**BOTNET METRICS**

**Basis and Outline of This Report**

The CSRIC III *Working Group Descriptions and Leadership* document (last updated November 15th, 2012),[[1]](#footnote-1) provides among other things that "[Working Group 7] shall identify performance metrics to evaluate the effectiveness of the ISP Botnet Remediation Business Practices at curbing the spread of botnet infections."

This report is provided in fulfillment of that requirement, and has the following structure:

I. Expected Audiences for This Report 2

II. Thinking Precisely About What Is and Isn't A "Bot" 6

III. What Sort of Botted "Things" Should We Be Trying to Count? 9

IV. Substantive Questions About Bots 15

V. Some Statistical Questions Associated With Botnet Measurements 22

VI. ISPs As A Potential Source of Botnet Data 23

VII. Sinkholing, DNS-Based Methods, Direct Data Collection and Simulations? 25

VIII. Recommendations 27

Appendices 28

**I. Expected Audiences for This Report**

While the primary audience for this report on botnet metrics is the FCC CSRIC itself, it is not the *only* such audience. Many other audiences also need or want botnet metrics, including ISPs, the public, other federal agencies, Senators and Representatives, public interest organizations, members of the media, security software and security hardware vendors, law enforcement agencies, cybersecurity researchers and governments abroad.

These various audiences may have widely varying botnet metric information needs:

(a) The CSRIC and the FCC may wonder, "Do we have empirical evidence that the Anti-Botnet

Code is helping, or do we need to try some other approach? If it is helping, how much?" "Overall,

can we say that the botnet problem is getting better, or getting worse?" "What anti-botnet

strategies have proven to be the most (and the least) successful?" "Are US ISPs doing as

well at tackling bots as their counterparts in Canada, Germany, Japan, Australia, etc.?"

(b) ISPs considering adoption of the code may wonder, "Is the Anti-Botnet Code for ISPs

worthwhile? What might it cost us to participate? What benefits might accrue to us if we do so?

Do the benefits exceed the costs? We need numbers!"

(c) The public may want data that will allow them to choose an ISP that takes the bot problem

seriously, and has taken effective steps to deal with bots targeting customers.

(d) Other federal agencies may wonder, "We've been fighting bots, too. How effective have the

FCC's efforts been compared to ours? Are there opportunities for us to collaborate on targeted

joint initiatives? If so, where's the low-hanging fruit?"

(e) Senators and Representatives may need botnet metrics to determine if new legislation is required,

or if existing legislation requires additional funding in order to be fully effective.

(f) Public interest organizations may want botted users to be protected from botnet threats, but only

in a way that's appropriate and privacy-respectful. A sense of the magnitude of the problem is

critical to sizing up what might be necessary, and metrics also provide programmatic

transparency.

(g) Member of the media will be interested in understanding and reporting on government efforts and

initiatives, will want to see documentation about how much work is being done on botnets, and

to what effect.

(h) Security software and security hardware vendors may view botnet metric requirements as

potentially driving new markets for new security gear -- or defining how well their existing gear

works. In a competitive marketplace, metrics may define "winners" and "losers" and be important

drivers for keeping existing customers and gaining new ones.

(i) Law enforcement agencies may eagerly seek botnet metrics to help them to target and optimize

their enforcement activities, endeavoring to target their limited cyber enforcement dollars in a

way that gives taxpayers the most "bang for their buck."

(j) Academic and commercial sector cybersecurity researchers might want access to raw empirical

data about botnets for use in their own analyses.

(k) Governments overseas may look at our bot metrics to see if this program is something that they

should be doing, too.

An unfocused/*ad hoc* botnet metrics program is unlikely to serendipitously meet the requirements of all those diverse audiences. The metrics that are needed now -- and that may be needed in the future -- must be explicitly and carefully defined, or we run the risk of finding ourselves with no evidence with which to answer critical operational and policy questions relating to bots.

At the same time, we must remain cognizant of the fact that collecting and reporting data about bots is potentially burdensome, intrusive, and expensive. Therefore, any data that is targeted for collection should be data that's needed and which will be used in meaningful ways that justify the cost of its acquisition.

A Specific NON-Audience: Botmasters and Other Cybercriminals: While there are many legitimate audiences that are welcome to industry botnet metrics, there is one explicit non-audience: botmasters (and other cyber criminals). We need to explicitly recognize that botnet metrics, done wrongly, have the ability to potentially help our enemies and undercut anti-botnet goals. A few simple examples:

(a) Some ISPs may worry that if publicly identified as working diligently to combat botnets, they

may be targeted for serious and ongoing Distributed Denial of Service (DDoS) attacks by

unhappy botmasters.

(b) Giving detailed and accurate information about where and when bot activity was observed may

be sufficient for a botmaster to identify (and subsequently avoid!) honeypots (or other data

collection infrastructure) in the future. If that happens, valuable (sometimes literally

irreplaceable) data collection sources and methods may be compromised.

