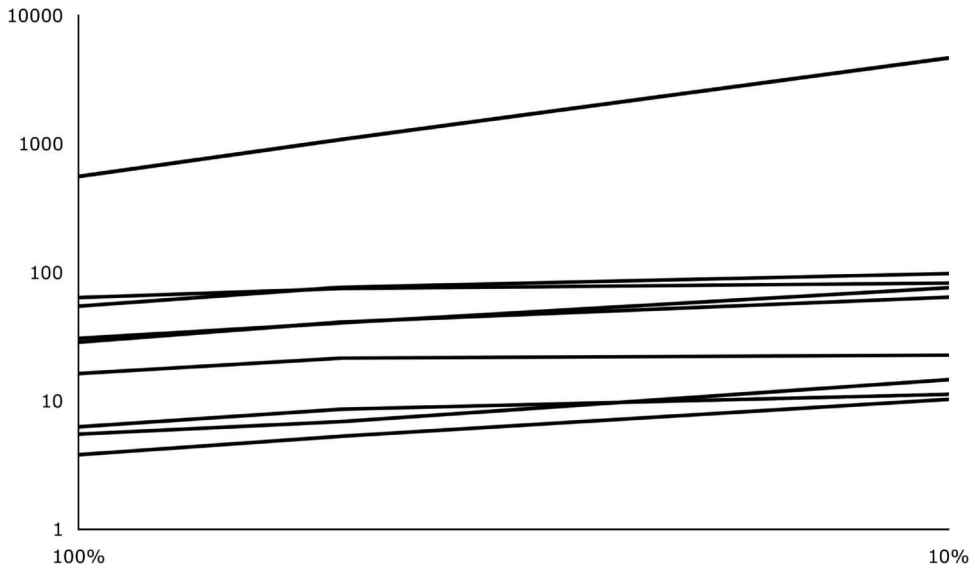
Base and technology firms with global reach, that the secondary targets were the primary targets' counterparties, and that the tertiary targets were any endpoints that could be used as staging areas for attacks on the primary and secondary targets. It is the latter group — the endpoints that have value as staging areas — that CyberGreen can measure. Nation states, like the U.S., will protect their primary targets and will not look to CyberGreen for assistance. It seems certain that risk-carrying interactions between primary and secondary targets will not be observable to CyberGreen in any useful way. Ergo, consistent with the previous paragraph, that leaves to CyberGreen the particular focus on entities that can be used to stage attacks, attacks that could, themselves, reach national consciousness in various nations.

CyberGreen's attention to opportunities for amplification attacks takes that conclusion to heart. Although slightly dated, Rossow's 2014 data on amplification attacks[3] serves as a good reminder of the need, and continuing opportunity, for data collection on the opportunity for amplification:

	BAF			PAF	
Protocol	all	50%	10%	*all*	Scenario
SNMPv2	6.3	8.6	11.3	1.00	*GetBulk* request
NTP	556.9	1083.2	4670.0	10.61	Request "monlist" statistics
DNS-NS	54.6	76.7	98.3	2.08	ANY lookup at author, NS
DNS-OR	28.7	41.2	64.1	1.32	ANY lookup at open resolv.
NetBios	3.8	4.5	4.9	1.00	Name resolution
SSDP	30.8	40.4	75.9	9.92	*SEARCH* request
CharGen	358.8	n/a	n/a	1.00	Character generation request
QOTD	140.3	n/a	n/a	1.00	Quote request
BitTorrent	3.8	5.3	10.3	1.58	File search
Kad	16.3	21.5	22.7	1.00	Peer list exchange
Quake 3	63.9	74.9	82.8	1.01	Server info exchange
Steam	5.5	6.9	14.7	1.12	Server info exchange
ZAv2	36.0	36.6	41.1	1.02	Peer list and cmd exchange
Sality	37.3	37.9	38.4	1.00	URL list exchange
Gameover	45.4	45.9	46.2	5.39	Peer and proxy exchange

TABLE III: Bandwidth amplifier factors per protocols; *all* shows the average BAF of all amplifiers, 50% and 10% show the average BAF when using the worst 50% or 10% of the amplifiers, respectively. The Packet Amplifier Factor is a function of the protocol, *per se*.

