Botnet Metrics

FCC CSRIC WG7 (Botnets)
9AM-3PM, 12/12/12, Washington DC

Joe St Sauver, Ph.D. (joe@oregon.uoregon.edu)

http://pages.uoregon.edu/joe/wg7-botnet-metrics/

Disclaimer: Any opinions expressed are solely those of the author and do not necessarily reflect the opinion of any other party.

1. Introduction

Our "Goals for the Day / Report"

- Present a model for metrics that is actually beneficial to the goals of the code and acceptable to participants

- Capture current state of metrics efforts across industry

- Describe the key barriers to effective collection/reporting of metrics at this time

- Offer a way forward (M3AAWG program, RFC, sample metrics, etc.)

Some Potential Botnet Metric Audiences

- **ISPs** may wonder, "Is ABCs for ISPs worthwhile? What's it cost us to participate? What benefits do we accrue from doing so?"
- **The FCC and DHS** may wonder "Is the Code working, or do we need to try some other approach? If it's helping, how much?"
- **The public and public interest organizations (including the media)** may want users to be protected from bot threats, but in a way that's appropriate and privacy-respectful.
- **Vendors** may view metrics as driving new markets for new gear
- **Law enforcement agencies** may eagerly seek botnet metrics to help them to target and optimize their enforcement activities
- **Cybersecurity researchers** might want access to empirical data about botnets for use in their own analyses.
- **Governments overseas** may look at our bot metrics to see if this program is something that they should be doing, too.

Opening the Kimono Too Far

- While there are many audiences that are welcome to industry botnet metrics, there is one explicit non-audience: **botmasters** (and other cyber criminals).
- Botnet metrics, done wrongly, have the ability to potentially help our enemies and undercut our own objectives.
- A simple example of this: giving detailed and accurate information about where and when bot activity was observed may be sufficient for a botmaster to identify (and subsequently avoid!) a researcher's honeypots (or other data collection infrastructure) in the future. If that happens, valuable (sometimes literally irreplaceable) sources and methods may be compromised.
- Another example: if our botnet metrics include "per-bot" cleaning and removal statistics, botmasters could use that feedback to learn what's bots have proven hardest to remove, information that they can then use to "improve" their evil products.

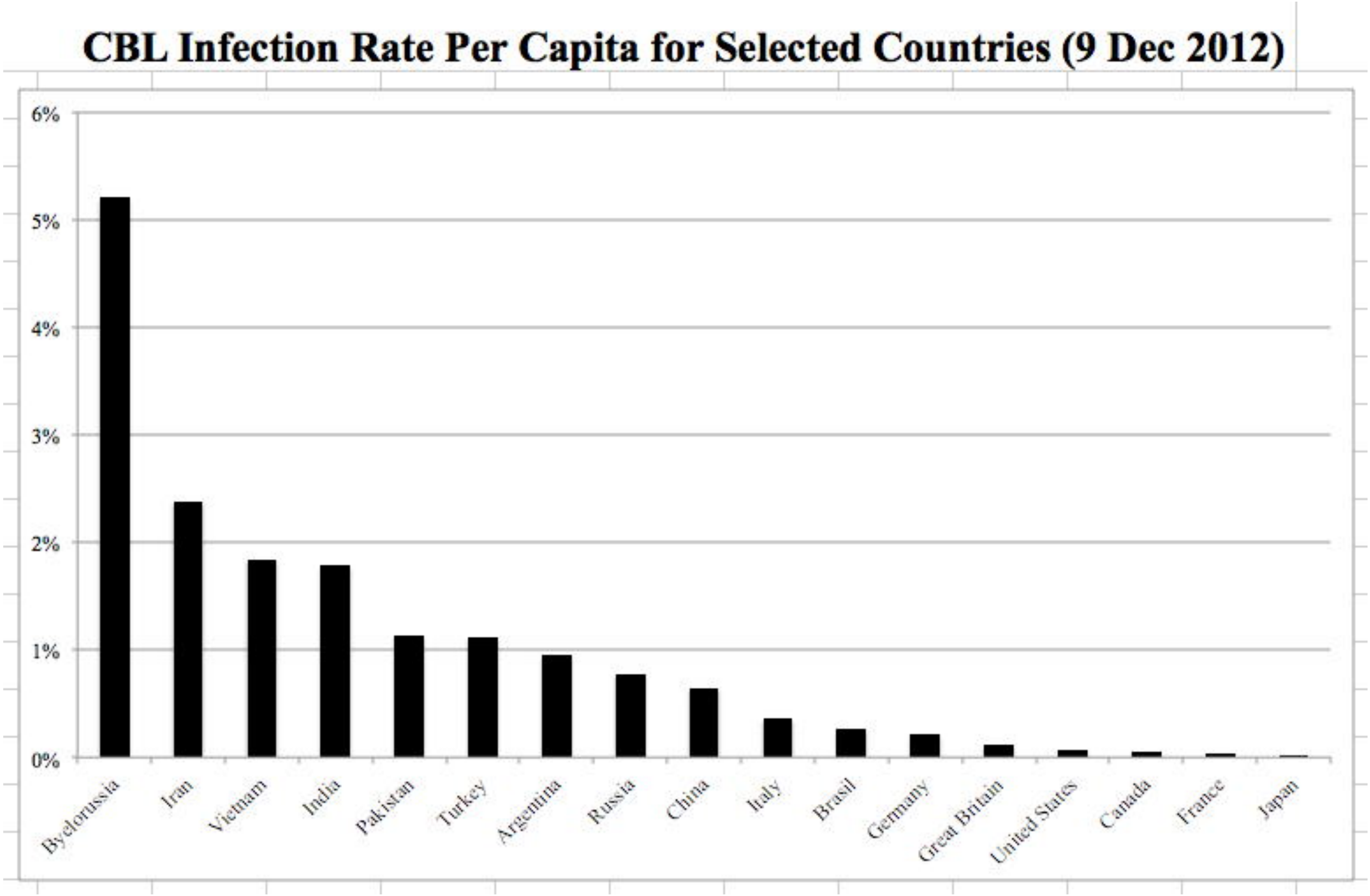
Lots of Questions, (Maybe) Some Answers?

- Today's session will include a lot of questions. Many of those questions have multiple "right" answers. We'd like to understand the full potential set of answers, even if we can't all support one answer in particular.
- We hope that the discussion today will kindle additional thinking, and provide fodder for the group's final report, but we do NOT want you to think that we'll be able to answer every question or address every issue. Some questions may need to be addressed by members of a future CSRIC working group.
- We also want to emphasize that botnets are highly dynamic and guided by some extremely smart opponents, so what works and is helpful today may NOT work or yield meaningful data tomorrow.
- Some topics may seem really picky or obscure, but by the time we're done, we hope that you'll see why the questions we're raising today are all potentially important ones.

Recognize That We May Already Be Succeeding

- There is a temptation to always approach cyber security topics with a negative eye, looking for evidence that "we've got a problem" that needs fixing.

- I don't want to be accused of sounding Pollyannaish, but I d hope we will be open to the possibility that, our collective "Eeyore-like" predispositions notwithstanding, we may actually be already doing pretty well when it comes to fighting the war on bots.



Source: http://cbl.abuseat.org/countrypercapita.html as of Sunday, Dec 9th, 2012

(Maybe) One Reason Why: Microsoft's MSRT

- Every month when your computer running Microsoft Windows automatically checks for patches, it also silently runs the Malicious Software Removal Tool.

- The MSRT doesn't remove ALL malware, but MSRT does remove the "worst of the worst" **including most bot malware families**. For a complete list of what bot families get removed by MSRT, see: www.microsoft.com/security/pc-security/malware-families.aspx

- **Because of MSRT, if a computer is set to automatically update, most common bot malware will automatically and silently be removed from the computer each month.**

- **Modern versions of Windows are set to auto update by default.**

Maybe Another Reason Why -- LE Bot Takedowns

- Another reason why we may be doing "better than expected" against bots is that law enforcement has clearly made bots a priority for enforcement actions. Just a few bot related examples:
  -- **Operation Bot Roast**, June 2007
  http://www.fbi.gov/news/stories/2007/june/botnet_061307
  -- **Operation Bot Roast II**, November 2007
  http://www.fbi.gov/news/stories/2007/november/botnet_112907
  -- **Mariposa Botnet Takedown**, July 2010, December 2012
  http://www.fbi.gov/news/pressrel/press-releases/fbi-slovenian-and-spanish-police-arrest-mariposa-botnet-creator-operators and
  http://www.fbi.gov/news/pressrel/press-releases/fbi-international-law-enforcement-disrupt-international-organized-cyber-crime-ring-related-to-butterfly-botnet
  -- **Coreflood Takedown**, April 2011
  http://www.fbi.gov/news/stories/2011/april/botnet_041411

All That Said, We Still Have Many Question

- So let's dig right in, beginning with a brief but important return to the question of "What's a bot?"

2. Defining "Bots" (And Why It Matters How We Define "Bots")

"Bots" As (Abstractly) Defined by WG7 So Far

- A malicious (or potentially malicious) "bot" (derived from the word "robot" [...]) refers to a program that is installed on a system in order to enable that system to automatically (or semi-automatically) perform a task or set of tasks typically under the command and control of a remote administrator (often referred to as a "bot master" or "bot herder.") Computer systems and other end-user devices that have been “botted” are also often known as "zombies".
- Malicious bots are normally installed surreptitiously, without the user's consent, or without the user's full understanding of what the user's system might do once the bot has been installed. Bots are often used to send unwanted electronic email ("spam"), to reconnoiter or attack other systems, to eavesdrop upon network traffic, or to host illegal content such as pirated software, child exploitation materials, etc.

http://www.maawg.org/system/files/20120322%20WG7%20Final%20Report%20for%20CSRIC%20III_5.pdf

Why A Precise Definition of "Bot" Matters

- If "bot" gets defined broadly to include virtually <u>any</u> sort of malware, the number of infected hosts we'd identify will be <u>far larger</u> than the number of botted hosts we'd find if "bot" was defined more narrowly.

- ***Many, many user systems are infected with, or WILL get infected with, adware or spyware, and while adware/spyware is malware, and bad, that adware/spyware is NOT bot malware. There are also plenty of other sorts of non-bot malware.***

- ***WE need/want to SOLELY focus on BOT malware.***

- For metrics, the distinction between malware in general and bots in particular can matter dramatically. For example...

Metrics and Sample Media Coverage of WG7

- "Seven months after a coalition of government and industry organizations announced a broad set of voluntary guidelines [e.g., the ABCs for ISPs] to help Internet service providers clean their broadband networks of malware, the effort has yet to produce measureable results. [...]

  "So far there is no evidence that the effort is producing meaningful results. In the third quarter of 2012, for example, **6.5 percent of North American households had *malicious software* on at least one computer**, according to the data from Kindsight's latest report. The rate is a slight increase from the **6 percent of households that showed signs of *malware infection*** in the first quarter of the year."

  "Anti-Botnet Efforts Still Nascent, But Groups Hopeful"
  http://www.darkreading.com/security-monitoring/167901086/security/news/240143005/anti-botnet-efforts-still-nascent-but-groups-hopeful.html [emphasis added]

Some Problems With The Preceding Analysis

1. **All types of malware got treated as if they represented "bots"**

2. **By looking at population wide (total) infection rates**, infection rates of code-subscribing ISPs ended up **comingled** with infection rates of non-subscribing ISPs. Given that comingling, an uptick in bots in non-subscribing ISP users might have offset any improvement in the number of bots seen in subscribing ISPs' user populations.

3. **That study looked at the infection rate for "North American" households**, presumably including households from **Canada** and **Mexico** as well as the United States, even though **the code only targets U.S. ISP and their customers.** That means that even if U.S. botnet numbers improved, domestic improvements may have been swamped and lost in a worsening of Canadian/Mexican numbers.

4. That study looked at infection rates for households, rather than computers. There might be a half a dozen computers in a household, but if even one is infected, the entire household is flagged as bad.

The Importance of Denominators

- If you're looking at a rate associated with an undesirable phenomena, there are two ways to worsen the value you obtain:

  -- drive the top value (the numerator) up
  -- concentrate the top value across a smaller bottom value (the denominator)

- If we are counting infected entities, and call any household where even one computer is infected an "infected household," that tends to increase our chances of finding "infected households."

- A less sensational result might come from counting infected computers, and reporting those infected computers as a fraction of all computers (a more natural denominator)

Using An Appropriate Numerator

- We also need to give a hard look at what gets tossed into the numerator.

- For example, we're the BOTNET working group, not the MALWARE working group, so it isn't fair to artificially inflate the numerator of our botnet infection rate by counting all kinds of malware, rather than just BOT malware.