(c) If our botnet metrics include "per-bot" cleaning and removal statistics, botmasters might be able

to use that feedback to learn what bots have proven hardest to remove, information that they can

then use to "improve" future bots, making them harder to mitigate or remove. We don't want

to teach our enemies how to better technically overcome our defensive strategies.

Our botnet metrics efforts must be structured so as to avoid accidentally benefiting our opponents or inhibiting our ultimate goals.

A Concrete Botnet Metrics Example From the Media: The following quote from the online cybersecurity industry news site Dark Reading[[2]](#footnote-2) nicely underscores why botnet metrics matter, and can be hard to do well:

*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.*

That anecdote nicely illustrates some of the metrics-related challenges the industry faces:

(a) All types of malware are treated as if they represented "bots," even though some of the most

common types of malware are not even remotely "bot"-like. We need to be very precise about

what is and isn't a bot if we're to collect any sort of useful numbers.

(b) By looking at population-wide (total) infection rates, the infection rates of code-subscribing ISPs

get comingled with the infection rates of non-subscribing ISPs. Given that comingling, an uptick

in bots among non-subscribing ISP users might offset any improvement in the number of bots

seen in subscribing ISPs' user populations.

It isn't surprising that researchers aren't splitting out their data according to subscribing ISPs vs.

non-subscribing ISPs since it is currently quite difficult to operationally tell what ISPs are   
 "in" and what ISPs are "out." While a number of ISPs have self-asserted that they are

participating,[[3]](#footnote-3) those self assertions (made just by company name), are less useful than a list of

specific ASNs,[[4]](#footnote-4) specific IP netblocks,[[5]](#footnote-5) or specific in-addrs[[6]](#footnote-6) associated with code-subscribers.

(c) That study looked at the infection rate for "North American"[[7]](#footnote-7) households, while the ABCs for

ISPs code only targets U.S. ISPs and their customers.

(d) That study looked at infection rates for *households*, rather than *computers*. There might be a half

dozen computers in a household, but if even one is infected, the entire household will get flagged

as bad. This can skew the proportion of a population that ends up getting reported as infected.[[8]](#footnote-8)

(e) On the other hand, what about infected devices other than just desktops or laptops? For example,

what about smart phones and tablets? Are we also counting infections on those devices? What

about other sorts of devices, such as "smart TVs" or Internet-connected gaming consoles?

(f) Not all broadband customers (nor all infected broadband customers) are "households."   
 For example, another important broadband customer segment might be small and medium-sized

businesses and similar organizations (such as broadband-connected primary and secondary

schools). Are infections in those populations "in scope" or "out of scope?"

(g) What if a botted computer is offline, and thus not "showing signs of an infection" (to use the

language from the article). Does that/should that "infected but offline" system still "count"?

(h) What constitutes a "meaningful" or "material" change for the better (or worse)? Is there some

level that we may eventually reach (even if it isn't zero) that we can all agree is "good enough?"

The answers to those questions largely shape the botnet metrics space, and those choices largely determine the answer that one ultimately finds.

We need to address these issues if we're to be able to provide meaningful metrics about the state of bots in the United States, and if we're to be able to measure the potential impact of the ABCs for ISPs code.

Let's begin with the issue of what is or isn't a bot.

**II. Thinking *Precisely* About What Is and Isn't A Bot**

What Exactly *Is* a Bot? In an earlier report,[[9]](#footnote-9) Working Group 7 provided a general definition of "what's a bot," stating:

*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*

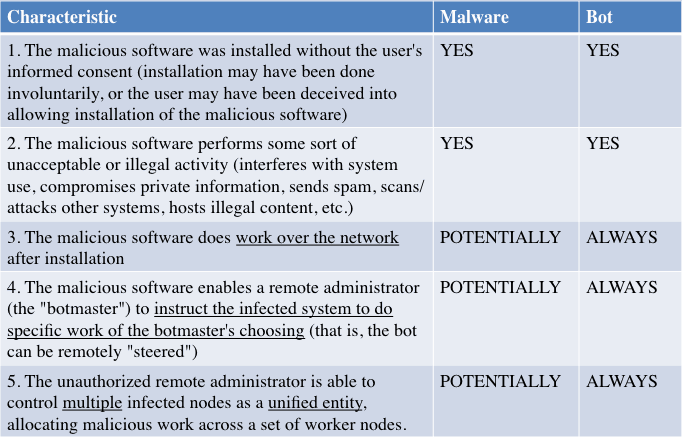
*send unwanted electronic email ("spam"), to reconnoiter or attack other systems, to eavesdrop upon*

*network traffic, or to host illegal content such as pirated software, child exploitation materials, etc.*

While that's a fine definition as far as it goes, it may not sufficiently emphasize one critically important point:

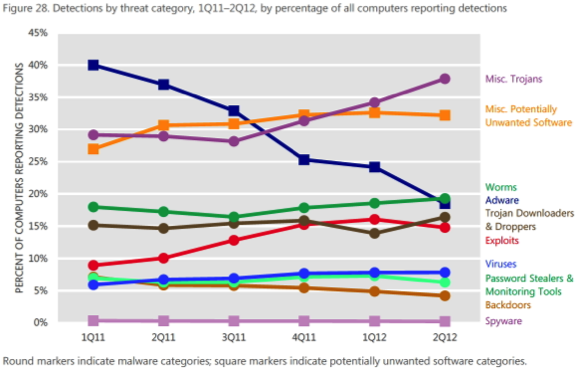
**Not all malware is bot malware.**

Characteristics that can be used to differentiate malware in general from bot malware in particular include:



An illustrative/non-exhaustive list of current and historical malware families that are generally agreed to be "bot malware" can be found in Appendix A to this report.

