Report From the Internet2 Data Driven Collaborative Security Workshop for High Performance Networks

Joe St Sauver, Ph.D.
joe@internet2.edu or joe@uoregon.edu
Security Programs Manager, Internet2

Internet2 Fall Member Meeting
San Antonio Texas

http://www.uoregon.edu/~joe/ddcsw-fmm

A Note On Format; Disclaimers

- Yes, this is another one of those oddly formatted "Joe talks." For those who haven't seen one of my talks before, I make them verbose so they;ll be readable after the fact for those who couldn't be here today, as well as for search engines, readers for whom english is a second language, the hearing impaired, etc. Please don't let my odd format shake you up. :-)
- This is also a good time for me to remind folks that all the opinions expressed in this talk represent solely my own perspective, and are not necessarily the opinion of Internet2, the University of Oregon, the Department of Justice (which provided funded for this workshop), the University of Maryland Baltimore County (where the workshop was held), or the meeting attendees themselves

Motivation

- Let's begin by talking a little about the framework for the workshop, including its motivation.

- Today's systems and networks are subject to continual cyber attacks including, *inter alia:*
  -- vulnerability scans and intrusion attempts;
  -- spam, phishing and other unwanted email;
  -- attacks via viruses, trojan horses, worms, rootkits, spyware and other malware;
  -- distributed denial of service attacks; and
  -- attacks on critical protocols such as DNS, BGP and even IP itself.
- Successfully combating those attacks (and other cyber threats) in a <u>scientific way</u> requires <u>hard data</u>.

Data

- Data about system and network attacks may come from a variety of sources, including:

  -- honeypots and dark space telescopes;
  -- deep packet inspection appliances;
  -- netflow/sflow data collectors;
  -- intrusion detection systems;
  -- passive DNS monitoring;
  -- BGP route monitoring systems;
  -- system logs and SNMP data; or even
  -- abuse complaints and other "human intelligence" sources, and
  -- our security and network colleagues.

Analysis

- Once we have data available, we can then analyze and better understand the phenomena we're experiencing. For instance, with the right data we may be able to:

  -- identify botnet command and control hosts;
  -- understand who's actually behind the spam that's flooding our users' accounts;
  -- use one bad domain to find other, related, equally bad domains;
  -- determine who's injecting more specific routes and hijacking our network prefixes;
  -- make decisions about problematic network ranges, including the potential consequences of filtering traffic to/from those problematic ranges.

Action

- Analysis and understanding ultimately enables action:

  -- firewall administrators can filter attack traffic;
  -- block list operators can list problematic IPs or domains;
  -- law enforcement can initiate investigations;
  -- private parties may commence litigation
  -- ISPs can terminate problematic customers for cause;
  -- or the community can even develop new protocols to address pressing concerns.

- But none of us can collect all the data that we'd like to have or that we need to have. We need to collaborate with each other by sharing data and other resources.

Collaboration

- Collaboration can be hard: data availability is often a matter of "feast or famine" -- we're either trying to "drink from the fire hose" without drowning, or we can find ourselves in a position where getting access to any data, or at least the right data, can be quite difficult.
- Data management can also be daunting -- storing, searching, and effectively using terabytes of data is a non-trivial undertaking.
- Simply deciding on a format to use to store or share data can sometimes be more of a problem than one might think: should we use IETF-standardized formats? What then if a major provider unilaterally decides to use their own proprietary format, instead?

That Was Our Framework

- That brief backgrounder should give you an idea of what the workshop was about, and the fundamental challenges we wanted to address:
  -- How can we better work together to share data and make a difference when dealing with operational cyber security issues?
  -- What are folks currently doing? What works well? What doesn't work well?
- It was our hope that attendees would:
  -- gain valuable new insights from the workshop,
  -- make useful professional contacts, and
  -- contribute to recommendations meant to facilitate future data-driven collaborative security initiatives.

Attendee Composition

- In order to encourage "cross pollination" among the various security communities, we intentionally and carefully invited attendees so we'd end up with about:
  -- 1/3rd folks from higher education IT
  -- 1/3rd folks from the private sector (security companies, not-for-profit entities, private security researchers, etc.), and
  -- 1/3rd folks from law enforcement and/or the government
- We also explicitly wanted a mix of both the "usual suspects" plus some less-well-known new faces
- We ended up with 55-60 folks, just the size group we were shooting for. An attendee roster is available at http://security.internet2.edu/ddcsw/

Workshop Site

- Jack Suess kindly offering to let us use the excellent University of Maryland Baltimore County (UMBC) Tech Center for our meeting.
- I think most folks know Jack, but just in case we have some new people in the audience, Jack is the CIO at UMBC, and is an active member of the Educause/ Internet2 Security Task Force (STF) Leadership Team, as well as serving as chair of the Internet2 Applications, Middleware and Services (AMSAC) Advisory Council.
- Jack and the entire UMBC crew were great to work with, and the UMBC Tech Center was a wonderful venue for this event.
- Thank you very much Jack and UMBC!