Notably, each of Rossow's top 10 read like power law curves:



Straight lines on a log-log graph => power law

While this is a small demonstration with someone else's thin data, it does raise a point. In a difficult paper,[4] Nassim Taleb trenchantly said that "[We are] undergoing a switch between [continuous low grade volatility] to ... the process moving by jumps, with less and less variations outside of jumps". Put differently, CyberGreen must be continuously on guard against implying that things are getting better and better if our argument for that is based on an assumption of Gaussian error and/or the Law of Large Numbers: When a distribution is fat-tailed, estimations of parameters based on historical experience will inevitably mislead. If CyberGreen can demonstrate power law relationships — and explain it well for those unfamiliar with the concept — then a real contribution will have been made.

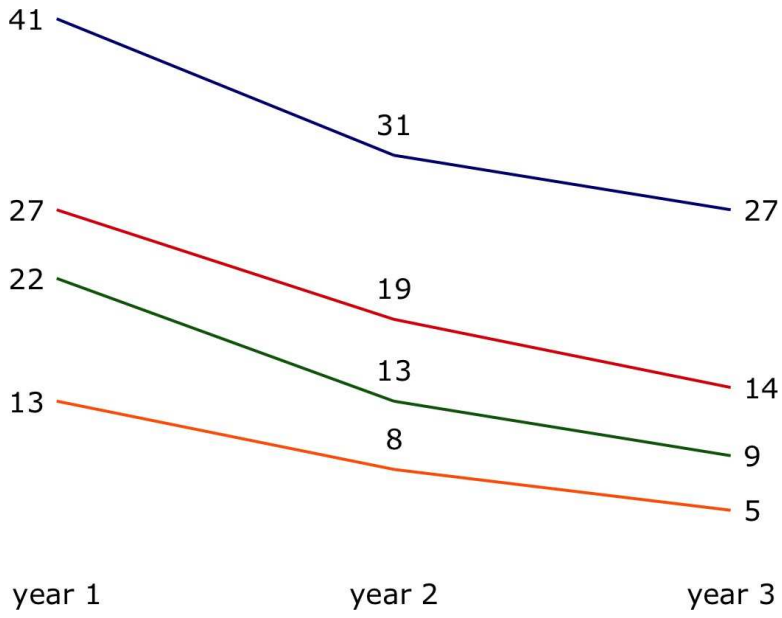
CyberGreen's base metrics today are

$$G = 1 - \frac{(C + V)}{2} \quad \text{and} \quad PR = \frac{L + (0.5 \times S)}{N}$$

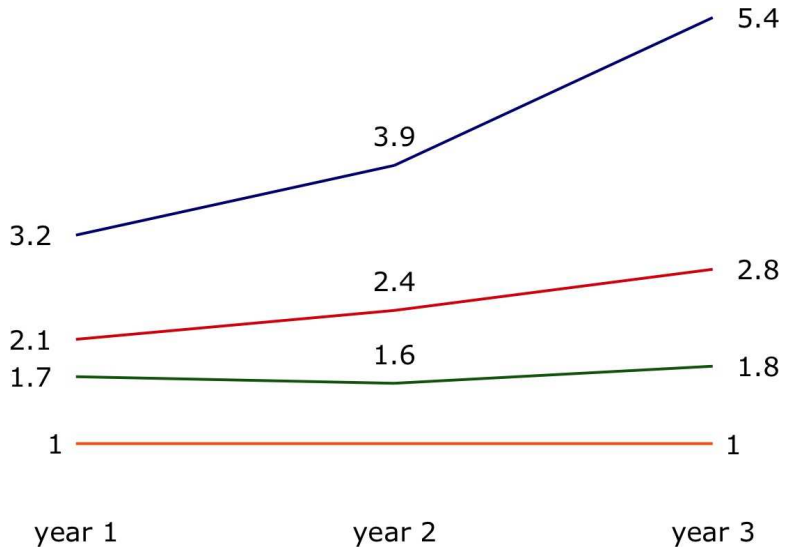
The "Percent Rank" (properly *Percentile Rank*) is mostly used, it would appear, in evaluating the results of educational testing. PR is independent of any underlying statistical distribution, which is