5. The extent to which the industry often fails to identify what malware is or isn't bot malware can be seen in this graph from the Microsoft Security Intelligence Report,[[10]](#footnote-10) which breaks out ten different types of malware, but makes no mention of what is or isn't a bot:



A Botnet Malware Registry? To help eliminate ambiguity over what is and isn't a bot, one option would be for the industry to create a voluntary botnet malware registry. An excellent foundation for a registry of this sort might be the site http://botnets.fr/ which currently catalogs over 300 botnet families by name.[[11]](#footnote-11)

Once an agreed upon bot registry is available, whether that's botnets.fr or something else, malware that has been found to be "bot" malware could then be listed in that registry.

While this might sound like a small step, it actually enables significant bot-related research. For instance, anti-malware vendors, when analyzing and cataloging malware they detect, could then potentially voluntarily add an "is this malware a bot?" attribute to their malware catalog entries (based on the registry), 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 hypothetically report on:

(a) The number of new bot malware families discovered that quarter,

(b) The percent of systems seen infected with each of the dozen most significant bots, and

(c) The total number of hosts detected as infected with one or more bots.

Having a common botnet definition would allow multiple reports of that sort to be compared: do all anti-malware vendors see the same number of new bot malware families? Do they see approximately comparable new levels of infection? Until we agree on what is and isn't a bot, it is impossible to tell if apparent differences are due to bot definitional differences, or other differences (such as a different customer base, differing detection efficacy, etc.)

"If It *Acts* Like a Botnet:" In some cases (for example, in the routine case of an ISP that does not have direct 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,[[12]](#footnote-12) or spewing spam, or contributing to a DDoS[[13]](#footnote-13) attack), even if a particular bot cannot be identified, the system should still get tagged as being botted.

Tagging botted systems based on their externally observable behavior may be necessary when direct access to the system isn't possible, but also in cases where systems are infected with malware that's so new that antivirus companies haven't yet had time to identify that malware.

Therefore:

**If a system acts like it is infected by a bot, even if it cannot be identified as infected by**

**a particular type of bot malware, tag it as botted.**

**III. What Sort of Botted "Things" Should We Be Trying to Count?**

Now that we've agreed on what is and isn't a bot, we're a large 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 deciding precisely what sort of botted "things" we're going to count.

*What* I'm *Able* to Measure Will Depend on My Role In the Ecosystem:

(a) If someone 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, those researchers would have the option to

count individual infections, or botted systems, or botted IP addresses, or botted users, or botted

households. While going "boots on the ground" might seem to provide the most flexibility and

most comprehensive data collection options, it is also the most potentially expensive option, and

it presumes access to a household's systems, access that might be viewed as intrusive and

routinely denied.

(b) On the other hand, if I'm an ISP, and I detect bots based on malicious network activity associated

with a particular IP address, I'm likely going to count botted IP addresses or botted subscribers.

(c) If I'm an antivirus company or an operating system vendor and I scan/clean end-user systems, I'm

going to end up counting individual infections,[[14]](#footnote-14) or perhaps botted systems.[[15]](#footnote-15)

(d) If I'm a survey research outfit, and I call people up on the phone and ask, "Have you ever been

infected with a 'bot'?" those survey researchers are going to end up counting botted users.[[16]](#footnote-16)

Different parties will contribute different views of the problem. While those views may be different, all are potentially valuable and important.

Desktops and Laptops Only? Are we going to measure all kinds of botted devices, or are we just going to count botted desktops and laptops? For example, consider smart phones in particular. The number of smart phones is now material, and malware is increasingly attacking and infecting at least some types of those devices.[[17]](#footnote-17) Users also have a growing number of tablets, Internet-connected "smart TVs" and set-top boxes, gaming consoles, and other devices that may be targeted for compromise. Should all those sorts of devices be counted, if botted? We think so, yes. To ensure a comprehensive botnet "picture," we suggest that any program of botnet metrics should include ALL types of Internet-connected devices, but, as that data is collected, it should include the type of device involved, thereby allowing analysts the option of reporting about all devices, or just some particular subset of all devices, such as just traditional laptops and desktops, or just smart phones.