Workshop Format

- The workshop, which ran for a day and a half, had a mixture of formal presentations, panels and discussion sessions, plus opportunities for private discussions during breaks and meals.
- ***Unlike many security-related workshops, we wanted to make sure that all presentations from this workshop could be publicly shared.*** Thus presenters were explicitly asked to build their slide decks for presentation and dissemination to a public cyber security audience.
- Presenters were also explicitly asked to NOT include any proprietary, FOUO ("for official use only") information or classified information, nor any information which might jeopardize ongoing investigations, prosecutions or sources and methods.

The Presentations and Breakout Sessions

- We had a dozen twenty minute-long presentations spread over the two day workshop (a time format which we frankly stole from the Internet2 Joint Techs meetings). Those presentations are now available in PDF format from http://security.internet2.edu/ddcsw
- We also wanted to make sure that we had a chance to hear the insights and perspectives of all attendees, so we also had two one and a half hour breakout sessions. Attendees could select one of three topics for each of the two breakout sections
- We also had a number of brief lightning talks at the end, another program element that should look familiar to Joint Techs or NANOG attendees.

Our Two Sets of 3 Breakout Session Topics

A) What Data Do We Still Need to Get?

B) What Are the Barriers to Collaboration Against Cyberthreats? How Can We Break Those Barriers Down?

C) Collaborative Data Driven Security in an International/Global Networking Environment

D) What Data Analysis and Data Manipulation Tools Are Missing?

E) What Obstacles Delay or Inhibit Action Against Cyberthreats and How Can We Overcome Those Obstacles?

F) Keeping High Performance Networks Secure -- While Insuring That They Also Remain High Performance Networks!

Detailed session descriptions are on the workshop website

Thanks To All of Our DDCSW Presenters, Panelists and Discussion Group Leaders

- Brian Allen, WUSTL
- Jeff Chan, SURBL
- Richard Cox, Spamhaus
- Andre' Di Mino, Shadowserver
- Brandon Enright, UCSD
- Andrew Fried, ISC
- Tom Grasso, FBI
- April Lorenzen, Server Authority
- Michael O'Rierdan, MAAWG
- Doug Pearson, REN-ISAC
- John Praed, Internet Law Group
- David A. J. Ripley, Indiana U
- Bill Stearns, SURBL
- Henry Stern, Cisco
- Mike Van Norman, UCLA
- Steve Wernikoff, FTC
- Wes Young, U Buffalo
- Sean Zadig, NASA OIG

Aand Thanks To Our Program Committee!

- Brian Allen, WUSTL
- Renee Frost, Internet2
- Terry Gray, University of Washington
- Minaxi Gupta, Indiana University
- Ken Klingenstein, Internet2
- Chris Misra, University of Massachusetts
- Jose Nazario, Arbor Networks
- Michael O'Rierdan, Comcast and MAAWG
- Doug Pearson, REN-ISAC
- Mark Poepping, Carnegie Mellon University
- Henry Stern, Cisco
- Joe St Sauver, Internet2 and U. of Oregon (chair)
- Michael Van Norman, UCLA
- Paul Vixie, ISC

Some Bits and Pieces From the Workshop

DATA: Actionable Data vs. Research Data

- Actionable operational security data needs to be:
  -- **timely** (even last week's data is way too old)
  -- **comprehensive**, not just sampled (we want to know about **all** our compromised hosts (so that we can track them down and get them fixed)
  -- **sufficiently specific** to allow the site to identify the systems/users which have been reported, and
  -- because of that specificity, the sharing of actionable data will usually be **limited to just data about one's own site** or sites (e.g., to stuff you can actually fix)
- Research data, however, is often:
  -- more often "**representative data**" (potentially sampled)
  -- at least partially **anonymized**
  -- global in scope (and not limited to a particular site)
- *We need to make sure we collect the right sort of data*

Data Can Require Real Time "Expansion"

- For example, consider a URL which has been found to lead to malware.
- That URL may chain through multiple additional intermediate sites, but you can only follow that chain of sites while that chain is live.
- Similarly, you need to resolve the URL to see what IP it is using before the domain gets taken down, etc.
- Obviously, if you plan to disassemble or reverse engineer malware, you need to retrieve a copy of the malware while it is still available to be downloaded
- *There's time urgency to these processes, and thus a need to automate the real time "expansion" of data. If you don't get it while its there, you may not be able to get it after the fact.*

Automatic Data Expansion Has Limits

- If you automatically and immediately follow all malicious URLs, you may potentially allow the bad guys to identify your analysis efforts (they may do this through things such as use of uniquely tagged domain names, or through carefully timed and monitored probes)
- Extensive use of automation may also result in analysts having less of a "feel" or less of a "holistic sense" for data that they no longer manually review
- Automation can also be vulnerable to being intentionally flooded or disrupted by being fed intentionally misleading data.
- Apparently most folks who are doing automated analysis have a number of manual safety features and/or "secret sauce" approaches to overcome these issues.