wise. An entirely similar construct — called a "diffusion index" — is found in stock trading where it is used in technical analysis to predict which way a market is going. Whereas the percentile rank phraseology seems largely confined to scoring some collection of one-time measures, the diffusion index phraseology seems largely confined to scoring some collection of trends. Here one might ask if CyberGreen's calculation ought to be on trends rather than current values. The mechanics of calculation are the same (the number in one direction plus one-half the unchanged divided by the total). If CyberGreen concentrated on trends, then any "F" score would be at once more deserved and less arguable.

A former employer, @stake, had a large enough customer list that we could pool our data without identifying any one customer. Amongst our customers, everyone was getting better but the best and the worst were diverging. One suspects that such divergence in the combined CyberGreen database will evidence itself as our time series grow. To illustrate, this is real data from that prior life:



Number of fatal security flaws per major application by quartile



The same data normalized to the first/best quartile

CyberGreen should probably report measures both of position and velocity, the former as it may permit some consumers to crow about their position while permitting other consumers to crow about their velocity of improvement. This makes sense — the best students will never get the "most-improved" award and *vice versa*. As to the velocity, the developed math around diffusion indices allows one to combine multiple timeseries into one timeseries so as to make prediction for the near term future, so CyberGreen should very much attend to this possibility.

In counter-proposal to various CyberGreen Version 1 screenshots, what might be most useful to a policy-making representative of a sovereign (country) is something more like this:

Your best score:	Open Recursive DNS Resolver
Next-best score:	Conficker Nodes
... :	Another issue
... :	Another issue
... :	Another issue
... :	Another issue
... :	Another issue
Your worst score:	UCE nodes

with perhaps some text to say that you'd do the most for your overall score by converting your worst scores into your best scores.

Accepting for the moment G, the existing Green Index, and its dependent relationship on PR, this author suggests that having to reverse G by subtracting the average of C and V from 1 is needless and should be repaired before much more time has elapsed. Instead, the PR equation is what should be reversed. This would mean, in effect, having C be the PR of "Uncompromised Nodes" and V the PR of "Invulnerable Nodes". In other words, using the variable names as currently given, the equations become

$$G = \frac{(C + V)}{2} \quad \text{and} \quad PR = \frac{H + (0.5 \times S)}{N}$$

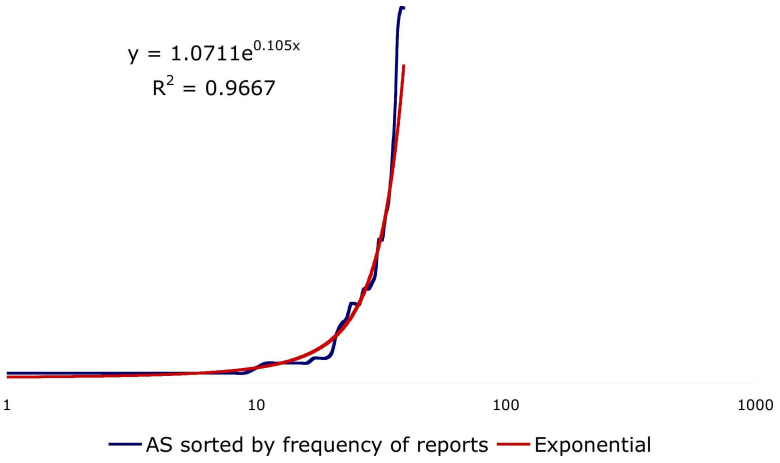
which is to say that PR should be based on $H = Number\ of\ higher\ rank$ rather than $L = Number\ of\ lower\ rank$. In other words, it would seem to this author that a metric that is non-intuitive can provide little decision support.

