Exploring with a Purpose

DAN GEER AND JAY JACOBS

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

Jay Jacobs is the co-author of *Data-Driven Security* and a data analyst at Verizon where he contributes to their Data Breach Investigations Report. Jacobs is a cofounder of the Society of Information Risk Analysts. jay@beechplane.com

> Think of [knowledge] as a house that magically expands with each door you open. You begin in a room with four doors, each leading to a new room that you haven't visited yet.... But once you open one of those doors and stroll into that room, three new doors appear, each leading to a brand-new room that you couldn't have reached from your original starting point. Keep opening doors and eventually you'll have built a palace.
>
> Steven Johnson, "*The Genius of the Tinkerer*" [1]

Learning pays compound interest; as a person studies a subject, the more capable they become at learning even more about the subject. Just as a student cannot tackle the challenges of calculus without studying the prerequisites, we must have diligence in how we discover and build the prerequisite knowledge within cybersecurity.

Before we discuss where we are heading, let's establish where we are. Until now, we (security metricians, including the present authors) could exhort people to "Just measure *something* for heaven's sakes!" It's safe to say that such measurement has largely begun. Therefore, we have the better, if harder, problem of the meta-analysis ("research about research") of many observations, always remembering that the purpose of security metrics is decision support.

Learning from All of Us

To understand how we are at processing our observations, we turn to published industry reports. It's clear that there are a lot more of them than even two years ago. Not all reports are equal; parties have various motivations to publish, which creates divergent interpretations of what represents research worth communicating.

We suspect that most data included in industry reports are derived from convenience samples—data gathered because it is available to the researcher, not necessarily data that is representative enough to be generalizable. Not to make this a statistics tutorial, but for generalizability you need to understand (and account for) your sampling fraction, or you need to randomize your collection process. It is not that this or that industry report has a bias—all data has bias; the question is whether you can correct for that bias. A single vendor's data supply will be drawn from that vendor's customer base, and that's something to correct for. On the other hand, if you can find three or more vendors producing data of the same general sort, combining them in order to compare them can wash out the vendor-to-customer bias at least insofar as decision support is concerned.

Do not mistake our comments for a reason to dismiss convenience samples; research with a convenience sample is certainly better than "learning" from some mix of social media and headlines. This challenge in data collection is not unique to cybersecurity; performing research on automobile fatalities does not lend itself to selecting random volunteers. Studying the effects of a disease requires a convenience study of patients with the disease. It's too

Exploratory before Explanatory

Exploratory research is all about hypothesis generation, not hypothesis testing. It is all about recognizing what are the unknowns within an environment. When that environment is complex or relatively unstudied, exploratory analysis tells you where to put the real effort. Exploratory research identifies the hypotheses for explanatory research to resolve. Exploratory research does not end with "Eureka!" It ends with "If this is where I am, then which way do I go?"

early to call it, but we think it infeasible to conduct randomized clinical trials, cohort studies, and case-control research, but the time is right for such ideas to enter the cybersecurity field (and for some of you to prove us wrong).

If we are going to struggle in the design of our research and data collection, we may be doomed never to produce a single study without flaws. Although that does not preclude learning, it means that we will have to accept and even embrace the variety of conclusions each study will bring while reserving the big lessons to be drawn after the appropriate corrections for the biases in each study are made and those results aggregated. This method of learning requires the active participation of researchers who must not only understand the sampling fraction that underlies their data but also must transparently communicate it and the methodology of their research.

Learning from Each Other

Industry reports are generally data aggregated by automated means across the whole of the vendor's installed base. The variety found in these aggregation projects is intriguing, because much of the data now being harvested is in a style that we call "voluntary surveillance," such as when all the clients of Company X beacon home any potential malware that they see so that the probability of detection is heightened and the latency of countermeasures is reduced for everyone. Of course, once the client (that's you) says "Keep an eye on me," you don't have much to say about just how closely that eyeball is looking unless you actually read the whole outbound data stream yourself.

What can you learn from industry reports? Principally two things: "What is the trendline?" and "Am I different?" A measurement method can be noisy and can even contain a consistent bias without causing the trendline it traces out to yield misleading decision support. As long as the measurement error is reasonably constant, the trendline is fine. Verizon's Data Breach Investigations Report (DBIR) [2] is not based on a random sample of the world's computing plant, but that only affects the generalizability of its measured variables, not the trendline those variables trace. By contrast, public estimates of the worldwide cost of cybercrime are almost surely affected by what it takes to get newspaper headline writers to look at you. Producers of cybercrime estimates certainly claim to be based on data, but their bias and value in meta-analysis efforts must be questioned.

That trendlines are useful decision support reminds us that an ordinal scale is generally good enough for decision support. Sure, real number scales ("What do you weigh?") are good to have, but ordinal scales ("Have you lost weight?") are good enough for decision making ("Did our awareness training hold down the number of detectable cybersecurity mistakes this year?").

Whether you are different from everybody else matters only insofar as whether that difference (1) can be demonstrated with measured data and (2) has impact on the decisions that you must make. Suppose we had the full perimeter firewall logs from the ten biggest members of the Defense Industrial Base. Each one is drawn under a different sampling regime, but if they all show the same sorts of probes from the same sorts of places, then the opponent is an opportunistic opponent, which has planning implications. If, however, they all show the same trendline except for yours, then as a matter of decision support your next step is to acquire data that helps you explain what makes you special—and whether there is anything to be done about that.

Standing on the Shoulders of Giants

Medicine has a lot to teach us about combining studies done by unrelated researchers, which is a good thing, because we don't have a decade to burn reinventing those skills. The challenges facing such meta-analysis are finding multiple research efforts

1. with comparable measurements;
2. researching the same time period (environment may change rapidly);
3. publishing relevant characteristics like the sample size under observation, the data collection, and the categorization scheme.