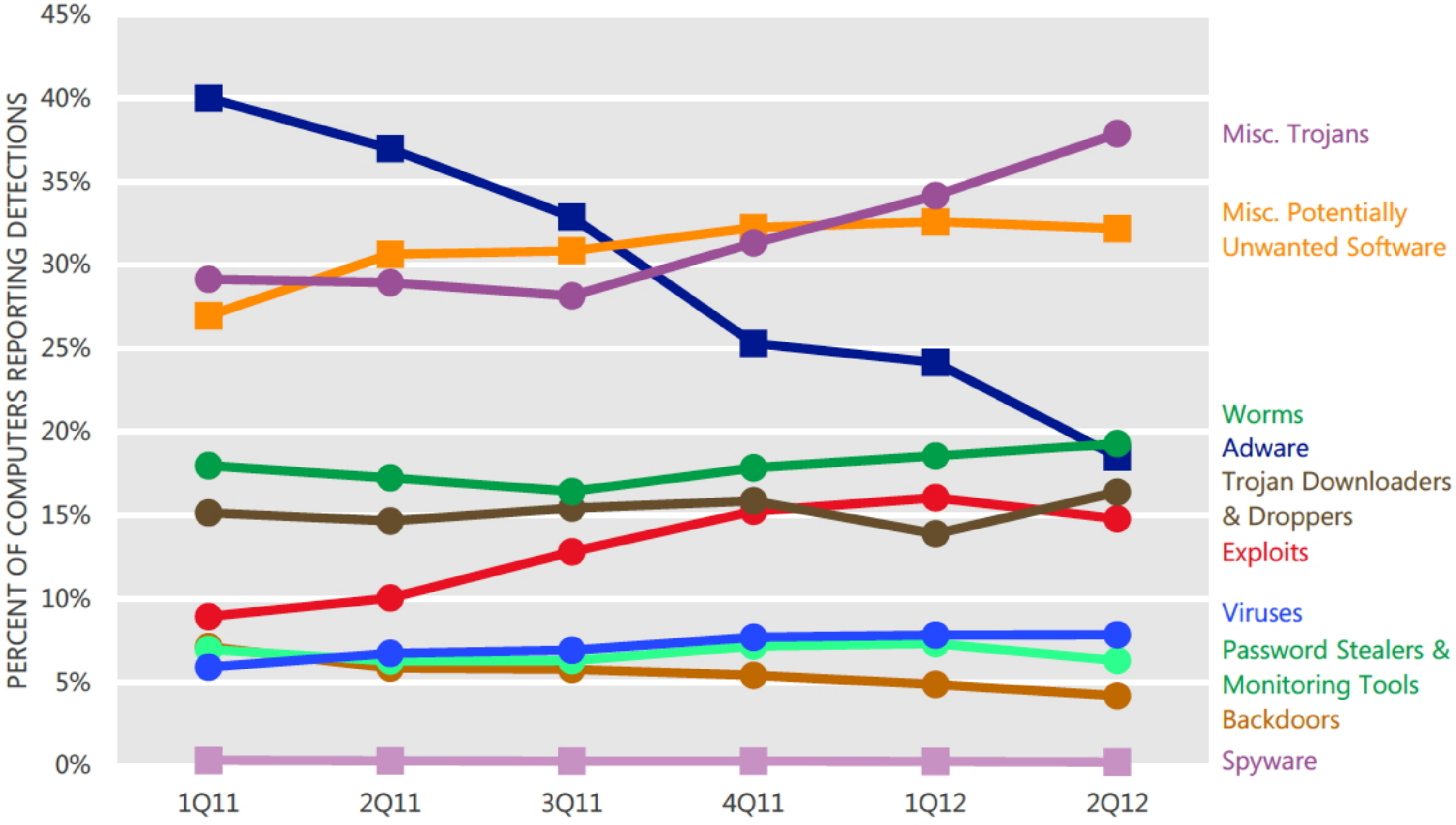
- What do we believe differentiates a bot from non-bot malware?

Malware In General vs. Bots In Particular

Characteristic	Malware	Bot
1. The malicious software was installed without the user's informed consent (installation may have been done involuntarily, or the user may have been deceived into allowing installation of the malicious software)	YES	YES
2. The malicious software performs some sort of unacceptable or illegal activity (interferes with system use, compromises private information, sends spam, scans/attacks other systems, hosts illegal content, etc.)	YES	YES
3. The malicious software does work over the network after installation	POTENTIALLY	ALWAYS
4. The malicious software enables a remote administrator (the "botmaster") to instruct the infected system to do specific work of the botmaster's choosing (that is, the bot can be remotely "steered")	POTENTIALLY	ALWAYS
5. The unauthorized remote administrator is able to control multiple infected nodes as a unified entity, allocating malicious work across a set of worker nodes.	POTENTIALLY	ALWAYS

Lots of Different Sorts of Malware...
How Much of This Is Actually "Bots"?

Figure 28. Detections by threat category, 1Q11–2Q12, by percentage of all computers reporting detections



Round markers indicate malware categories; square markers indicate potentially unwanted software categories.

Source: Microsoft Security Intelligence Report, Volume 13

A Bot Malware Registry?

- To explicitly clarify what we mean when we refer to "bots," the industry should create a voluntary bot malware registry, listing malware that has been found to be "bot" malware.

- Anti-malware vendors, when analyzing and cataloging malware, could then potentially voluntarily add an "is this malware a bot?" attribute to their malware catalog entries, and potentially employ that attribute as part of their periodic malware reporting.

- For example, in addition to any other statistics an anti-malware vendor might share, an anti-malware vendor might also report on:
  -- number of new families of bots discovered that quarter
  -- percent of systems seen infected with the most significant bots
  -- total number of hosts detected as infected with at least one bot

A Few Examples of Current / Historical Malware Families That Could Properly Be Called "Bots"

- Agobot/Phatbot
- Bagle
- Coreflood
- Cutwail/Pushdo/Pandex
- Dirt Jumper/Russkill
- Donbot
- Festi/Spamnost
- Grum/Tedroo
- Kelihos/Hlux
- Koobface
- Kraken/Bobax
- Lethic
- Maazben
- Mariposa
- Mega-D/Ozdok
- Ogee
- Rustock
- SDBot
- Srizbi
- Storm
- Waledac
- Zeus/Zbot

- http://en.wikipedia.org/wiki/Agobot
- http://en.wikipedia.org/wiki/Bagle_%28computer_worm%29
- http://en.wikipedia.org/wiki/Coreflood
- http://en.wikipedia.org/wiki/Cutwail
- http://ddos.arbornetworks.com/2011/08/dirt-jumper-caught/
- http://en.wikipedia.org/wiki/Donbot_botnet
- http://blog.eset.com/wp-content/media_files/king-of-spam-festi-botnet-analysis.pdf
- http://www.theverge.com/2012/8/5/3220834/grum-spam-botnet-attack-fireeye-atif-mushtaq
- http://en.wikipedia.org/wiki/Kelihos_botnet
- http://en.wikipedia.org/wiki/Koobface
- http://en.wikipedia.org/wiki/Kraken_botnet
- http://en.wikipedia.org/wiki/Lethic_botnet
- http://labs.m86security.com/2009/10/maazben-best-of-both-worlds/
- http://en.wikipedia.org/wiki/Mariposa_botnet
- http://en.wikipedia.org/wiki/Mega-D_botnet
- http://riskman.typepad.com/perilocity/2012/03/what-other-asns-were-affected-by-botnet-ogee-in-february-2012.html
- http://en.wikipedia.org/wiki/Rustock
- http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp_sdbot_irc_botnet_continues_to_make_waves_pub.pdf
- http://en.wikipedia.org/wiki/Srizbi_botnet
- http://en.wikipedia.org/wiki/Storm_botnet
- http://en.wikipedia.org/wiki/Waledac_botnet
- http://en.wikipedia.org/wiki/Zeus_botnet

Are There Other Major Bot Families That We Should Include On (Or Omit From) That List?

ADD:

- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________

DROP:

- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________
- ______________________________

"If It Acts Like a Bot..."

- In some cases (for example, in the case of an ISP that does not have administrative access to a customer's system), **if a system exhibits empirically observable bot-like behaviors** (such as checking in with a botnet command and control host, or spewing spam, or contributing to a DDoS attack), **even if a particular bot cannot be identified, the system should still get tagged as being botted.**

- This would also be true in cases where access to the machine is possible, but the malware on the machine that's exhibiting bot-like behaviors is so new that antivirus companies have not yet had time to identify and investigate it.

- **Suggestion: if a system acts like a bot, even if it cannot be identified as infected by bot malware, tag it as botted.**

Once We Agree About What Is/Isn't A Bot...

- After we agree on what is and isn't a bot, we're part of the way to being able to ask meaningful/measurable questions about them....

- However, we also need to decide one other critical issue, and that's the "**unit of analysis.**"

3. Picking A Unit of Analysis (Also Known As The "So What Are We Counting?" Question)

Are We Going to Measure All Kinds of Botted Devices, Or Just Botted Desktops and Laptops?

- In the old days, everyone used a desktop or laptop computer, and that's what got botted.

- These days, people also have tablets, smart phones, and all sorts of other devices. Those sort of devices can now get botted, too.

- Do we count all those sort of devices, or just desktops and laptops? Since we are seeing smart phones getting botted just as desktops and laptops are, I'd argue that we should include ALL types of devices.

- We also need to recognize that servers are also getting botted these days.

Are We Only Going to Count Botted Devices That Are Online?

- Imagine that we're an enterprise that actively scans its systems with Nessus or a similar scanning tool in an effort to identify systems that appear to be botted.
- Obviously, if a system isn't connected to the network, or isn't powered up when our scan happens, a potential infection on that system can't be scanned and found. If we scan and can't reach a potentially botted host, should we schedule it to be periodically retried unless/until it can be?
- That is, are we only interested in counting botted systems that are "live"/"that we can see" on the network?
- Sidebar: from the POV of an ISP, if a botted system has been detected as being botted, and successfully put into a walled garden where it can't hurt others, should it still be counted as "botted"?

When/For How Long Will We Measure?

- If a botted system is only powered on/connected occasionally, when we scan it (or how many times we try to rescan a system) may matter. (Or should scans be ongoing/continual?)
- Similarly, when/how long might we collect bot-related Netflow records? Should we collect Netflow continually? Do ISPs have the *capacity* to do so on a sustained un-sampled basis?
- Some home systems may only be used in the evening. Some work systems may only be used during the work day. Some bots may only be activated during particular periods. Therefore, if you try to count botted hosts during too short a time period, you might miss some of them.
- In other cases, a botted host might show up on one dynamic IP address now, but another address later. If you count over too long a time period, you may count the same botted host more than once.

What I'm Able to Count Depends On Who I Am...

- If I'm an antivirus company or an operating system vendor and I scan/clean systems, I'm going to end up counting botted systems. (One user might have 2 or 3 systems, or one system, multiple users)
- If I'm an ISP, and I detect bots based on malicious network activity associated with a particular IP address, I'm going to count botted IP addresses. (A single "botted IP address" might represent a NAT box in front of half a dozen different devices)
- If I'm a survey research outfit, and I interview people, asking "Have YOU ever had a computer that's been infected with a 'bot'?" those survey researchers are going to end up counting botted users.
- If you were to go "boots on the ground" and actually check all the devices in a number of households to see whether any device is infected, you might end up counting systems, users or households.
- Those various measures will obviously NOT be comparable.

In General, We'd Prefer Data From The "Finest Grained" Unit of Analysis That's We Can Get

- If we collect data correctly, we can always "roll up" data collected at a finer level of analysis into a more aggregate level.
- For example, if we collection infection status data at the level of individual systems, (but also note the IP address or subscriber ID to which those individual systems belong), we can always *post hoc* aggregate those detailed record according to the IP address they share or their subscriber ID.
- Sometimes, unfortunately, we may not be able to collect data at the unit of analysis we might prefer. For example, if an ISP collects their bot measurements by looking at network flow data, traffic from multiple systems behind a customer NAT box may be indivisibly comingled. That ISP would have no option but to try to work with that more highly aggregated data since all systems at that customer's house generate traffic with the same shared IP address.

Note: Not All Data May Roll Up to Households

- While we expect most code participants are broadband ISPs serving residential customers, that may not always be true.
- For example, imagine a broadband ISP that has small and medium business customers as well as residential customers. In the small and medium business environments, multiple IPs may roll up to a single store, and then multiple stores may roll up to a corporation. In that case, how could we "roll up to the household level?" Obviously, we couldn't.
- We could roll up to the *customer* level in this case, but does it always make sense to do so? In particular, note that all customers might not be comparable size units: a "mom and pop" corner bodega with cable model connectivity is a different entity than a regional chain store running its own corporate LAN, for example.
- Aggregation can be tricky, but it *will* routinely be done.

"Subscribers"
(The Code Is An ISP Activity, Right?)

- The MAAWG metrics program focuses on "subscribers" as the unit of analysis, and counts the number of subscribers as of the last day of the month.
- MAAWG's metric program tallies the number of unique subscribers that have been found to be infected one or more times during the month. (What is/isn't an infection isn't explicitly defined, except to say that it should be an "infection" that's serious enough to motivate the ISP to contact the user about the infection)
- As a metric, note that this implicitly incorporates some measurement compromises, e.g., given the definition of this metric, we can't talk about how many infected customer systems may be present, for example.

A "System" Might Not Always Be The Most Fine-Grained Unit of Analysis

- There's a temptation to think of "systems" as individual desktop or laptop or smartphone, and for a consumer broadband ISP, often that will be a reasonable assumption.
- However, another possibility might be that a single "system" might be a **large shared system**, such as high density web hosting site, or a Unix box with thousands of shell accounts. In that case, you'd need to know a specific URL or a specific username to know which of many customers on that large shared host is botted.
- **Virtual machines** (particularly virtual machines with individual network interfaces) pose similar problems. If a system has two dozen virtual hosts on it, is that one "system," or two dozen?
  (I think a virtual machine with its own network interface should be treated as its own system, even if it shares physical hardware)

Bot Malware Infections:
A Still-Finer Unit of Analysis?

- Contrast two different systems:

  System A: infected with one bot.

  System B: infected with seven different bots.