*Online* Devices Only? We must recognize that in most cases we can only count botted systems that are "live" or "that we can see" on the network. Other devices may be botted, but remain undetected/unknown as a result of how we detect botted hosts.

An easy way to understand this is to imagine that we're an enterprise that actively scans its corporate systems with Nessus[[18]](#footnote-18) or a similar security 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 network scan takes place, a potentially infected system won't be able to be found with that tool.[[19]](#footnote-19)

Similarly, if an ISP identifies botted hosts based on spam (or other network artifacts visible to the ISP "on the wire"), a botted host that's offline or in a walled garden won't be seen "making noise" or "causing problems" on the ISP's backbone[[20]](#footnote-20) where instrumentation exists, and also won't/can't be noted as being botted by external parties relying on externally visible symptoms to tag systems as being botted.[[21]](#footnote-21)

Measurement Window-Related Decisions Will Have A Material Impact on Bot Detection Rates: If a botted system is only online occasionally, *when* and *how long* we look for bots (e.g., our "measurement window") may strongly impact how many bots we find.

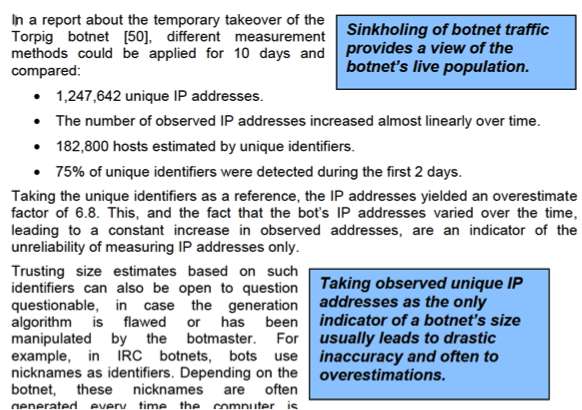
For instance, if an ISP is detecting bots based on characteristic botnet network activity, when/how long the ISP collects bot-related network flow records will strongly influence botnet detection rates. To understand why, note that if we only watch for botted hosts during a brief window during the business day, we might miss home systems that are turned off except when a family member is using them at night. On the other hand, if we collect bot data during evening "prime time" hours, we will likely miss any botted work systems that may only be on and in use during the normal 8 to 5 work day. Therefore, if you try to count botted hosts during too brief a time period, you may miss some bots.

If we go to the other extreme, and *continually* watch for botted hosts, we will virtually certainly see the same botted host more than once, and since we can't tell one bot apart from another, we run the risk of counting a single bot more than once, simply because in most cases there's no unique identifier that we can use to track a particular botted host from one sighting to the next time we see it.

Unique Identifiers: If each botted host *did* have a unique identifier, we could collect data over a protracted period and not have to worry about counting the same botted host multiple times. Unique identifiers for botted hosts would also greatly simplify the process of aggregating (or "rolling up") fine grained records appropriately: for example, we could tag a record about each individual infection with the unique identifier associated with that botted host, and then we could easily consolidate that data if/when we wanted to do so.

Unfortunately, if we don't or can't use unique identifiers, our measurements may end up profoundly flawed.

Consider one example mentioned in the ENISA report on *Botnets: Detection, Measurement, Disinfection & Defence:[[22]](#footnote-22)*



Clearly, simple IP addresses are far from "unique identifiers!"

One Reason Why IP Addresses Are Not Unique Identifiers: DHCP-Related Issues: While non-technical Internet users might assume that the IP address their computer uses is constant (or "static"), in most cases that will not be true. Most users will receive a dynamic IP address via DHCP,[[23]](#footnote-23) and that address may be reassigned when the DHCP lease expires.

This means that a given botted host might have a succession of different IP addresses over time, making it appear that there are *more* botted hosts than are actually the case.

It also means that potentially two or more different botted hosts may use the same IP address in succession, making it appear as if there are *fewer* botted hosts than are actually the case.

While ISPs normally can and do identify which customer is using a particular dynamic IP address at a particular time if/when they need to do so,[[24]](#footnote-24) that process often is somewhat cumbersome and scales poorly as a method to be used in conjunction with thousands of botted hosts.

Another Reason Why IP Addresses Are Not Unique -- Network Address Translation (NAT): Another potential complication when it comes to mapping bots to IP addresses is the use of NAT. NAT is a public IP address-conserving technology that allows multiple systems to share a single public IP address. Because multiple systems share a single public IP address, malicious traffic from multiple systems may appear to come from just one IP address, resulting in an underestimate of the number of truly infected systems.

IPv6 Addresses: As the Internet runs out of traditional IPv4 network addresses, ISPs are beginning to use IPv6 addresses to supplement rapidly depleting stocks of IPv4 addresses. IPv6 significantly complicates the process of measuring botnets via their network traffic. Let's just mention a few of many reasons why this is true:

(a) Many ISPs may not have IPv6 network flow monitoring that's on par with their IPv4 monitoring

capabilities. These IPv6 network flow monitoring deficiencies may leave many ISPs fully (or at

least partially) "blind" when it comes to monitoring IPv6 network activity, including IPv6 botnet-

related network activity.