In any case, this is the moment to beg the question: Is it more important to give persons of authority a chart with a dot that says "You are here" or is it more important to give persons of authority a chart with a line that says "This is your future unless you do something about it"? Perhaps both, but this reviewer would argue that the latter — the implications of trajectory — is what really matters to policy-setters.

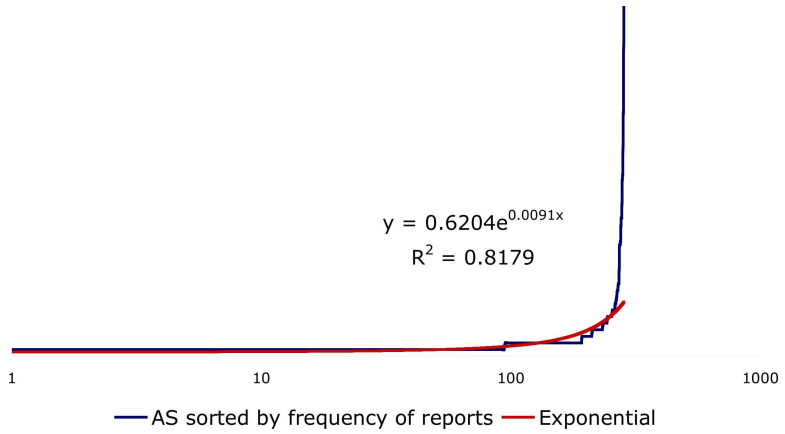
Nevertheless, if we can assume that compromised nodes and vulnerable nodes are non-overlapping sets, *i.e.*, we don't double count them, then we know that all the compromised nodes were once vulnerable (but uncompromised) nodes. Therefore, if C is the number of currently compromised nodes and V is the number of currently vulnerable nodes, then the figure $C/(V + C)$ tells us what fraction of now or previously vulnerable nodes have been compromised. If we are lucky, this might be a correlate of attack pressure within that entity's jurisdiction. This is a figure that is self-normalizing insofar as total node count does not, even by implication, factor in. We still need total node count, of course, but for the moment having an estimate of the attack pressure is at hand even if we have no bloody idea what the sampling fraction of C and/or V really are. We might even

get to doing an analysis of variance (with a categorical variable for jurisdiction and a continuous variable for attack pressure). In any case, we certainly need to understand the volatility of C and V. As always, the volatility of their ratio will be higher and analysis of that volatility (variance) is likely to be instructive.

The CyberGreen data already makes quite clear one thing: the most telling unit of observation is that of per-ASN risk. Because of data volume, let's just take two samples of the data, one for February 20th and one for May 15th, both for the same three hour clock-time. Because CyberGreen is receiving more and more data per unit time, there are more reports for May 15th than for February 20%. From each of those two samples, we count unique ASN occurences, sort those counts into rank order, and plot the result, first for February 20th:



and secondly for May 15th:



The exponential lines are there merely to underscore how much some ASNs are worse than others, and that that disparity is perhaps increasingly so. In any case, there are policy implications here.

When trying to tease out subtle trends, ratios are indeed the tool of choice, just as with our existing *PR* and *G*. Odds ratios have particular appeal. Quoting colleague Paul Vixie, he is flatly correct that "epidemiology has shown high utility in understanding broad scale events, but rarely aids in risk management by individual potential victims". The combination of Vixie's remark and the present author's favor of odds ratios has to lead somewhere. So to illustrate a point hinted at earlier, let's start with a simple 2x2 table with indicative labeling:

	compromised	exposed
country #i	V_i	$C_i + V_i$
others	$V_{\bar{i}}$	$C_{\bar{i}} + V_{\bar{i}}$

and fill it in with not so very mock data[5]