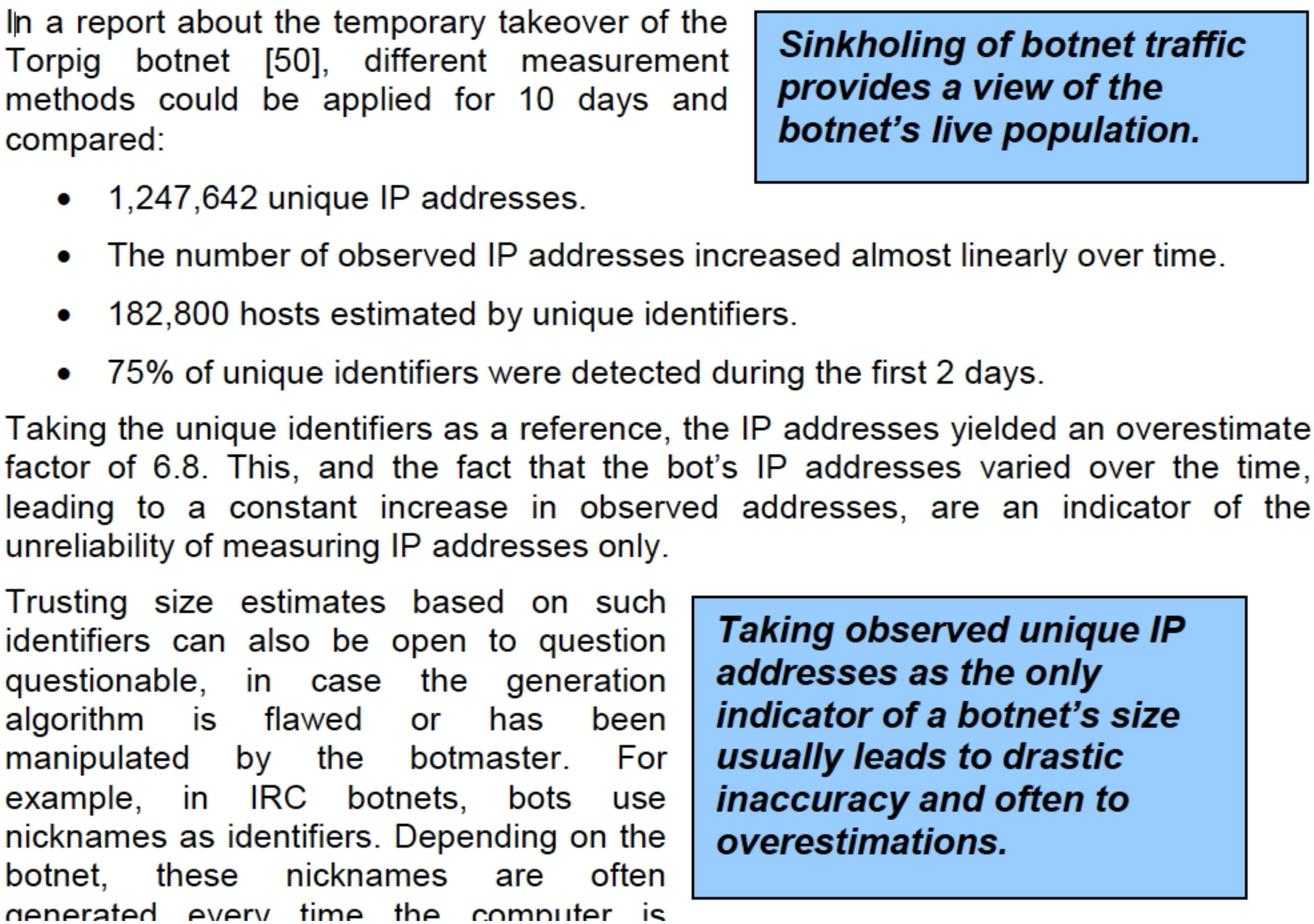
- Should each of those systems be counted as "one" infected system?

- Or should one generate one record (for System A), and generate seven records (one for each bot infection on System B), thereby allowing each individual type of bot infection to be separately tracked?

Ensuring That Fine-Grained Records Can Be Aggregated ("Rolled Up") Appropriately

- As records get collected on a fine grained basis, it will be critical to ensure that those fine grained records can be aggregated ("rolled up") appropriately. In order for this to be possible, related records need to share a common identifier.
- For example, if you want to roll up multiple infections associated with a single system, you might want to have something like the system's hardware Ethernet (MAC) address (or some other unique system identifier) as part of each infection record. Alternatively, if you want to roll up records associated with a given subscriber, you'd need the subscriber identifier as part of each record.
- The challenge: use (and sharing!) of unique identifiers may raise privacy concerns. But, if we don't us unique identifiers, our measurements may be flawed. For example...

Unique Identifiers vs. Unique IP Addresses



In a report about the temporary takeover of the Torpig botnet [50], different measurement methods could be applied for 10 days and compared:

Sinkholing of botnet traffic provides a view of the botnet's live population.

- 1,247,642 unique IP addresses.
- The number of observed IP addresses increased almost linearly over time.
- 182,800 hosts estimated by unique identifiers.
- 75% of unique identifiers were detected during the first 2 days.

Taking the unique identifiers as a reference, the IP addresses yielded an overestimate factor of 6.8. This, and the fact that the bot's IP addresses varied over the time, leading to a constant increase in observed addresses, are an indicator of the unreliability of measuring IP addresses only.

Trusting size estimates based on such identifiers can also be open to question questionable, in case the generation algorithm is flawed or has been manipulated by the botmaster. For example, in IRC botnets, bots use nicknames as identifiers. Depending on the botnet, these nicknames are often generated every time the computer is

Taking observed unique IP addresses as the only indicator of a botnet's size usually leads to drastic inaccuracy and often to overestimations.

Source: "Botnets: Detection, Measurement, Disinfection & Defence," ENISA, page 70.

Measuring Bots On IPv6

- If we measure bots on a per-IP basis, how will we handle bots connecting via IPv6?

- In IPv6, hosts may have an automatically assigned address that leverages a reformatted version of the system's hardware MAC address, making it easily possible to track a botted host using that style address over time and even across multiple ISPs.

- However, IPv6 hosts can also use constantly changing IPv6 "privacy addresses." If a botted host is using IPv6 privacy addressing, how will we track those sort of botted hosts over time?

Some Bot Infections May Not Even Be Knowable

- Botnets are inherently stealthy/strive to hide! Some do this well!
- A new piece of bot malware might not immediately get detected by the antivirus product that a user may be running.
- Moreover, once a piece of malware is detected by antivirus software and is removed, all we can talk about is former infections... that's not as interesting as unmitigated infections.
- Speaking of former infections, a system may get silently botted, but then also get silently cleaned up (e.g., by something like Microsoft's Malicious Software Removal Tool), so only the botmaster and the party automatically cleaning up that bot (e.g., Microsoft) may even know that it was ever there...
- A spam bot may attempt to send spam, but that spam may be blocked by an ISP, so an external researcher may not know that system is infected (even if the ISP may see/log that spam)

Some Apparent "Bot" "Hits" May Not Be <u>Real</u>

- For example, imagine researchers investigating a botnet: it is conceivable that they might attempt to "register" or "check in" fake "bots" of their own creation in an effort to figure out how a botnet operates, sometimes in substantial volume.

- In other cases, imagine a antibot organization that is attempting to proactively interfere with a bot by "poisoning" it with intentionally bogus information about fake botted hosts, hoping that the system will try to rely on the injected fake bots (which might not actually do anything) rather than using real bots.

- Thus, if you're measuring bots by counting the hosts that "check in" rather than the bots that are actually seen "doing bad stuff" you run the risk of overestimating the number of bots that actually exist.

4. Some Substantive Questions We Might Have About Bots

Let's Agree: It Isn't Hard to Come Up With Lots of Interesting Questions About Bots

- If we were just sitting around brainstorming, it isn't hard to come up with a whole bunch of interesting questions relating to botnet metrics...

One Basic Question: "What's The Order of Magnitude of the Bot Problem...?"

- **If botted hosts are rare, we likely don't need to worry about them.** On the other hand, if ISPs are being overrun with botted hosts, we ignore those botted hosts at our peril.

- If we don't (or can't!) at least roughly measure botnets, we won't know if bots are a minor issue or a huge problem, and if we don't know roughly the size of the problem, it will be impossible for industry or others to craft an appropriate response.

- Note: when we talk about "order of magnitude," we're NOT talking about a precise measurements, we're just asking, "Are 10% of all consumer hosts botted? 1% of all hosts? 1/10th of 1% of all hosts?" etc... We should at LEAST be able to do that, right?

There Are Many Order of Magnitude Bot Estimates Already In Circulation... For Example...

- "Household Botnet Infections," Gunter Ollmann, March 26, 2012, www.circleid.com/posts/20120326_household_botnet_infections/

> *Out of the aggregated 125 million subscriber IP addresses that Damballa CSP product monitors from within our ISP customer-base from around the world, the vast majority of those subscriber IP's would be classed as "residential" — so it would be reasonable to say that* ***roughly 1-in-5 households contain botnet infected devices.*** *[...]*
> *Given that the average number of devices within a residential subscriber network is going to be greater than one (let's say "two" for now — until someone has a more accurate number),* ***I believe that it's reasonable to suggest that around 10% of home computers are infected with botnet crimeware.***

If 1-in-5 Households Was Infected in the US...

- There are 81.6 million US households with broadband connectivity as of 10/2010 (see http://www.census.gov/compendia/statab/cats/information_communications/internet_publishing_and_broadcasting_and_internet_usage.html (table 1155))

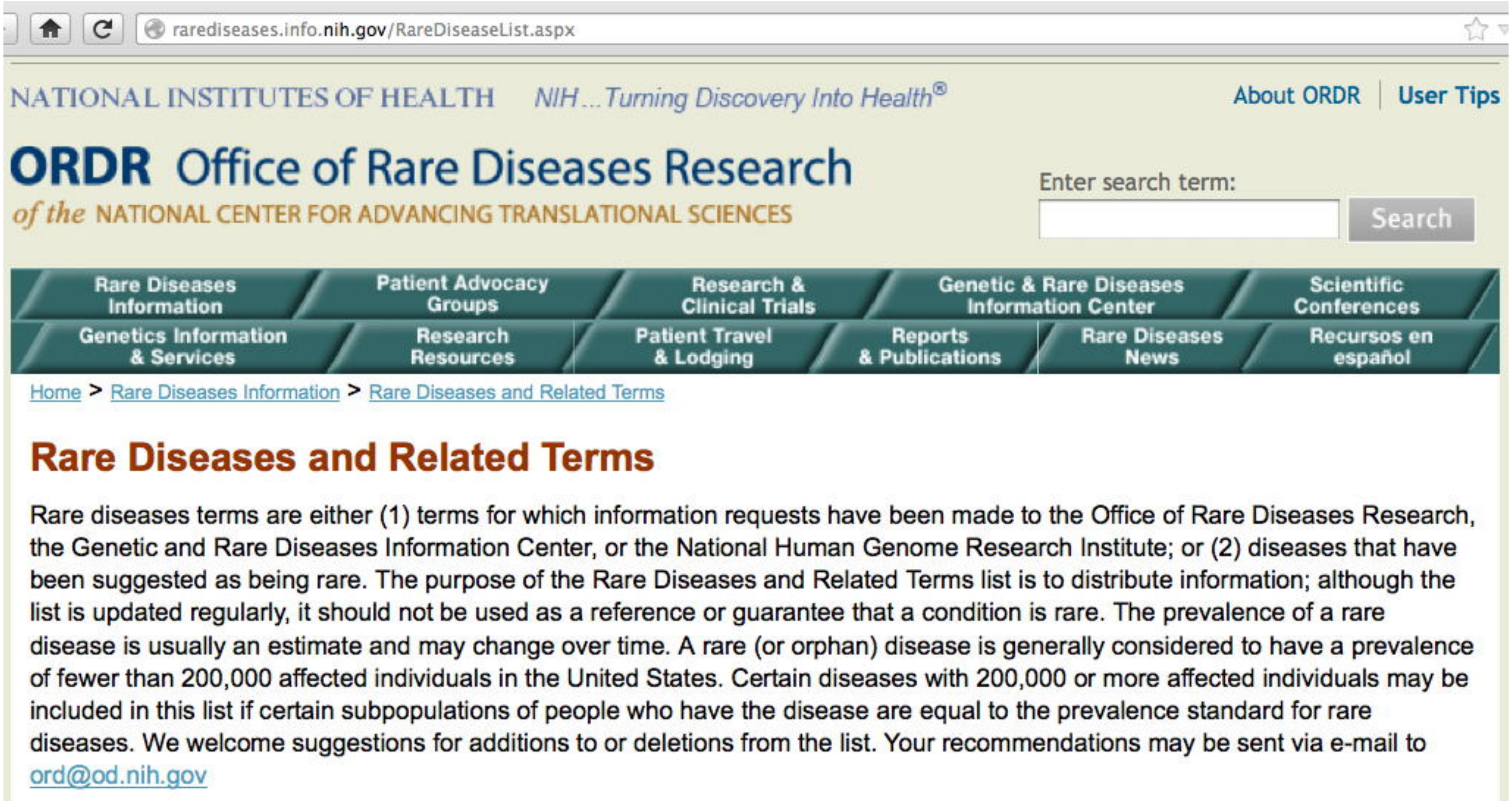
- If 20% of 81.6 million US broadband households were actually to be botted, that would imply that there are 16 million+ bots in the US alone...

- I'm not sure that I "buy" that.

Let's Consider Another Estimate, From the CBL

- On Sunday December 9th, 2012, the Composite Block List knew about 174,391 botted host IPs in the United States.
  See http://cbl.abuseat.org/countrypercapita.html
- There are 245,000,000 Internet users in the US as of 2009 according to the CIA World Fact Book.
- 174,391/245,000,000*100=**0.0711% of all US Internet users are potentially botted** [assuming 1 computer/user]
- Worldwide, that puts the US near the bottom of all countries, in 149th place. On a per capita-normalized basis, **the US is among the least botted of all countries** as measured by the CBL.
- Wonder which nations are the worst? Looking just at countries with 100,000 or more listings, the most-botted countries are **Byelorussia** (137,658 listings, with 5.2% of its users botted), **Iraq** (196,046; 2.4%), **Vietnam** (431,642; 1.85%), and **India** (1,093,289; 1.8%).