(b) In other cases (as when ISP customers obtain tunneled IPv6 connectivity from a third party

provider), the customer's "home" ISP may have little or no visibility into the content of IPv6

tunneled traffic. Any third party network researchers taking network measurements would see

that customer's IPv6 traffic (including any IPv6 *bot* traffic) emerge from the tunnel provider's

network infrastructure, not the home ISP's network infrastructure. This would obviously

complicate any global botnet measurement work involving IPv6 connectivity.

(c) IPv6 network address assignment technologies also play a role in potentially simplifying -- but

more likely complicating -- botnet measurement efforts. To understand how this can be true,

understand that in IPv6, IPv6 addresses may be assigned in multiple different ways (including,

but not limited to):

(i) *Static IPv6 address:* If a botted host uses a static IPv6 address, it is easy to track it over

time, just as it is easy to track a botted host that's using a static IPv4 address. Static IPv6

addresses are expected to be rare, however, except for manually configured servers.

(ii) *SLAAC*: Hosts may have an automatically assigned[[25]](#footnote-25) IPv6 address that leverages a

reformatted version of the system's hardware MAC address. When systems use addresses

of that sort, it becomes possible to potentially track a botted host over time and even

across multiple ISPs. This is an example of how IPv6 addresses can simplify botnet

measurement activities.

(iii) *Privacy Addresses:* IPv6 hosts can also use constantly changing IPv6 "privacy

addresses."[[26]](#footnote-26) Hosts using IPv6 privacy addresses periodically and automatically

change their IPv6 addresses in an effort to make it hard for those systems to be

systematically tracked by marketers or others. If a botted host is using IPv6 privacy

addressing, it becomes very difficult for researchers to accurately follow the IPv6

addresses that that system may be using over time, and as a result, we may potentially

dramatically over-estimate the actual number of IPv6-connected bots.

(iv) *DHCPv6:* In a fourth scenario, IPv6 using customers may receive IPv6 addresses via

DHCPv6, the IPv6 analog of DHCP for IPv4. When DHCPv6 is used, ISPs can map a

particular IPv6 address to a particular customer, but doing so remains somewhat

awkward, and won't typically be practicable for large numbers of botted hosts.

Counting Infections vs. Counting Botted Systems: Let's set aside IPv6 issues for now, and just 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 a botnet analyst generate one measurement for System A, and seven measurements, one for each bot infection on System B, thereby allowing each individual type of bot infection to be separately tracked?

This may be particularly relevant on large shared systems, such as high density web hosting sites, or "timesharing" Unix boxes with thousands of shell accounts, all sharing a single IP address. In that case, a single system might have multiple independent customers, and multiple independent bot infections (including potentially multiple copies of the same bot!), all running in parallel.

Not All Botted Hosts Are Equally "Potent:" To understand what's meant by this consider two hypothetical botted hosts:  
  
 *System C:* An ancient consumer system connected via a legacy 56Kbps dialup connection

*System D:* High end server with multiple CPUs/multiple cores, lots of RAM, and gigabit

ethernet connectivity

Should each of those systems simply be counted as one "botted host"? There would certainly be a huge difference when it comes to the amount of spam or the volume of DDoS traffic that each of those two systems might respectively deliver... Does this meant that we should weighting botnet detections by some measure of capacity (such as their average spam throughput, or their average DDoS output)?

Some Apparent "Bot" "Hits" May Not Be Real: 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. 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 "real" bots that actually exist.

What One Nascent Metrics Program Chose to Count: The recently estsablished MAAWG malware metrics program focuses on "subscribers" as the unit of analysis.

Because the number of subscribers will fluctuate over the course of a typical month, MAAWG decided to use the number of subscribers as of the last day of the month.

Participant ISPs then report 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)

Participant ISPs will also report the number of unique subscribers that have been notified of a problem by whatever means (SMS, phone call, email, web redirection/browser notification etc.) Multiple notices to the same subscriber count as one. This does not imply that the subscriber has read/received the notice.

Given those values, one can compute the percentage of subscribers that have been found to be infected one or more times during a given month, and the percentage that have been notified of that infection.

As a metric, note that this metric implicitly accepts some compromises, e.g.:

(a) Given the definition of this metric, we can't talk about how many infected customer systems may

be present, nor how many distinct infections were seen, nor can we talk about whether a

particular customer was repeatedly reinfected, or if the infection was on a laptop, smart phone,

gaming console, etc.

(b) The MAAWG program doesn't focus solely on bots, since many ISPs want to protect their

customers from all types of serious malware infections, not just bots.

(c) Choice of a month-long window means that day-to-day or week-to-week infection trends won't be

able to be identified.

(d) There are many other potential measurements related to infected customers that aren't getting

reported (for example, how much customer effort was required to disinfect and harden a

typical infected system?)