	C	*V*	
US	164,846	127,147,768	127,312,614
others	9,581,997	1,154,778,261	1,164,360,258

$$Prob(US) = 0.129\% = 164,846 \,/\, 127,312,614$$

$$Prob(others) = 0.823\% = 9,581,997 \,/\, 1,164,360,258$$

leading to the ratio of probabilities

$$\frac{Prob(others)}{Prob(US)} = \frac{0.823\%}{0.129\%} = 6.36 \equiv \text{Relative Risk (RR)}$$

$$\frac{RR-1}{RR} = \frac{6.36-1}{6.36} = 84\% \equiv \text{Attributable Risk Percent (ARP)}$$

which means that if you have in your hands a packet from the US and a packet from somewhere else on the planet, the one from somewhere else is 6.36 times as likely to be slime. Further, if you are accepting a flood of packets from everywhere at all times, 84% of your (risk of) slime will not be from the US. In short, RR and ARP seem like numbers that could get some attention without having to have an enormous educational effort for the recipients. By sticking this sort of measure, we do not have to defend our base numbers as guaranteed correct — so long as all our sampling is wrong with some sort of consistency their wrong-ness doesn't change the inferences drawn from them.

For the simple reason that this is a dynamic situation, everything that is put on a screen or printed out needs to have an effective date in plain view, "This page created on 15 May 2016" or the like. To make the CyberGreen website really excellent, we need a way to say "Show me the data for 21 November 2015" and the like so that the reader can look at CyberGreen data "as of" such and such a date. If that requires structure that is not now in the database, then that structure needs to happen before it is too late to do it at modest labor cost.

Perhaps the reader is already aware of how good the report from Qualys' SSL Labs is;[6] there are things in it to copy, perhaps especially how they say that "Because you use XYZ, your grade is capped at C." Joe St. Sauver's work for the FCC has many useful and perceptive ideas.[7] The CBL at `abuseat.org` is pretty good though the lead-in paragraphs and the actual display do not match well. For CyberGreen, on any of its data pages a worked example will be essential; even though the fraction of readers who actually do RTFM is doubtless small, if the worked example is not clear or not correct, the person reading it will conclude disparaging things about the rest of the site. (Reminds one of why airlines wash their planes.) For extra credit, the worked example could use current data so that as the data changes so does the worked example.

The (US) National Weather Service has a nice strategy for forecast information. There is a graphic for the US as a whole. There is a seven-day forecast by postal zip-code that has both icons and terse text. There is a vertical array of forward looking hour-by-hour graphs for all the major weather components. There is a long, unashamedly technical prose discussion of how the (named) forecaster put together the forecast and where in the forecast there is uncertainty — and that prose section is itself in three parts: general, aviation, and marine. Historical data is available by month by weather station and can be gotten in CSV form. Everything can be bookmarked. Javascript is not required. CyberGreen could (read, MUST) hold such a site as a goal state.

This is a very exciting time. We've got a lot to do. Lets get to it.

---

[1] Voydock V & Kent S, "Security Mechanisms in High-Level Network Protocols," *Computing Surveys,* 15:2, June 1983

[2] Hathaway M, White House, "Cyberspace Policy Review," 8 May 2009, `www.whitehouse.gov/assets/documents/Cyberspace_Policy_Review_final.pdf`

[3] Rossow C, "Amplification Hell: Revisiting Network Protocols for DDoS Abuse", Symposium on Network and Distributed System Security, 2014; Legend edited slightly for contextual clarity. Retrieved from `www.internetsociety.org/sites/default/files/01_5.pdf`

[4] Taleb N, "On the Super-Additivity and Estimation Biases of Quantile Contributions", `www.fooledbyrandomness.com/longpeace.pdf`

[5] The numbers, retrieved May 15, 2016 from `www.abuseat.org/countrytraffic.html`, are

the total "Listings" for the US (164,846),
the total "Listings" for the rest of the world (9,851,127 - 164,846 = 9,417,151),
one half the US "IPop" (254,295,536 / 2 = 127,147,768), and
one half the total "IPop" ((2,560,930,603 / 2) - 127,147,768 = 1,154,778,261).

That "one half" is the maximum ignorance estimate of the fraction of all entities that are compromised.

[6] See `www.ssllabs.com`

[7] Example: `www.stsauver.com/joe/maawg-orlando-talk/maawg-orlando-talk.pdf`