If There Are Only 175K Bots in the US, Are Bots Now A "Rare Disease" Here? In Medicine, Rare Diseases Are Those With Less Than 200,000 Affected People in the US...


rarediseases.info.nih.gov/RareDiseaseList.aspx
NATIONAL INSTITUTES OF HEALTH
NIH...Turning Discovery Into Health®
About ORDR
User Tips
ORDR Office of Rare Diseases Research
of the NATIONAL CENTER FOR ADVANCING TRANSLATIONAL SCIENCES
Enter search term:
Search
Rare Diseases Information
Patient Advocacy Groups
Research & Clinical Trials
Genetic & Rare Diseases Information Center
Scientific Conferences
Genetics Information & Services
Research Resources
Patient Travel & Lodging
Reports & Publications
Rare Diseases News
Recursos en español
Home > Rare Diseases Information > Rare Diseases and Related Terms

Rare Diseases and Related Terms

Rare diseases terms are either (1) terms for which information requests have been made to the Office of Rare Diseases Research, the Genetic and Rare Diseases Information Center, or the National Human Genome Research Institute; or (2) diseases that have been suggested as being rare. The purpose of the Rare Diseases and Related Terms list is to distribute information; although the list is updated regularly, it should not be used as a reference or guarantee that a condition is rare. The prevalence of a rare disease is usually an estimate and may change over time. A rare (or orphan) disease is generally considered to have a prevalence of fewer than 200,000 affected individuals in the United States. Certain diseases with 200,000 or more affected individuals may be included in this list if certain subpopulations of people who have the disease are equal to the prevalence standard for rare diseases. We welcome suggestions for additions to or deletions from the list. Your recommendations may be sent via e-mail to ord@od.nih.gov

Coming Back to Our Two Estimates, How Can Those Values Be So Different?

- Key point #1: the two estimates are for **different sorts of things** (households that are detected as being botted vs. botted IPs seen sending spam)

- Key point #2: the two estimates also measure **different populations** (users worldwide who are connected via ISPs with enough of a bot problem that those ISPs are motivated to purchase a commercial network security solution vs. users in the US (where a real push to control bots has been underway for years))

- **CONCLUSION: A key part of our work today on bot metrics will be thinking more carefully, more precisely, if you will, about exactly what it is we want to measure.**

Some Other Questions Are Also Pretty Basic, E.G., "How Many Families of Bots Are Out There?"

- That is, are there three main types of bots actively deployed right now? Thirty? Three hundred? Three thousand? The proposed malware registry should allow us to answer this question...

- The number of unique types of bots tells us a lot about how hard it might be to get the "bot problem" under control.

- Closely related, how many <u>botmasters</u> are out there?

  We might expect that the number of botmasters would roughly track the number of unique bots, but one bot "code base" might be "franchised" and in use by multiple botmasters, or one botmaster might run multiple different bots.

Another Basic Question: How Many Users Are Covered by The ABCs for ISPs Code?

- I know that some organizations have attempted to identify the number of users covered by the ABCs for ISPs Code, but it can be hard to dig out subscriber estimates for participating ISPs.

- Should one of our "metrics" simply be a clean report of how many subscribers are covered by the code?

"Are There Any Trends Relating to Bots?"

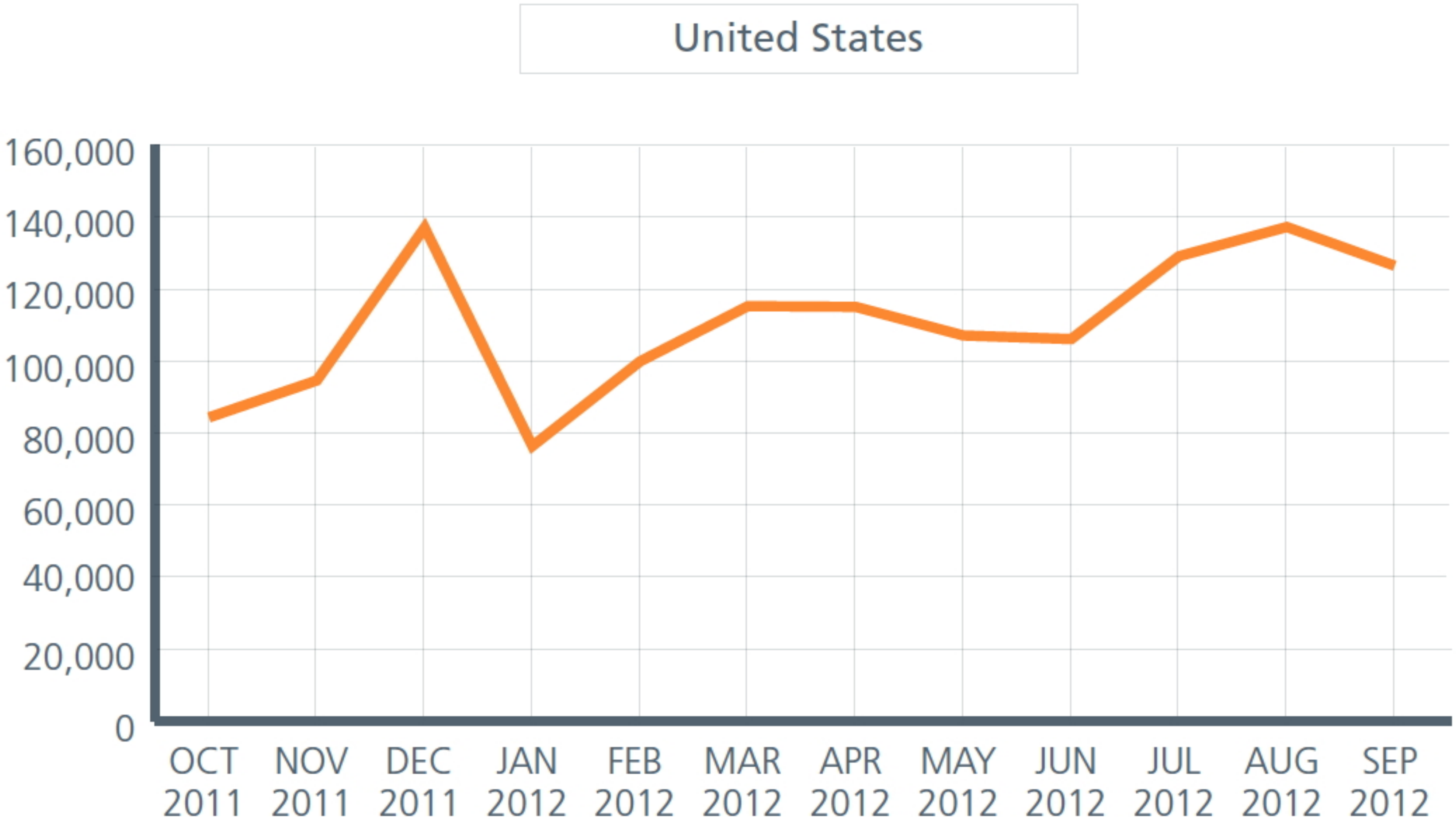
- Another very reasonable general question to ask is,

  "In general, is the bot problem getting better or worse over time?"

- Note that trends of this sort could be shared even if (for whatever reason) a company is unable or unwilling to share absolute measurements (conceptually imagine a graph that has no units on the vertical axis).

- However, we know that we can already get examples of graphs over time WITH measurements...

Graph of US Spam Bots Over Time from McAfee



Source: McAfee Threat Report, Third Quarter 2012, PDF page 30

"Do Bots Show Any Sort of Operational ***Patterns***?"

- That is, for example, hypothetically, does most botnet spam get sent "overnight" when US anti-spam folks are asleep but Europeans have already woken up? (remember, Europe is +7 or +8 relative to the US Pacific Time, right?)
- Does the number of bots increase during the weekend, and then go back down during the week? (This might be the case if a regularly employed botmaster just ran his or her botnet as a way to supplement his or her income on weekends, or if fewer anti-botnet people were paying attention/whacking bots on weekends)
- Does the number of bots increase at the start of the month when people get paid, or peak in the month before Christmas (when people are most likely to be Christmas shopping), perhaps?
- If law enforcement takes down a botnet, can we see a noticeable drop in the amount of spam sent, or do other botnets immediately step up and fill that now vacant niche in the botnet ecosystem?

The Effect of One Law Enforcement Takedown

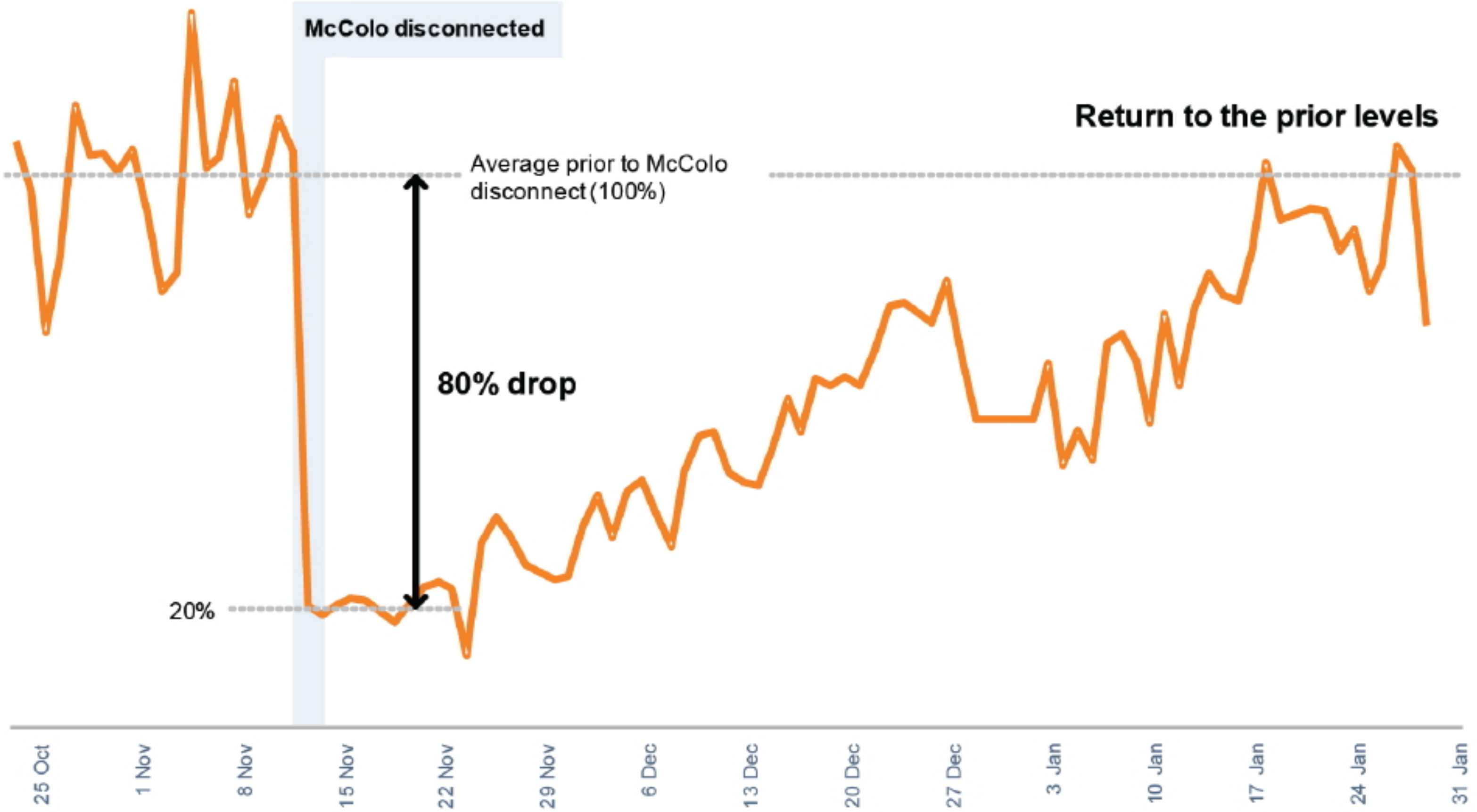
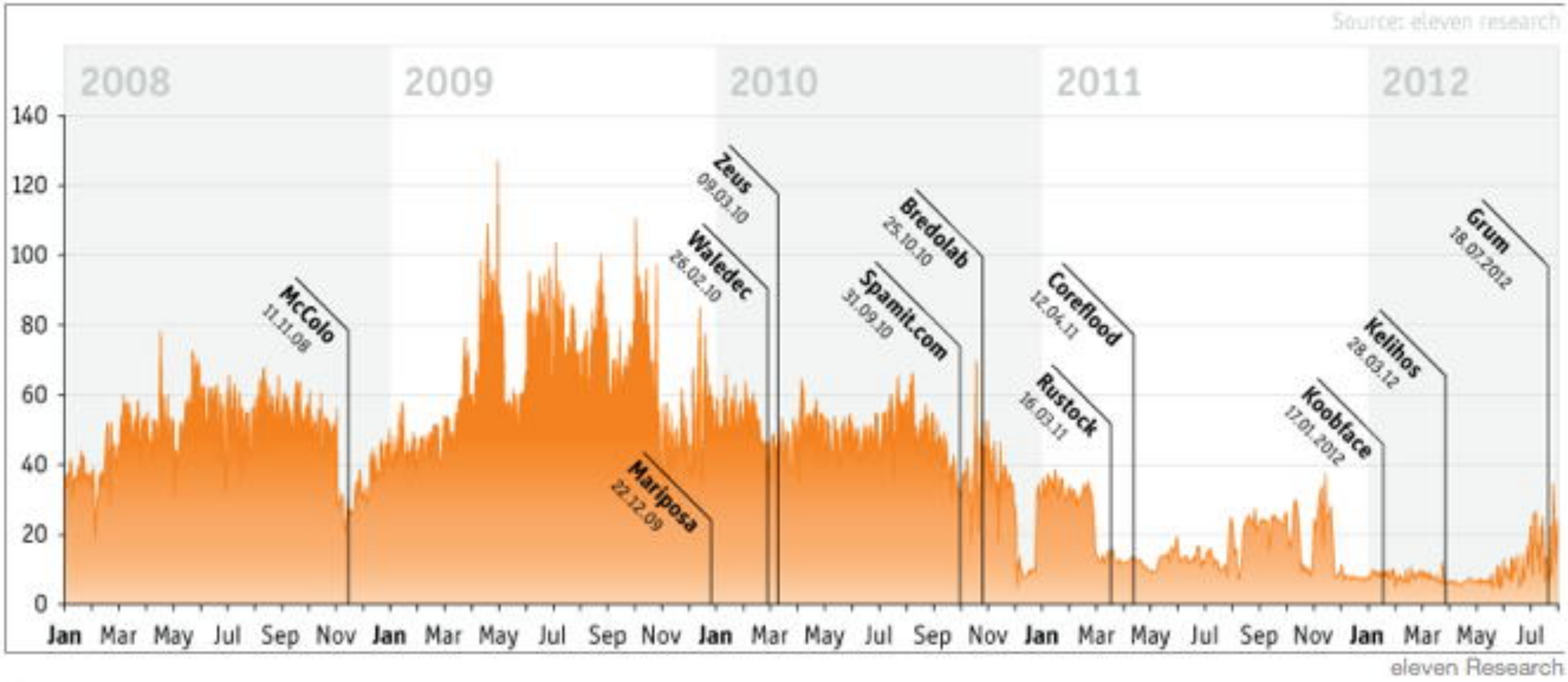


Figure 13: Temporary impact of the shutdown of hosting provider McColo on spam e-mail (Diagram by MessageLabs, Symantec Hosted Services). [209]

Source: "Botnets: Detection, Measurement, Disinfection & Defence," ENISA, page 109.