These and other limitations were explicitly recognized and accepted by MAAWG as part of its pragmatic program design decisions, recognizing that if it made the malware metrics reporting program too difficult or too time-consuming or too complex, many ISPs would simply opt out of participating. Keeping the program simple and easy to participate in increases the number of participating ISPs.

Another example of a pragmatic measurement choice was the decision to focus on the number of unique customer detections and customer notifications rather than the number of customer systems that have been cleaned up or rebuilt. (Because customers may use third party services to clean up or rebuild their systems, ISPs may not know if a customer's system has been cleaned up, rebuilt, replaced, or remains infected (but is offline/dormant)).

**IV. Some Substantive Questions About Bots**

The MAAWG example just mentioned will yield some botnet related metrics. However, what are the other substantive questions about bots that we might like to be able to answer?

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 all 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 that 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?

One example of such an estimate can be seen in Gunter Ollmann's "Household Botnet Infections,"[[27]](#footnote-27)

*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.*

There are 81.6 million US households with broadband connectivity as of 10/2010.[[28]](#footnote-28)
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.

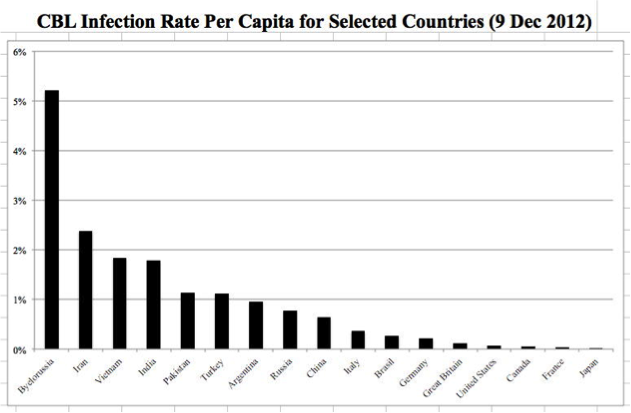
Let's consider another estimate, from the Composite Block List ("CBL"). On Sunday December 9th, 2012, the Composite Block List knew about 174,391 botted host IPs in the United States.[[29]](#footnote-29)

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 the CBL. On a per capita-normalized basis, that means that the US is among the least botted of all countries as measured by the CBL.[[30]](#footnote-30)

See the graph at the top of the following page.



In fact, if there are only 175,000 bots here in the U.S., botted hosts have effectively become a "rare disease" in that when it comes to traditional medicine, the U.S. definition for a rare disease is one which afflicts fewer than 200,000 people in the United States.[[31]](#footnote-31)

How Could Those Two Estimates Be So Vastly Different? We think that there are two main reasons for that vast discrepancy in those estimates:

(a) The two estimates count different sorts of things (households that are detected as being botted vs.

botted IPs seen sending spam)

(b) The two estimates measure different populations (users worldwide who are connected via ISPs

with enough of a bot problem that those ISPs have been motivated to purchase a commercial

network security solution vs. users in the United States (where a real push to control bots has

been underway for years)

Another Very Basic Question: How Many Different 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...

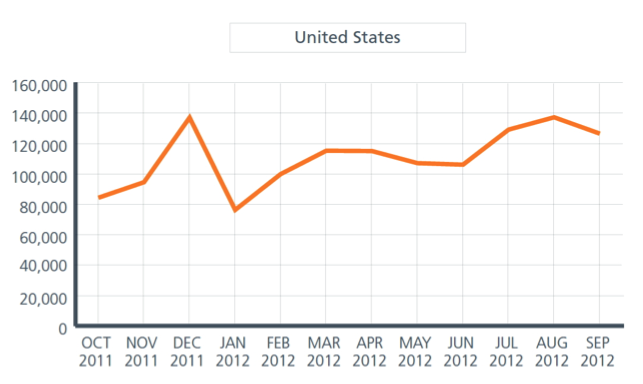
The number of unique types of bots is important because it tells us a lot about how hard it might be to get the "bot problem" under control. If there are only a handful of major bots, concerted effort should allow government authorities to shut them down, if the government makes doing so a priority. Conversely, if there are three thousand different types of bots out there, getting all those bots under control would be far harder.1

Closely related to the question of how many types of bots are out there, how many botmasters 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, eroding a direct one-to-one relationship between the two metrics.

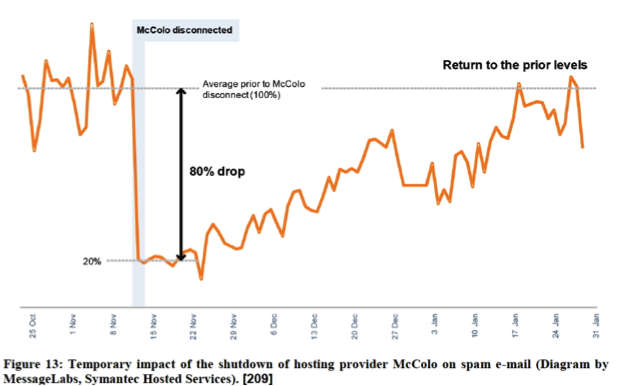
How Many Users Are Covered by the ABCs for ISPs Code? While some organizations have attempted to identify the number of users covered by the ABCs for ISPs Code, it can be hard to dig out subscriber estimates for participating ISPs. Should one of our "metrics" simply be a clean report of how many subscribers are covered by the Code?