Meta Data

- As we accumulate more and more data, **data about the data we collect ("meta data") becomes increasingly important,** particular for data which isn't inherently "self-documenting."
- For example:
  - -- **Where** and **when** was the data collected?
  - -- Was it **sampled**? If so, how and at what rate?
  - -- Was it **summarized**, **anonymized** or otherwise postprocessed after it was collected?
  - -- What **sharing constraints** apply to use of this data?
  - -- Are there **data dependencies** reflected in this data?

- *We need to be better about creating and using metadata*

Longitudinal Data

- We are also increasingly coming to recognize that there is value to longitudinal or historical data.
- For example, consider fast flux domain names. Fast flux domain names allow miscreants to host web sites on a pool of compromised broadband hosts, replacing old hosts with new ones as the old ones get turned off or cleaned up. Mechanical data-driven fast flux identification formulas exist, such as the so-called "Mannheim Formula," which tabulate the number of IPs and ASNs seen in conjunction with a fully qualified domain name, taking into account not just the IPs and ASNs used right now, but also any previously observed IPs and ASNs if known.
- *As a community we should be timestamping and archiving more data!*

Commercial Data

- Commercial security companies often collect extremely interesting actionable data. Unfortunately, however, often that data is treated as "proprietary" and a corporate asset -- something to be "monetized" rather than something to be freely shared in the interest of improving our collective security
- In general, Internet security "do-gooders" (to use a term that emerged during the meeting), usually do not have the money to buy that commercial security data.
- Commercially collected data which doesn't get sold is a wasting asset; *is there some way the "do-gooders" could perhaps get access to at least "non-saleable" bits of data?*

Data Taxonomy

- Increasing amounts of data also challenges our ability to keep track of data types and sources -- we need a taxonomy, or organized arrangement of data collections, so an analyst can find relevant data, see how it fits with and differs from other data sources, etc.

- An excellent example of this sort of thing (albeit for Internet tools rather than Internet data) is CAIDA's taxonomy of tools, see http://www.caida.org/tools/taxonomy/

Description of A Specific Data Source: SURBL

- Another useful part of the workshop was a case study of how some specific security data sources work.
- For example, Jeff Chan and Bill Stearns talked about how the SURBL works.
- The SURBL doesn't list spam sources, it lists spamvertised URLs and IP addresses seen in the body of spam messages
- Working through a specific resource in detail was very instructive because it helped to explain the objectives and concerns of that resource's operator, including things like the SURBL's real emphasis on minimizing false positives.

ANALYSIS: Transmission & Storage of Data

- Currently a lot of security-related data is transferred from one site to another by email.
- While email is ubiquitous and great for talking with friends or colleagues, it is suboptimal for transferring large volumes of data in real time (and is even more suboptimal for working with that data in the future)
- Programmatically searching for and finding related items across multiple messages may be surprisingly difficult.
- Or malware samples sent by email may be filtered as, well, malware!
- Structured, standardized, extensible transmission and storage of security data is required. SES, the Security Event System, a project which was also funded by this DOJ grant, is an example of an effort to do just that.

Batch vs. Continual Flow Processing

- We're all familiar with batch and flow processes at home:
  -- batch: we accumulate laundry until we have enough dirty clothes for a load of wash
  -- flow: our air conditioners continually produce cool air, our hot water heaters continually produce hot water
- The traditional data analysis paradigm has been to accumulate discrete batches of network data to analyze, just like a pile of clothes, processing that data only after an entire batch has been received and saved.
- Improved data access as well as other factors are now increasingly resulting in replacement of "batch" analyses with continual flow analyses.
- *We must learn to retool and adapt our analytical process to work with continual data flows*

Case Study: ISC SIE

- Andrew Fried of the Internet Systems Consortium did a great job describing the Security Information Exchange (SIE), a trusted, private framework for information sharing where participants can run real time sensors which upload/inject live data to SIE, and where other participants can monitor, query or download relevant bits of that data.
- Thirteen real time channels are currently available, including a variety of DNS data, netflow data, spam, bad URLs and other data. His talk describes this further, and provides helpful examples of what's available.
- ISC SIE is an excellent example of a resource that is driving flow (rather than batch) processing.

Visualization of Security Data

- As the volume of security-related data increases, it becomes harder to make sense of raw numerical data.

- Visualization has the potential to help us see trends and to identify interesting departures from normal activity.

- Visualization also helps us to "package" security data in ways that may appeal to non-technical audiences

- The trick is making visualizations useful rather than just mesmerizingly "cool" (albeit uninterpretably so!)