Botnets and Spam Development

Shutting Down Botnets and the Effect on Spam Volume



Spam volume and botnet shutdowns

Source: http://www.eleven.de/botnet-timeline-en.html

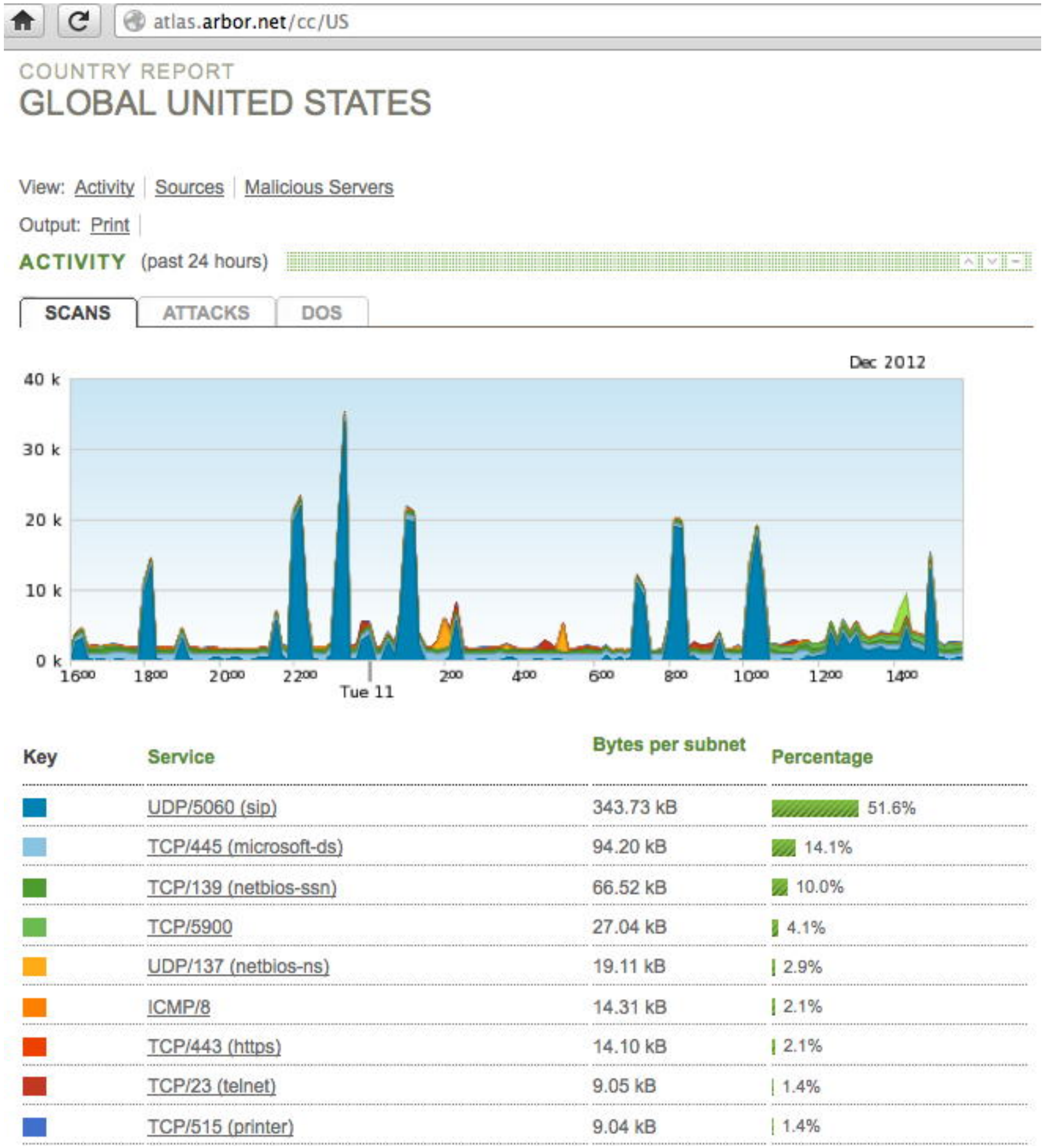
"HOW Are Bots Being Used?"

- That is, on average, what percentage of all bots send spam? (And how many spams get emitted from each single botted host? And are those bots running "flat out" or just "loafing along?")
- What percentage of bots get used to conduct DDoS attacks?
- What percentage of bots get used to scan network-connected hosts for remotely-exploitable vulnerabilities?
- What percentage of all bots get used to host illegal files?
- What percentage of all bots get used to steal private information?
- What percentage of all bots get used as part of a compute farm to crack passwords or mine Bitcoins, etc.?
- Are any bots installed and "live" but totally unused?
- **How bots get used** determines, in part, how we can **measure** those bots (if bots aren't sending spam, for example, we shouldn't attempt to measure bot counts based on observed spam, right?)

Non-Spam Bots

- Not all bots are spam bots.
- Some bots may be used for doing scans, or in an attempt to exploit known vulnerabilities, or for distributed denial of service attacks, or to steal credentials (just to mention a few examples)

Arbor's "Scan" Report for the United States



COUNTRY REPORT

GLOBAL UNITED STATES

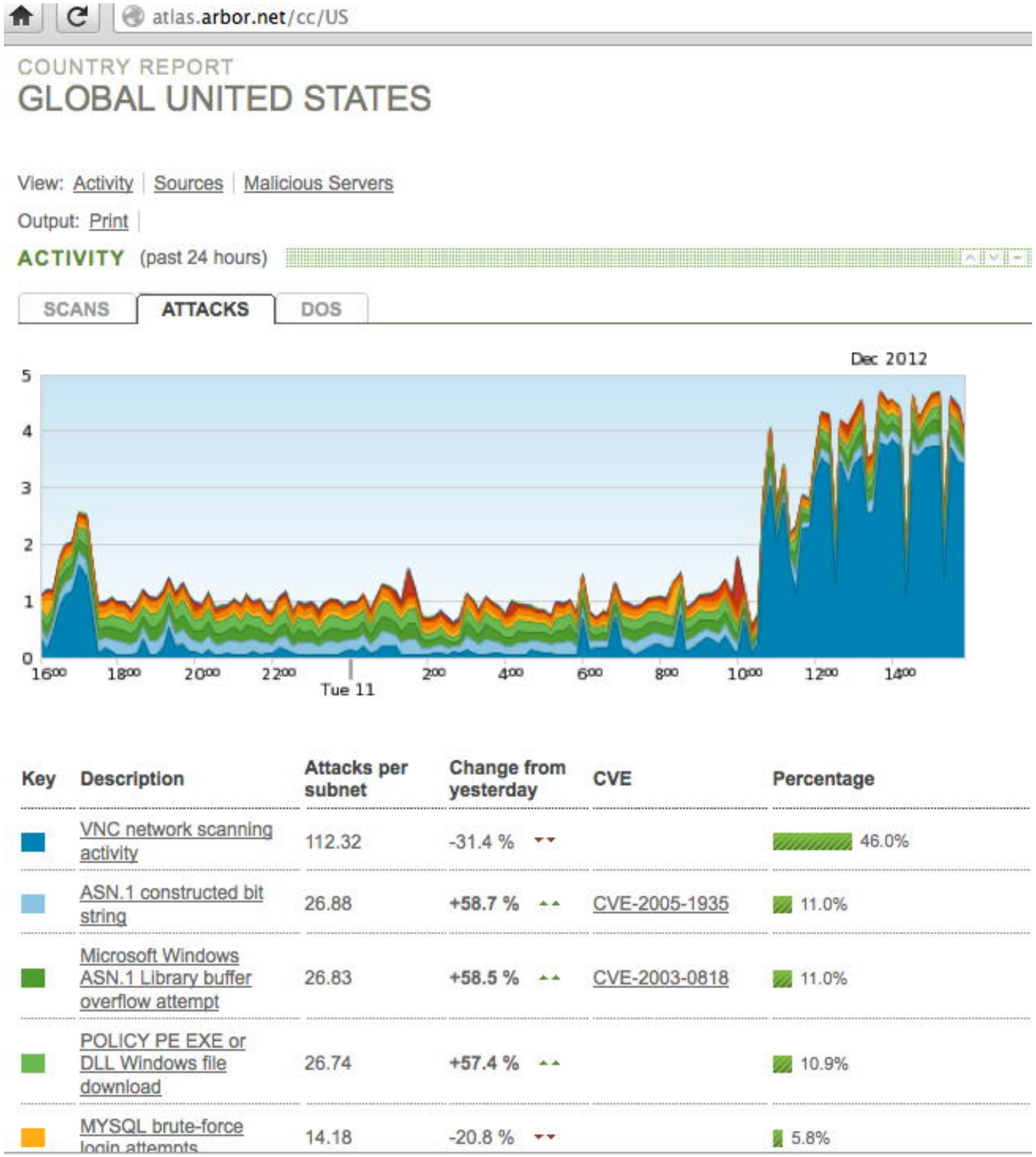
View: Activity | Sources | Malicious Servers

Output: Print |

ACTIVITY (past 24 hours)

Key	Service	Bytes per subnet	Percentage
🟦	UDP/5060 (sip)	343.73 kB	51.6%
🟦	TCP/445 (microsoft-ds)	94.20 kB	14.1%
🟩	TCP/139 (netbios-ssn)	66.52 kB	10.0%
🟩	TCP/5900	27.04 kB	4.1%
🟨	UDP/137 (netbios-ns)	19.11 kB	2.9%
🟧	ICMP/8	14.31 kB	2.1%
🟥	TCP/443 (https)	14.10 kB	2.1%
🟥	TCP/23 (telnet)	9.05 kB	1.4%
🟦	TCP/515 (printer)	9.04 kB	1.4%

Arbor's "Attacks" Report for the United States



COUNTRY REPORT

GLOBAL UNITED STATES

View: Activity | Sources | Malicious Servers

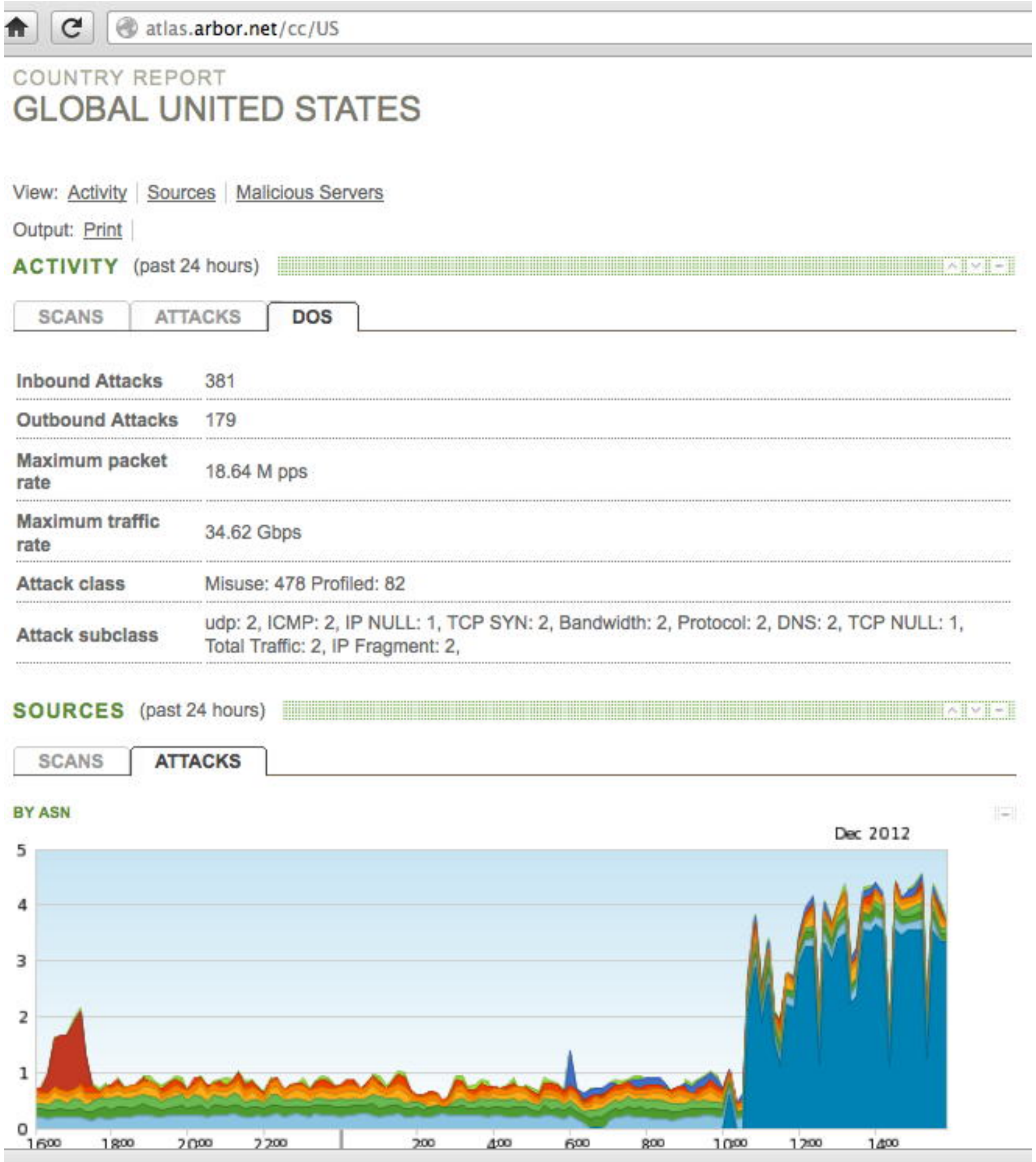
Output: Print |

ACTIVITY (past 24 hours)