Are There Any Trends Relating to Bots? That is, in general, is the bot problem getting better or worse over time?

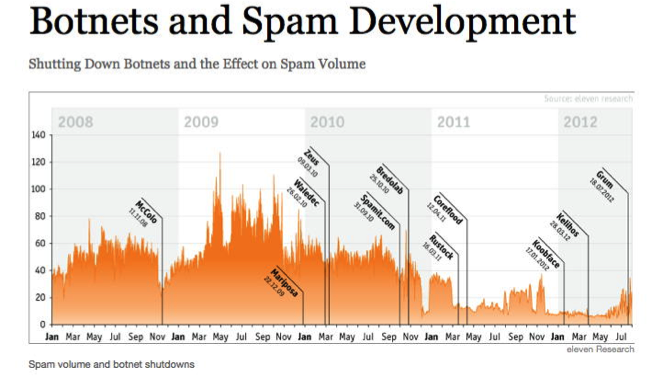
Some anti-malware companies are already sharing data of this sort, at least for some types of bots. See for example the following graph from McAfee for the United States:[[32]](#footnote-32)



Do Bots Show Any Sort of Operational Patterns? 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) 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 and have money to buy spamvertised products, or does it peak in the month before Christmas (when people are most likely to be Christmas shopping), perhaps? If law enforcement arrests a botmaster or 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 bot ecosystem? Here's one interesting graph of that last sort...[[33]](#footnote-33)



Another example of an interesting long term botnet graph:[[34]](#footnote-34)



HOW Are Botnets Being Used? While bots can be rapidly reconfigured from one purpose to another, do we have a clear understanding of how bots are currently being used, or how they've been used over time? That is, what fraction of all botnet capacity is used for:

(a) Sending spam? (and how many spams get emitted from each individual botted host? are those

bots running "flat out," or just "loafing along?")

(b) Participating in DDoS attacks?

(c) Scanning network-connected hosts for remotely exploitable vulnerabilities?

(d) Hosting illegal files?

(e) Stealing private information?

(f) Cracking passwords, mining BitCoins or other compute-intensive tasks?

(g) Are there bots that are installed but totally idle? If so, why? Excess capacity?

Understanding how bots are being used will help us to figure out we how we should try to measure bots.

For example, if bots are not longer being widely used to send email spam, we shouldn't attempt to measure botnet populations based on the amount of email spam we observe, right?

HOW Bots Are Being Used May Change Who's Interested in Them: Hypothetically, if bots are no longer in widespread use for spamming, anti-spammers may lose interest in bots. On the other hand, if bots start to be widely used to conduct distributed denial of service attacks against critical government sites, that change might increase interest in bots in the homeland security and national security communities.

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

"Comparatively Speaking..." Another set of potentially interesting botnet metrics are comparative metrics:

(a) Are American computers getting botted more (or less) than Canadian computers, or computers in

Great Britain, France, Germany, Japan, Russia, China, Brazil, India or \_\_\_\_\_\_\_?

(b) 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?)

(c) 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

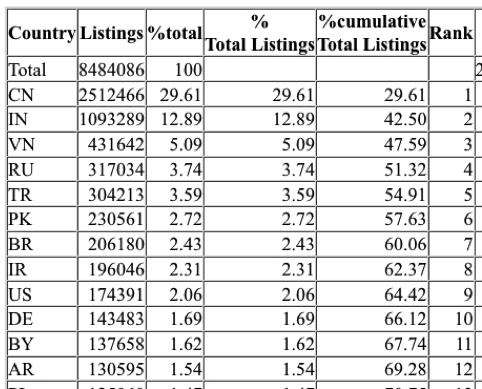
fewer bots than other ISPs that don't adopt it?

(d) Are there other important 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

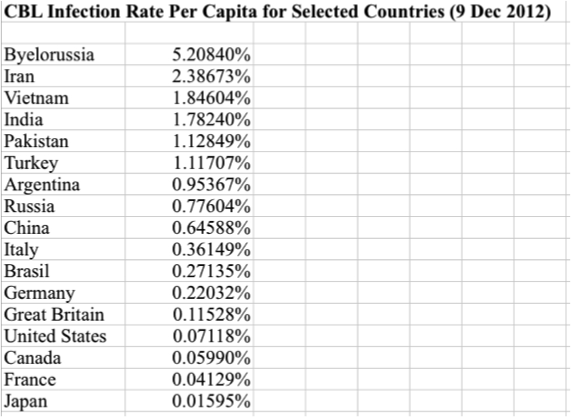
or web browser or email client or operating system people use?