ACTION: Commercial Data-Driven Activity

- Henry Stern of Cisco talked about Cisco's new plans to turn the table on botnets using Cisco IPS 7.0 with Global Threat Correlation
- They have 200K+ potential sensors
- IP reputation data is used to adjust risk ratings
- Sensors collect and aggregate event data (500GB/day of sensor data!), coordinated through Cisco Security Intelligence Operations
- They're exploring novel scalable algorithms for attack detection and botnet correlation
- Seeking related grant proposals from the community for the Cisco Foundation

Data-Driven Civil Litigation

- One of the presentations that most-energized the attendees was a talk by Jon Praed, an attorney with the Internet Law Group.
- Jon explained how civil legal processes can be used to collect and preserve cybercrime-related data, leverage government resources, and expose enablers and cyber threats, thereby resulting in strategic solutions to some of the threats we face online.
- He had many interesting observations, including the fact that about a dozen gangs are responsible for over 80% of all cyber misbehavior.
- He also urged "rebalancing" the allocation of security resources between defensive measures (e.g., more/better captchas) and offensive measures (e.g., civil lawsuits)

Government Enforcement

- Steve Wernikoff of the FTC was good enough to go over a spam-related case which had recently taken place.
- It was very instructive to see the sorts of data the FTC and others acquired as part of that specific investigation, and the sort of further leads that were found as a result. For example, following tracking ids and monitoring affiliate cookies, etc., are examples of some new key data.
- At the same time, it may be noteworthy that at least some enforcement entities may be moving away from cases built on "technical intelligence" to cases built using other mechanisms (such as "follow the money" or human intelligence approaches).

ISP And Their Responsibility to Act

- Some ISPs, registrars, and other Internet actors are very responsive when provided with trustworthy information about abuse.
- Other parties, however, may refuse to act unless compelled to do so by force of law (e.g., law enforcement appears with a court order and seizes systems, or LE forces the production of records via compulsory process).
- Fundamental problem: "abuse handling is not a profit center" and "there's no law that says we have to deal with this, we're just a pipe."
- The process of obtaining cooperation from ISPs and other key Internet parties remains a critically challenging issue, particularly when even marginal customers may be seen as a critical source of desperately needed income.

Making Volunteer Efforts Sustainable

- Many key system and network security efforts are run by volunteers as a public service.
- While volunteer efforts may start out fueled by caffeine, noble intentions and personal credit cards, its hard to sustain that momentum.
- There are real costs to one's personal finance, one's careers and to one's relationships with families and friends as a result of doing side security projects in addition to one's $DAYJOB.
- *Sustsainable long-term business models (or at least long-term funding sources!) are required or we risk losing key resources we all depend on.*

COLLABORATION: Where Should Reports Go?

- It can be surprisingly difficult at times to figure out who to talk to/where reports should go, and to keep reporting contacts current. Some obvious possibilities include:
- CERTs and ISACs
- ASN Owners, CIDR Owners, Domain Name Owners
- Law enforcement
- Researchers
- Security groups
- Obviously we don't want to send data to the bad guys (or to people who could care less about getting reports)
- Should we be prioritizing incident reports which will definitely get worked if noticed and reported? ("magic ASNs")

What Should We Be Reporting?

- Spam sources (largely compromised hosts)
- Spamvertised URLs
- IPs seen doing scans, probes and brute force attacks
- DDoS participants (compromised hosts, open recursive resolvers, etc.)
- Botnet C&C hosts
- MD5s and samples of poorly detected malware
- Info gleaned from the network traffic generated when sandboxed malware is executed
- Etc., etc., etc.

Collaboration Case Study: NCFTA

- The National Cyber Forensics and Training Alliance (NCFTA) is a non-profit, non-government entity located in Pittsburgh PA. It…
  -- is a joint effort between law enforcement, industry and academia to fight cyber crime
  -- is a neutral space where subject matter experts can work with law enforcement on focused initiative
  -- NCFTA Information Analysts are students from local universities (Pitt, CMU, WVU, Robert Morris, Duquesne, Slippery Rock) in discliplines such as security intelligence studies, information science, business & law
  -- Law enforcement participants include FBI, Postal Inspection and others
- Work on bots, spam, phishing, id theft, child exploitation

Summary

- Thanks to the hard work and participation of a lot of great folks, the DDCSW workshop went very well.
- The data, analysis, action and collaboration framework provided a good basis for considering the issue.
- We didn't have time in today's brief slot to describe all the findings and outcomes of the workshop, however even the selected items mentioned here should be quite encouraging.
- The meeting also catalyzed some emerging collaborative efforts which are already paying off.
- We hope to have DDCSW II this coming spring, either somewhere on the west coast, or in the midwest.
- Thanks for the chance to talk - are there any questions?