SCANS | ATTACKS | DOS

Key	Description	Attacks per subnet	Change from yesterday	CVE	Percentage
🟦	VNC network scanning activity	112.32	-31.4 % ▾▾		▨ 46.0%
🟦	ASN.1 constructed bit string	26.88	+58.7 % ▴▴	CVE-2005-1935	▨ 11.0%
🟩	Microsoft Windows ASN.1 Library buffer overflow attempt	26.83	+58.5 % ▴▴	CVE-2003-0818	▨ 11.0%
🟩	POLICY PE EXE or DLL Windows file download	26.74	+57.4 % ▴▴		▨ 10.9%
🟧	MYSQL brute-force login attempts	14.18	-20.8 % ▾▾		▨ 5.8%

Arbor's "DOS" Report for the United States



atlas.arbor.net/cc/US

COUNTRY REPORT

GLOBAL UNITED STATES

View: Activity | Sources | Malicious Servers

Output: Print |

ACTIVITY (past 24 hours)

SCANS | ATTACKS | DOS

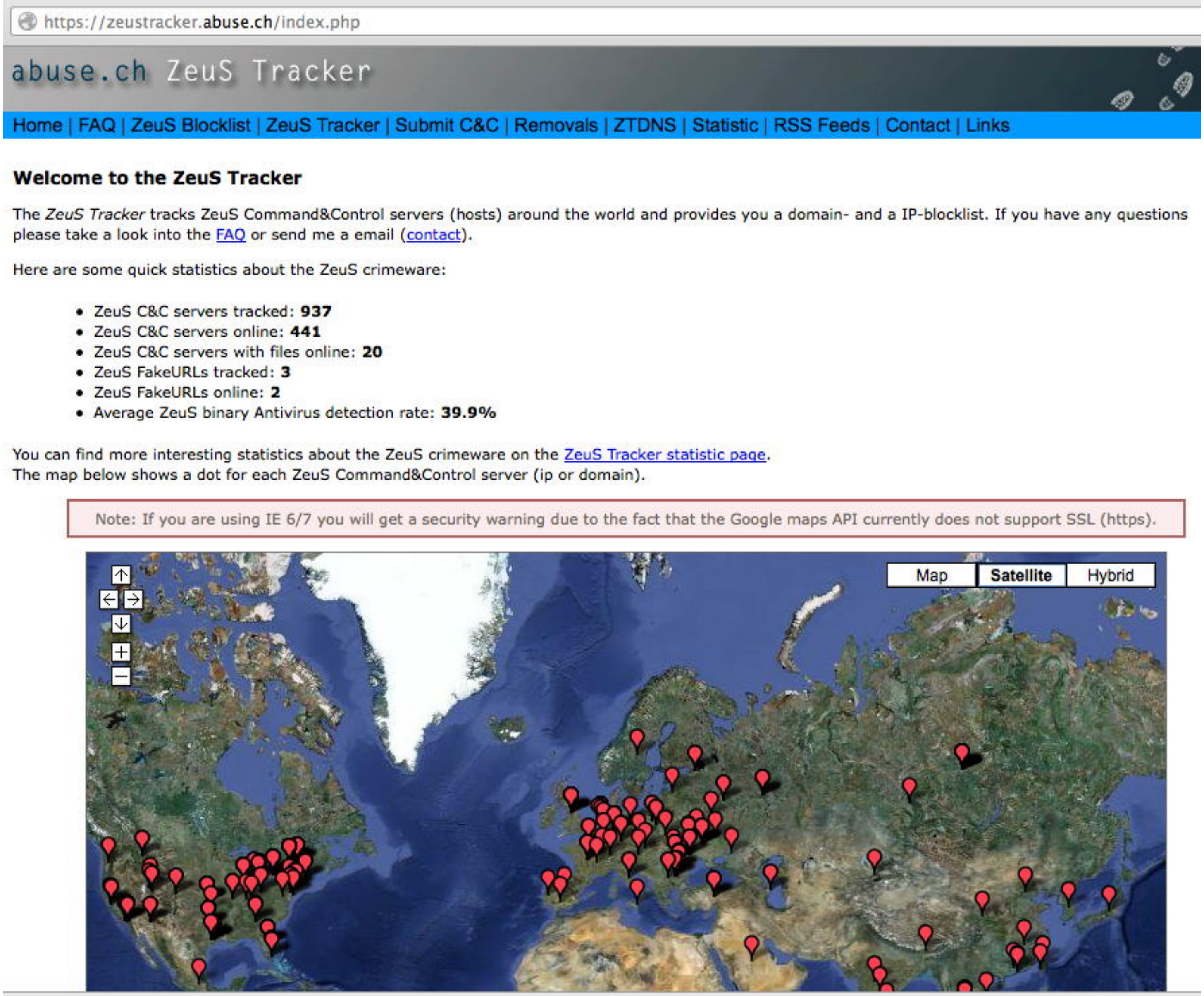
Inbound Attacks	381
Outbound Attacks	179
Maximum packet rate	18.64 M pps
Maximum traffic rate	34.62 Gbps
Attack class	Misuse: 478 Profiled: 82
Attack subclass	udp: 2, ICMP: 2, IP NULL: 1, TCP SYN: 2, Bandwidth: 2, Protocol: 2, DNS: 2, TCP NULL: 1, Total Traffic: 2, IP Fragment: 2,

SOURCES (past 24 hours)

SCANS | ATTACKS

BY ASN

Zeus Tracker



https://zeustracker.abuse.ch/index.php

abuse.ch ZeuS Tracker

Home | FAQ | ZeuS Blocklist | ZeuS Tracker | Submit C&C | Removals | ZTDNS | Statistic | RSS Feeds | Contact | Links

Welcome to the ZeuS Tracker

The *ZeuS Tracker* tracks ZeuS Command&Control servers (hosts) around the world and provides you a domain- and a IP-blocklist. If you have any questions please take a look into the FAQ or send me a email (contact).

Here are some quick statistics about the ZeuS crimeware:

- ZeuS C&C servers tracked: **937**
- ZeuS C&C servers online: **441**
- ZeuS C&C servers with files online: **20**
- ZeuS FakeURLs tracked: **3**
- ZeuS FakeURLs online: **2**
- Average ZeuS binary Antivirus detection rate: **39.9%**

You can find more interesting statistics about the ZeuS crimeware on the ZeuS Tracker statistic page.
The map below shows a dot for each ZeuS Command&Control server (ip or domain).

Note: If you are using IE 6/7 you will get a security warning due to the fact that the Google maps API currently does not support SSL (https).

Another Reason Why The "How Are Botnets Being Used?" Question May Really Matter...

- If I'm an anti-spam person, and bots are no longer being used for spamming, I may stop worrying about bots (happy days!)

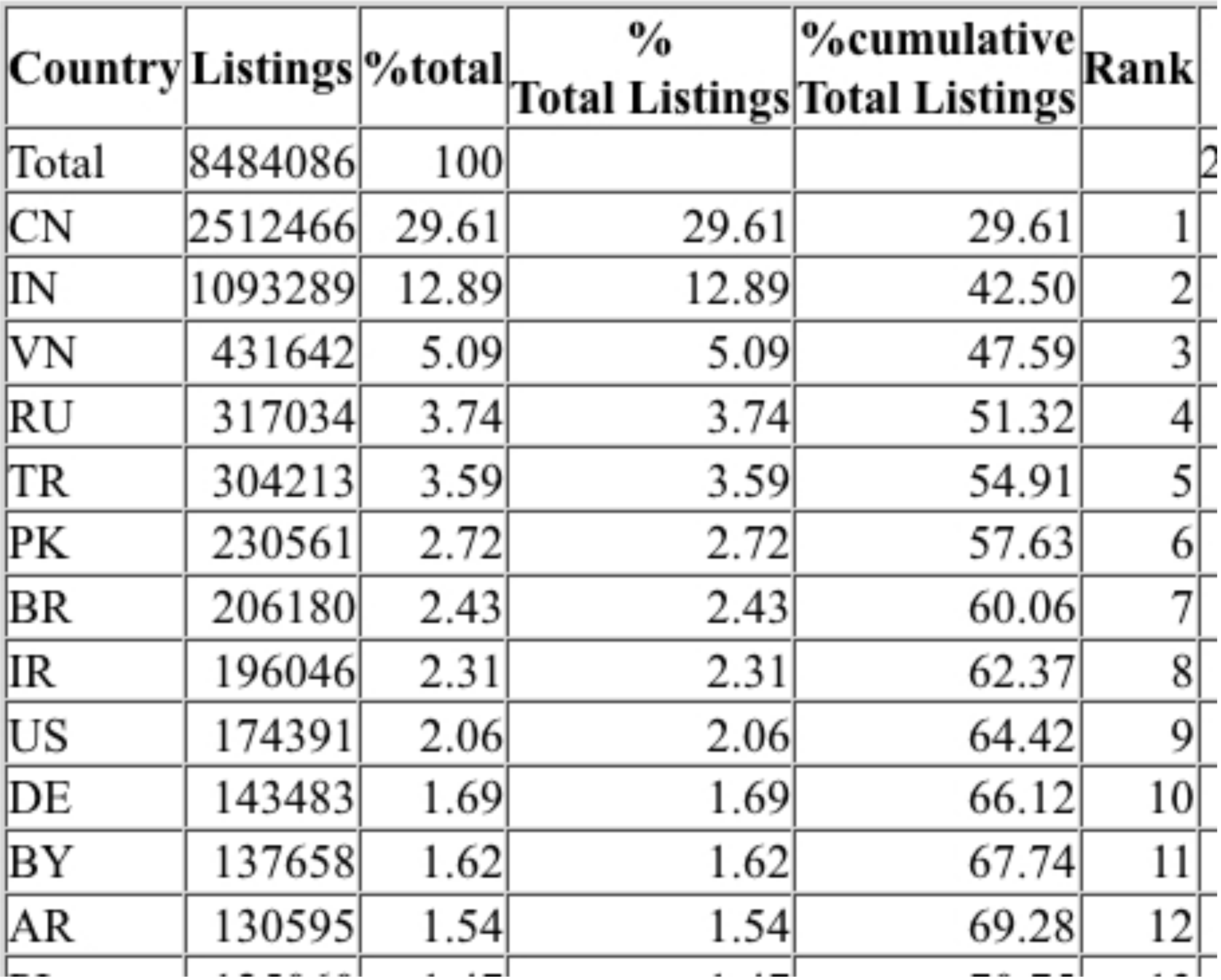
- On the other hand, if bots start to become predominantly used for some other purpose, such as to conduct distributed denial of service attacks against critical government sites, that change in use might INCREASE interest in identifying and mitigating those bots among others...

- We really need to understand/monitor the bot workload profile as seen in the wild, recognizing that this can change as quickly as the weather...

"<u>Comparatively</u> Speaking..."

- Are American computers getting botted more (or less) than Canadian computers, or computers in Great Britain, France, Germany, Japan, Russia, China, Brazil, India or _______?
- Not all countries are the same size. Should we normalize botnet infection rates by the population of each country (or by the number of people in each country who have broadband connectivity?)
- Are all ISPs within the United States equally effective at fighting bots, or are some doing better than others? For example, if an ISP adopts the voluntary "ABCs for ISPs" code do they have fewer bots than other ISPs that don't adopt it?
- Are there other comparative differences that we can identify? For example, are older users (or younger users) more likely to get botted? Does it seem to matter what antivirus product or web browser or email client or operating system people use?