Without the combination of all three, comparison and meta-analysis (and consequently our ability to learn) becomes significantly more difficult. To illustrate, we collected 48 industry reports; 19 of them contained a reference to "android," and five of those 19 estimated the amount of android malware to be:

- 405,140 android malware through 2012 (257,443 with a strict definition of "malware") [3]
- 276,259 total mobile malware through Q1 2013 [4]
- 50,926 total mobile malware through Q1 2013 [5]
- 350,000 total number of android malware though 2012 [6]
- over 200,000 malware for android through 2012 [7]

That's a broad range of contrast. But do not mistake the range and contrast for an indication of errors or mistakes—their studies are exploring data that they have access to and are an example of the variety of conclusions we should expect. That exploration is exactly what should be happening in this field at this point in time. What we (the security metrics people) must now do is learn how to do meta-analysis in our domain, and if we are producers, learn how to produce research consumable by other security metrics people. We have to learn how to deal with our industry's version of publication bias, learn how adroitly to discount agenda-driven "results," and learn which indicators enable us to infer study quality.

This task will not be easy, but it is timely. It is time for a cybersecurity data science. We call on those of you who can do exploratory analysis of data to do so and to publish in styles such that the tools of meta-analysis can be used to further our understanding across the entire cybersecurity field.

Thanks in advance.

Resources

[1] Steven Johnson, "The Genius of the Tinkerer," *Wall Street Journal*, September 25, 2010: tinyurl.com/ka9wotx.

[2] Verizon's Data Breach Investigations Report: http://www.verizonenterprise.com/DBIR/.

[3] F-Secure Threat Report H1 2013: retrieved from tinyurl.com/ob6t3b4.

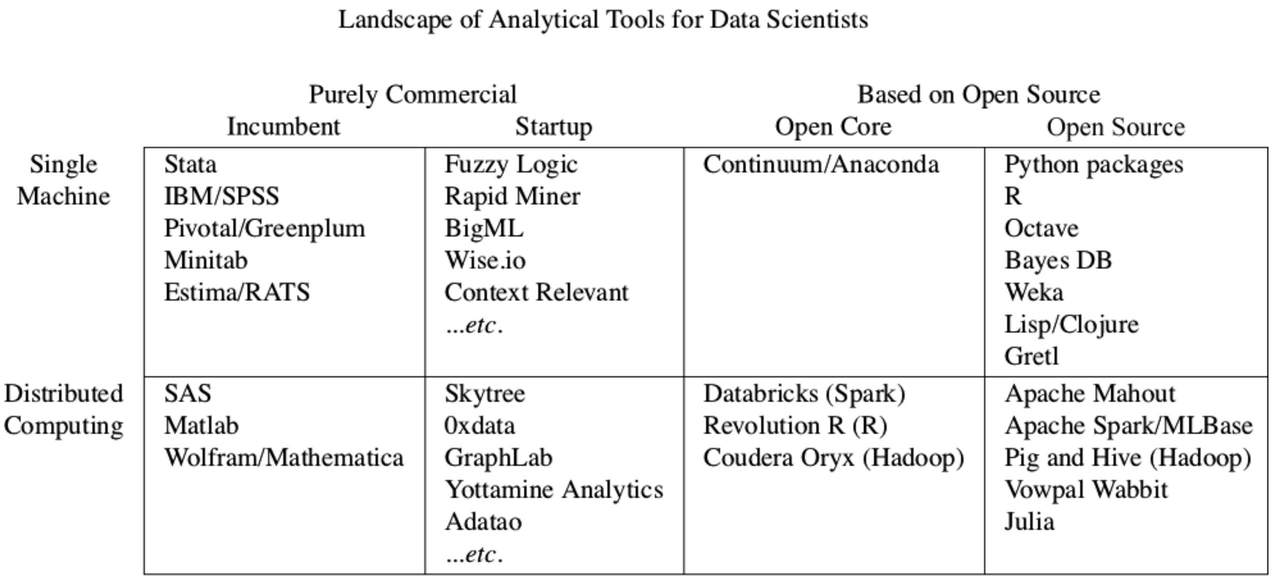
[4] Juniper Networks Third Annual Mobile Threats Report, March 2012 through March 2013: retrieved from tinyurl.com/nf654ah.

[5] McAfee Threats Report: Q1 2013: retrieved from tinyurl.com/ovtrbmg.

[6] Trend Micro: TrendLabs Q1 2013 Security roundup: retrieved from tinyurl.com/lfs2uvt.

[7] Trustwave 2013 Global Security Report: retrieved from tinyurl.com/ljt73qj.

For more resources, please visit: http://dds.ec/rsrc.



Landscape of Analytical Tools for Data Scientists

	Purely Commercial		Based on Open Source	
	Incumbent	Startup	Open Core	Open Source
Single Machine	Stata IBM/SPSS Pivotal/Greenplum Minitab Estima/RATS	Fuzzy Logic Rapid Miner BigML Wise.io Context Relevant ...*etc*.	Continuum/Anaconda	Python packages R Octave Bayes DB Weka Lisp/Clojure Gretl
Distributed Computing	SAS Matlab Wolfram/Mathematica	Skytree 0xdata GraphLab Yottamine Analytics Adatao ...*etc*.	Databricks (Spark) Revolution R (R) Coudera Oryx (Hadoop)	Apache Mahout Apache Spark/MLBase Pig and Hive (Hadoop) Vowpal Wabbit Julia

Addendum: Data science tools as of this date