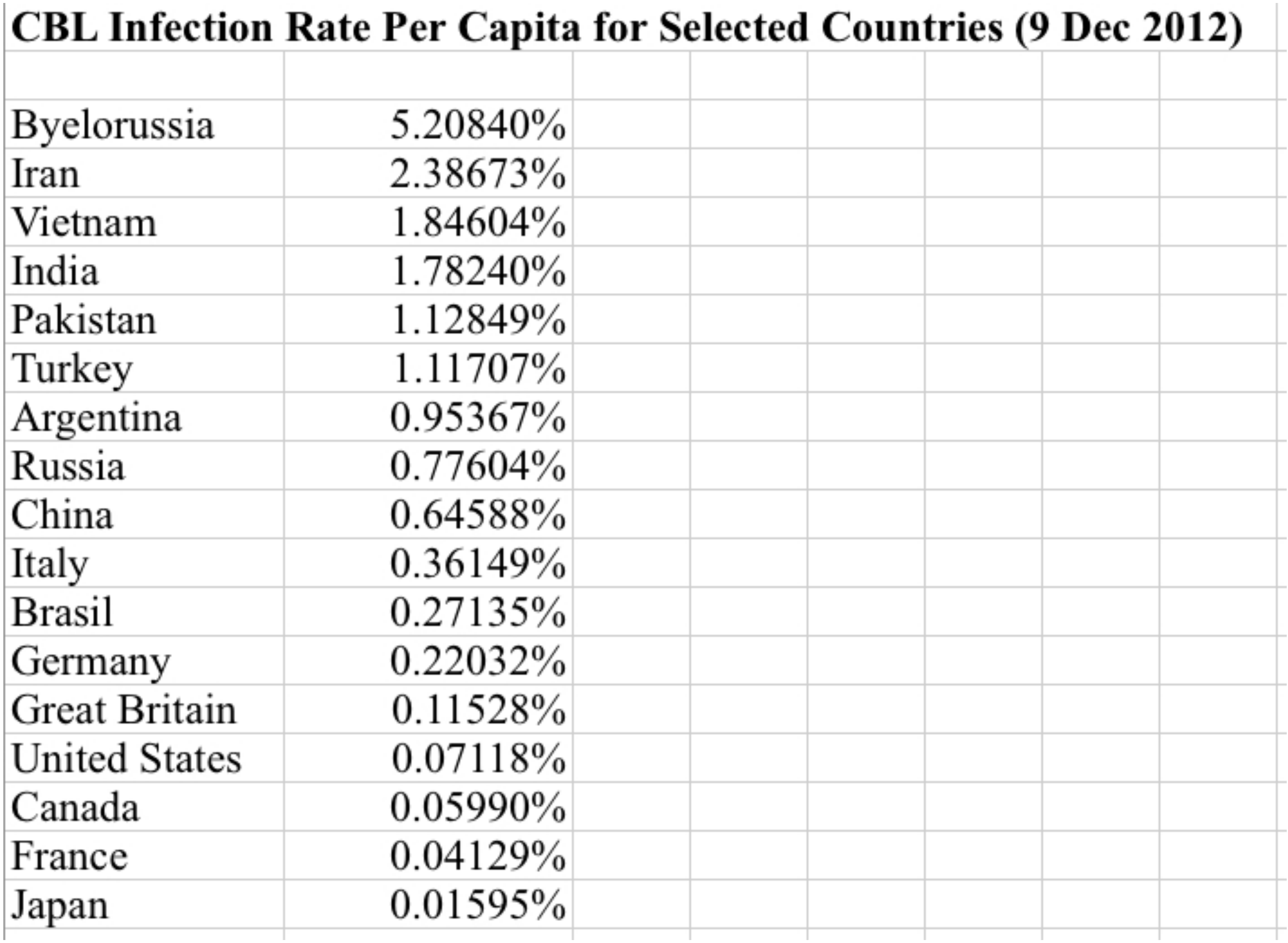
Comparative Bot Levels from the CBL



Country	Listings	%total	% Total Listings	%cumulative Total Listings	Rank
Total	8484086	100			
CN	2512466	29.61	29.61	29.61	1
IN	1093289	12.89	12.89	42.50	2
VN	431642	5.09	5.09	47.59	3
RU	317034	3.74	3.74	51.32	4
TR	304213	3.59	3.59	54.91	5
PK	230561	2.72	2.72	57.63	6
BR	206180	2.43	2.43	60.06	7
IR	196046	2.31	2.31	62.37	8
US	174391	2.06	2.06	64.42	9
DE	143483	1.69	1.69	66.12	10
BY	137658	1.62	1.62	67.74	11
AR	130595	1.54	1.54	69.28	12

Source: http://cbl.abuseat.org/country.html as of Sunday, Dec 9th, 2012

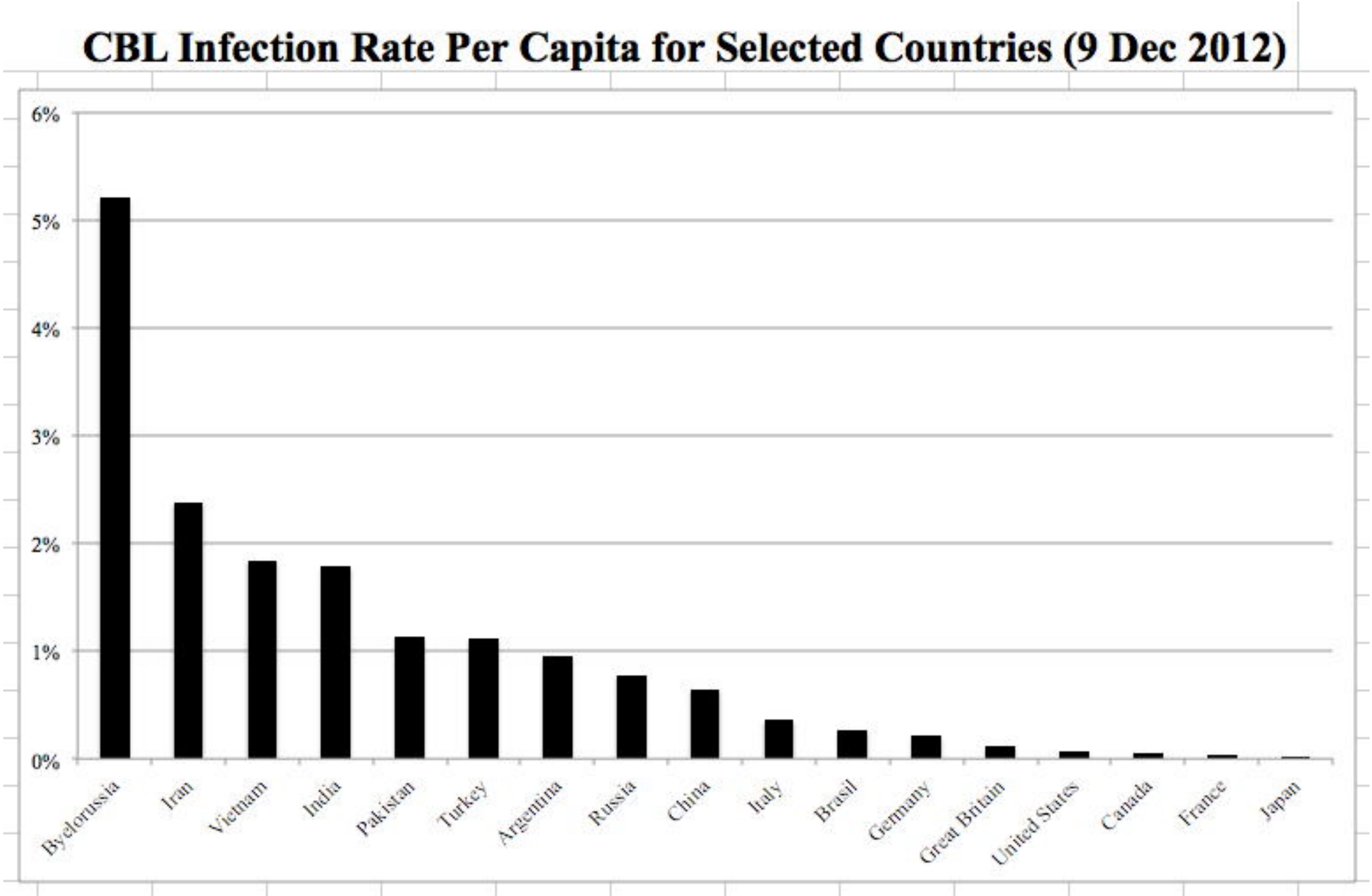
Selected CBL Listings Normalized Per Capita



CBL Infection Rate Per Capita for Selected Countries (9 Dec 2012)

Byelorussia	5.20840%
Iran	2.38673%
Vietnam	1.84604%
India	1.78240%
Pakistan	1.12849%
Turkey	1.11707%
Argentina	0.95367%
Russia	0.77604%
China	0.64588%
Italy	0.36149%
Brasil	0.27135%
Germany	0.22032%
Great Britain	0.11528%
United States	0.07118%
Canada	0.05990%
France	0.04129%
Japan	0.01595%

Source: http://cbl.abuseat.org/countrypercapita.html as of Sunday, Dec 9th, 2012



CBL Infection Rate Per Capita for Selected Countries (9 Dec 2012)

Source: http://cbl.abuseat.org/countrypercapita.html as of Sunday, Dec 9th, 2012

Comparisons → Scalable Access to "Attributes"

- As an outside researcher, perhaps I wonder, "If an ISP subscribes to the "ABCs for ISPs" code, do they have lower levels of botnet infections than ISPs that don't subscribe?"
- In order to be able to make that comparison, I need to know *specifically* which ASNs (or netblocks, or in-addr domain name patterns) represent systems in each category. I do NOT believe that this is currently publicly disclosed for "ABCs for ISPs"-subscribing entities.
- Potential Recommendation: create a machine-accessible list of ASNs (or netblocks, or in-addr domain names, etc.) that represent ISPs adhering to the "ABCs for ISP" code.
- Of those possible options, ASNs will be the easiest to work with/ keep up-to-date, assuming code adoption is uniform within ASNs. (Patchwork adoption might drive a need for listing CIDR blocks)

"How Many Systems Are Infected with ***Bot Foo***?"

- In addition to measurements made about overall bot infectivity, we also need the ability to "drill down" and get more precise estimates on a bot-family-by-bot-family basis, ideally both at any given point in time, and historically.

- Per-bot-family measurements might include the number of systems infected with each particular major bot, but also related measurements such as:

  -- the amount of spam attributed to each particular spam botnet
  -- the volume of DDoS traffic attributed to each DDoS botnet
  -- the number of command and control hosts that a bot uses
  -- the geographic distribution of hosts infected with each bot

Per-System Bot-Related Measurements

- Not all metrics are macroscopic measurements related to botnet infection rates. Some might be per-system micro values:

  -- **What does it cost to rent a bot** on the open market?

  -- **How long does it take/what does it cost** to de-bot a single botted host? What determines that range?

  -- When a system is found to be botted, does it tend to be botted with **just one type of bot**? If co-infections are routinely found, can we identify "clusters" of bot malware that are routinely found together?

  -- **If a user's botted once, does that make them more (or less) likely to get botted again?** That is, can we expect that that a once-botted user will become less likely to be rebotted as a result of that experience? Or are some types of users just inherently more prone to get themselves cr*pped up? If users do end up rebotted, what's the time till reinfection?

Coming At This From a Different Direction: What's The Expected Lifespan of a Botted Host?

- For example, hypothetically assume that you're running a blocklist, and you list the IP addresses of botted systems when you see those systems send spam or check in with a C&C you're monitoring.

- If you don't observe any subsequent activity from a botted and blacklisted system, when could you "safely" remove it?
  After a day? After a week? After a month? After 90 days? Never?

- Some botnet blocklists deal with this issue by simply rolling off the oldest entries after the list reaches some target maximum size (after all, if the system turns up being bad again, you can always freshly relist it)...

Measuring Botnet Backend Infrastructure

- While we've been talking about botted end user hosts, another potential target for measurement is botnet backend infrastructure, such as botnet command and control hosts (see the Zeus Tracker example shown earlier in this section).

- Potentially one could also track authoritative name servers associated with bot-related domains, and sites known to be dropping bot malware, and a host of other botnet-related things (other than just botted hosts).

- *A philosophical aside:* is there any risk that focusing on backend botnet infrastructure (including potentially doing C&C takedowns) will result in interference with ongoing legal investigations?

How Precisely Do We Need To Measure Our Answers to All These Questions?

- "High precision" answers cost more than "rough" answers. (Think of this as the width of a confidence interval around a point estimate)

- If you want to estimate a value within +/- 10%, that requires less work than if you want to know that same value within +/- 5% or even +/- 1%

- How precise do our measurements need to be?

How Much Confidence Do We Need That Our Estimates Include the Real Value?

- For example, if we need 99% confidence that our estimate includes the real value for a parameter of interest, we can get that level of confidence, however, getting 99% confidence might require accepting broader bounds around an estimate (or drawing more observations) than we'd need if we could live with just a 90% level of confidence.

- Notice the interaction between (a) the required precision, (b) the required confidence, and (c) **the cost of obtaining those answers** (typically the number of observations required).

- **Most people want high precision and high confidence and low cost, but you <u>can't</u> have all three at the same time.**

Budget

- Let me emphasize that if need bespoke hard numerical answers to questions about botnets, it's going to **cost money**.

- How much are we willing to spend to get those answers?

- If the answer is "zero," then I would suggest that in fact these questions are just a matter of simple curiosity, and not something that's actually valuable ("value" implies a willingness to pay)

- If we also don't have a budget for data collection, our ability to prudently set the required level of precision (and the required confidence in our estimates) is also going to be impaired.

5. ISPs As A Potential Source of Botnet Data

The ISP As A Potential Source of Botnet Data

- **The WG7 metrics presumption has inherently been that ISPs themselves might be a potential source of bot data about their botted customers.** While this is an understandable assumption, it might be problematic in practice for multiple reasons:

- **Collecting botnet metrics requires time and effort. Who will reimburse ISPs for the cost of this work, or for the capital costs associated with passively instrumenting the parts of the ISP's network that may not currently be set up to gather the required data?**

- There are many ISPs around the world. Many will not participate. Incomplete participation (even simple addition and subtraction of participating ISPs) will complicate data interpretation and analysis.

The ISP As A Potential Source of Bot Data (2)

- ISPs may also be reluctant to share customer bot detections because of the distinct possibility that bot detection statistics will be misinterpreted. For example, if ISP A has a higher level of bot detections than ISP B, does that mean that ISP A is "better" at rooting out botted customers than ISP B? Or does it mean that ISP A customers are inherently "less secure" than ISP B customers? Or does it mean that the bad guys are simply attacking ISP A customers more aggressively than ISP B?

- Customers, third party privacy organizations, and some governments (rightly or wrongly) may view ISP data sharing about botnets as a potential infringement of ISP customer privacy, even if all customer data is highly aggregated. (Notice the tension between customer privacy and the methodological ideal of gathering fine grained data with unique identifiers)

The ISP As A Potential Source of Bot Data (3)

- Different ISPs may measure botted customers differently (efforts at standardization notwithstanding), undermining the comparability of inter-ISP botnet metrics.

- Self-reported and unaudited data bot may be subject to error or manipulation, or at least the perception by some that it might not be fully candid/fully accurate/fully trustworthy.

- Finally, we need to recognize that most bots are not domestic, while ABC for ISP-code participants are. Thus, US ISPs are poorly positioned to provide detailed botnet intelligence on most of the bots that are actually hitting US targets. You need other entities, entities with a global footprint.

Entities With Global Footprints

- There are entities other than ISPs with consistent global visibility into the bot status of Internet systems and users:

  -- Operating system vendors that do periodic patching and malware removal (classic example: Microsoft with its Malicious Software Removal Tool that runs every month at patch time)

  -- Anti-virus or anti-spam companies with large international customer bases

- **These entities already produce public reports about the data that they've collected. Are we taking adequate advantage of what's already been published? If not, why not? We should NOT be "reinventing the wheel," particularly if we don't have budget to collect botnet data carefully and consistently.**

Examples of Some Cybersecurity Data Reports

- **Composite Block List Statistics**
  http://cbl.abuseat.org/statistics.html
- **Kaspersky Security Bulletin/IT Threat Evolution**
  http://www.securelist.com/en/analysis/204792250/IT_Threat_Evolution_Q3_2012
- **McAfee's Quarterly Threats Report**
  http://www.mcafee.com/apps/view-all/publications.aspx?tf=mcafee_labs&sz=10&region=us
- **Microsoft's Security Intelligence Report (SIR)**
  http://www.microsoft.com/security/sir/
- **Shadowserver Bot Counts**
  http://www.shadowserver.org/wiki/pmwiki.php/Stats/BotCounts
- **Symantec's Internet Security Threat Report (ISTR)**
  http://www.symantec.com/about/news/resources/press_kits/detail.jsp?pkid=threat_report_17

A Particularly Noteworthy Existing One Time Botnet Report

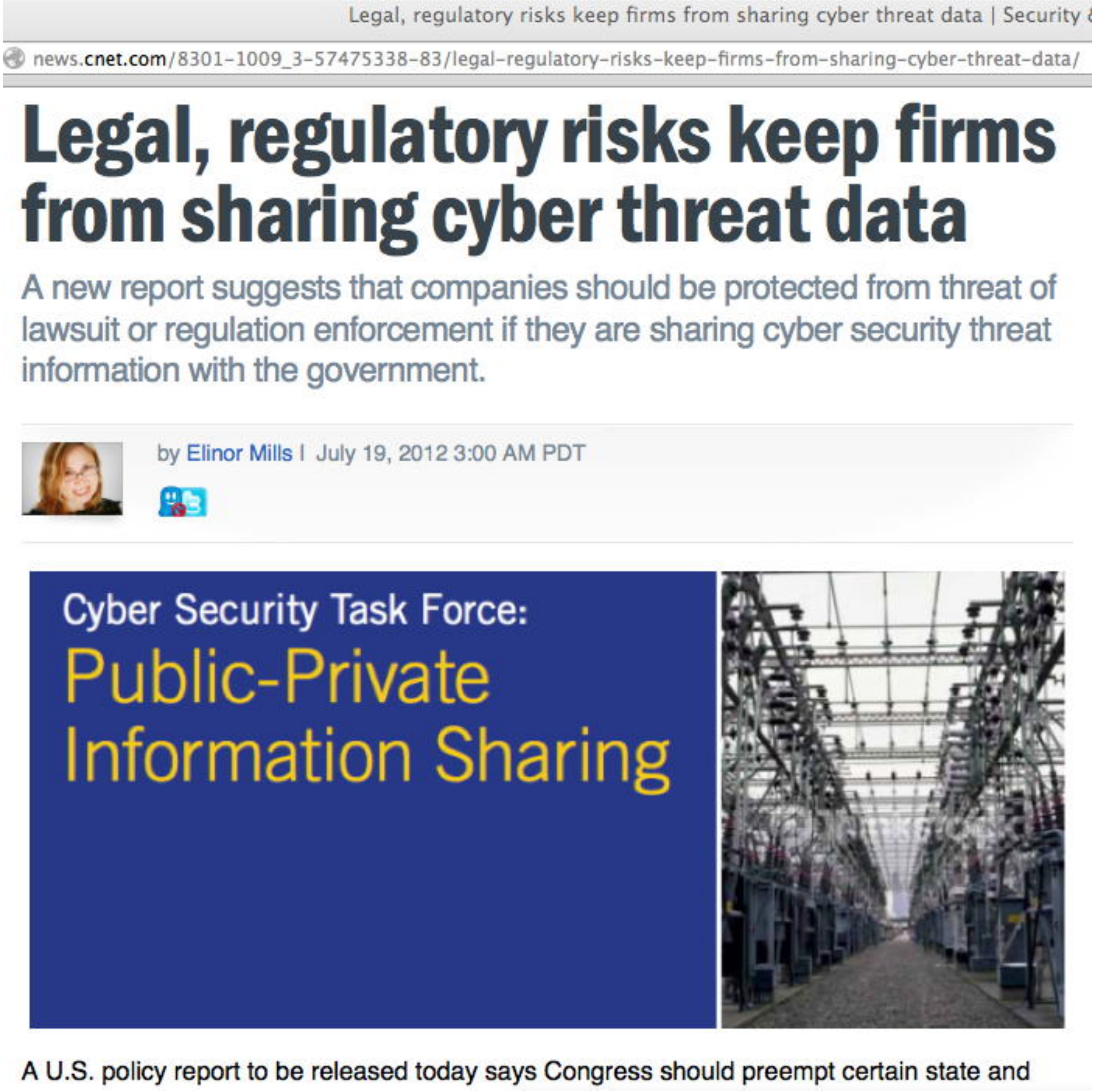
- *Botnets: Detection, Measurement, Disinfection & Defence*, European Network and Information Security Agency, 7 Mar 2011, http://www.enisa.europa.eu/activities/Resilience-and-CIIP/critical-applications/botnets/botnets-measurement-detection-disinfection-and-defence

Pre-ABC Code/Post-ABC Code "Per-ISP" Studies

- Given that it may be difficult to compare bot-related statistics collected by ISP A with bot-related statistics collected by ISP B, another option might be to study botnet stats longitudinally, within an individual ISP, over time.
- For example, assume the FCC would like to know if an ISP has fewer botted customers after adopting the ABCs for ISPs than before (this is what some might call a "pre/post" study). If so, we'd expect to see a downward sloping curve...
- Fortunately or unfortunately, many of the most important/most interesting ISPs have already implemented important parts of the ABCs for ISPs code.
- Thus, we cannot get a "clean" "pre" "baseline" profile for many ISPs because the ISPs have ALREADY begun doing what the ABCs for ISPs code recommends. ("No fair being so proactive/on the ball when it comes to dealing with your botted customers!" ☺)

How Might the FCC Encourage ISPs to Voluntarily Submit Their Own Stats?

- Hypothetically, assume that no ISPs voluntarily submit metrics on their botnet experiences to the FCC. [BTW has the FCC explicitly said that they'd be interested in receiving these, and told ISPs about an address/department to which such data might be sent?]
- In that case, having no other option, let's assume that the FCC begins to look at publicly available third party data sources, and begins to use that data as a basis for evaluating ISP performance when it comes to combatting bots.
- Let us further assume that the 3rd party data the FCC obtains is inconsistent, or the 3rd party data they obtain is radically different from what ISPs believes to be accurate. Those data discrepancies might potentially motivate an ISP to voluntarily contribute data supporting their alternative (and more authoritative) perspective.



Legal, regulatory risks keep firms from sharing cyber threat data | Security

news.cnet.com/8301-1009_3-57475338-83/legal-regulatory-risks-keep-firms-from-sharing-cyber-threat-data/

Legal, regulatory risks keep firms from sharing cyber threat data

A new report suggests that companies should be protected from threat of lawsuit or regulation enforcement if they are sharing cyber security threat information with the government.

by Elinor Mills | July 19, 2012 3:00 AM PDT

Cyber Security Task Force:
Public-Private Information Sharing

A U.S. policy report to be released today says Congress should preempt certain state and

Sharing: Where's the Mutuality/Reciprocity?

- If we view ISPs and the government as partners that share a common interest in tackling bots and improving cyber security, and if we believe that both parties believe in collaborative data driven security work, it would be terrific if operational data sharing could be bidirectional.
- That is, if ISPs are good about sharing botnet metric data with the FCC, how will the FCC reciprocate and share data back with the ISPs? Data sharing partnerships should not be just unidirectional, just industry to government, for the pleasure of the government!
- Yes, I know, there are limits to what data can be shared by the government to ISPs, and there was legislation proposed to deal with this issue, but that legislation hasn't passed to-date.

6. Alternatives to ISP Provided Bot Data and Bot Data From Global-Footprint Third-Party Commercial Entities:

Sinkholing, DNS Based Methods, Surveys, and Simulations

Sinkholing Specific Botnets

- Sometimes a researcher or the authorities are able to gain control of part of a botnet's infrastructure. When that happens, the researcher or government person may be able to direct botnet traffic to a sinkhole, and use the data visible via that sinkhole to measure the botnet.

- Some might hope that sinkholing would provide a general purpose botnet estimation technique. Unfortunately, because this is a bot-by-bot approach, and requires the researcher or authorities to "inject" themselves into the botnet's infrastructure, it will not work to get broad ISP-wide botnet measurements for all types of botnets.

DNS-Based Approaches

- Another approach that's sometimes mentioned is measuring botnets by their DNS traffic. That is, if you know that all botted hosts "check in" to a specific fully qualified domain name, if an ISP see a customer attempt to resolve the magic bad domain name, there's a substantial likelihood that that customer is botted (unless that customer happens to be a security researcher ☺)
- Big botnets might generate more DNS traffic of this sort than small botnets, but remember that DNS traffic patterns can be weird due to things like caching effects, and not all bots even use DNS.
- Some might also be tempted to using an RPZ-like approach (as implemented in BIND) to "take back" DNS and prevent bots from using DNS as part of their infrastructure. [Obligatory double black diamond warning/steep and dangerous run ahead: remember SOPA/PIPA and its attempts to leverage DNS for policy-driven purposes]

Surveying Internet Users To Find Botted Hosts?

- Assume that you want a direct information-gathering approach that doesn't rely on ISPs providing data, or on third party data. That is, you want to go out and collect your own data, much as survey research groups survey entities about political viewpoints, consumer spending, etc.

- How many individuals might you need to survey to get sufficient data about botted users?

- The required number of users will depend on the breakdown between botted and non-botted users, and the number of ISPs whose customers you'd like to directly track.

1.5% Botted Users? 0.0711% Botted Users?

- If you don't have "hints" about who's a botted user *a priori*, and you just need to discover them "at random," you may be facing a daunting task if bots are indeed a "rare disease."
- Let's arbitrarily assume you want 350 botted users to study.
- If 1.5% of all users are botted, on average you'd see 15 botted users per thousand. Given that you want 350 botted users, that would imply you'd likely need to check (350/15)*1,000=23,300 users in order to find the 350 botted users you needed.
- But now if just 0.0711% of all users are botted (recall that this is the current CBL rate for the US), on average you'd see just 0.711 botted users per thousand. To get 350 botted users to study, you'd probably need to check (350/.711)*1,000=492,264 users (ouch)
- Now assume that you want 350 users PER ISP, and assume you're interested in a dozen ISPs... that's a LOT of users to look at!

"Hi, I'm From the FCC. I'd Like To Check Your Computer for Bots..."

- Assume that you were charged with going out and checking 492,264 computers to see if those systems had been botted. To keep this simple, let's assume that we'll call a system botted if a bot is found when we run a commercial antivirus product on that system.
- If we assume that it would take an hour to run a commercial antivirus program on one machine (probably a low estimate given the increasing size of consumer hard drives today), and techs work 40 hours a week, it would take 492,264/40 = ~12,307 "technician weeks" to scan all those systems. If a tech works 50 weeks a year, that would be 246.14 "technician years" worth of work. If we assume an entry-level antivirus tech earns even $50,000/year (salary plus benefits), and neglecting all other costs (managerial/ supervisory salary costs and software licensing costs and travel costs, etc.), our cost would be 246.14*50,000=**$12.3 million**

"Hey, You're Not Going to Scan MY Computer"

- I suspect that many users would not be willing to allow a random government-dispatched technician to "scan their computer."
- Personally owned computers often have intensely private files -- financial information (tax records, brokerage information, etc.); medical records; private email messages; goofy photographs, etc.
- In other cases, users may even have out-and-out illegal content on their systems (simple example: pirated software/movies/music).
- Given these realities, many users would probably simply refuse to allow their computer to be checked, even if they thought that their system might be infected.

What If We Just Asked For Volunteers?

- Some users who think that their systems might be infected might welcome the opportunity to have their systems scanned.

- However, a "convenience sample" of that sort would not result in data that would allow us to generalize or extrapolate from the sample to the population as a whole.

<u>Simulating</u> Bot Infections in a Lab/Cyber Range?

- Another option, if we wanted to avoid the problems inherent in surveying/checking users (as just discussed), might be to try simulating bot infections in a lab or on a so-called "cyber range."

- This might not be easy. For example, the fidelity of the results from such a simulation will depend directly on researchers ability to:

  -- replicate the full range of systems seen on the Internet (operating systems used, antivirus systems used, applications used, patching practices, etc. – do we have the data we'd need to do that?)
  -- replicate the range of botnet malware seen on the Internet
  -- accurately model ISP response to the malware threat

  I believe this is a fundamentally impractical approach.

Appendix A.

Mere Nuisance or National Security Threat?

- While botnets are often thought of purely as a nuisance, e.g., a source of spam and similar low grade unwanted Internet traffic, bots *have* also been used to attack government agencies and Internet-connected critical infrastructure. Viewed in *that* light, bots might properly be considered a threat to national security.
- If bots are indeed a threat to national security, "other government agencies" may be able to directly apply "national technical means" to collect intelligence about bots, including per-ISP estimates.
- Such information, once collected, might then be able to be shared with appropriately cleared colleagues in the FCC and/or in DHS, assuming they have a legitimate need-to-know.
- If domestic collection mechanisms aren't an option or appropriate, it may also be possible to make estimates about domestic bot populations based on data collected by international counterparts.