Comparative Raw Bot Levels Per Country From the CBL:[[35]](#footnote-35)



China looks pretty bad in that list, but then again, remember, that China's a big country. How do they look, comparatively, once we adjust for their population?

Selected CBL Listings By Country, *Normalized Per Capita:*



Once we've normalized per capita, China is no longer leading the list (that dubious "honor" now goes to Byelorussia), but China is still fully an order of magnitude more botted than the United States is, even if China is fully an order of magnitude less botted than Byelorussia is.

Pre/Post Longitudinal 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 track 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 as the number of bots drop over time.

In practice, it may be difficult to do a study of that sort since 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.

Drilling Down on a Per-Bot-Family Basis: 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:  
  
 (a) the amount of spam attributed to each particular spam botnet

(b) the volume of DDoS traffic attributed to each DDoS botnet

(c) the number of command and control hosts that a bot uses

(d) the geospatial distribution of hosts infected with each bot

Micro as Well as Macroscopic Measurements: Not all metrics are macroscopic measurements related to botnet infection rates. Some measurements of interest might be per-system micro values:

(a) What does it cost to rent a bot on the open market?

(b) How long does it take/what does it cost to de-bot a single botted host? What factors make a

system take more or less time to de-bot? Can we build a standardized cost model? For example,

what's it worth to have a clean backup of a botted system? Does that make it significantly easier

to get a botted system cleaned up and hardened?

(c) 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, so that an anti-malware technician can then be told, "If you find bot A on a

system, also be on the lookout for bot B, too?"

(d) 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

presumably unpleasant experience? Or are some types of users just inherently more prone to get

themselves reinfected, perhaps because of a failure to apply available patches, or inherently risky

online activity patterns? If users do end up getting rebotted, what's the typical time till

reinfection?

Looking at This From A Different Direction: How Long Will A Typical Bot Live? 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 that 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.[[36]](#footnote-36) 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? If third parties don't target botnet backend infrastructure, can the Internet community be confident that law enforcement will in fact track and take down those botnet-critical resources? Are there ways that we can deconflict this work without compromising operational security?

**V. Some Statistical Questions Associated With Botnet Measurements**

How *Precise* Do Our Answers to These Questions Need To Be? "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%. Exactly how precise do our measurements need to be, and why?

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 can't have all three at the same time.

*Budget:* We really need to emphasize that if bespoke hard numerical answers to questions about botnets are needed, **it's going to cost money to obtain those values**. How much are we willing to spend to get those answers?

If the answer is "zero," then I would suggest that in fact all our substantive questions about bots 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 rationally set the required level of precision (and the required confidence in our estimates) is also going to be impaired.

**VI. ISPs As A Potential Source of Botnet Data**

The CSRIC 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:

(a) 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?

(b) There are many ISPs in the United States. There are even more ISPs in other countries. Many of

these ISPs will not participate. Incomplete participation (even simple addition and subtraction of

participating ISPs) will complicate data interpretation and analysis.

(c) ISPs may also be reluctant to share data about 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?

(d) 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)

(e) Different ISPs may measure botted customers differently (efforts at standardization

notwithstanding), undermining the comparability of inter-ISP botnet metrics.

(f) Self-reported and unaudited data bot may be subject to error or manipulation, or at least the

perception by cynics that it might not be fully candid/fully accurate/fully trustworthy.

(g) 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, if you want consistent data with a global scope.

Entities With A Global Footprint: There are entities *other than ISPs* with consistent global visibility into the bot status of Internet systems and users:  
  
 (a) 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)  
  
 (b) 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 data that's already been published?**

If not, why not? We believe that the FCC should NOT be "reinventing the wheel," particularly if there's no funding that can be used to ensure that botnet-related data is collected carefully and consistently. A list of some existing cyber security reports can be found in Appendix B to this report.

How Might the FCC Encourage ISPs to Voluntarily Submit Botnet Metrics? Hypothetically, assume that no ISPs voluntarily submit metrics on their botnet experiences to the FCC.[[37]](#footnote-37)

In that case, having no other option (short of mandating reporting, which would likely be resisted by ISPs and others), 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 -- ***if*** those companies could be assured that having shared botnet data, they'd be safe from threats of lawsuits, involuntary public disclosure of shared data, or eventual compulsory reporting.

Sharing: Where's the Mutuality/Reciprocity? If we view ISPs and the government as partners that share a common interest in tackling botnets and improving cyber security, and if we believe that both parties are committed to collaborative data driven security work, it would be terrific if operational data sharing was 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![[38]](#footnote-38)

**VII. Sinkholing, DNS-Based Methods, Direct Data Collection 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 as a result of that sinkhole to measure a particular 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. Many modern botnets now also take special care to prevent or deter efforts at sinkholing.

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 a "magic" known-bad domain name, there's a substantial likelihood that that customer is botted.[[39]](#footnote-39)

Big botnets might tend to make more extensive use of DNS than smaller and less sophisticated botnets, but caching and other subtleties associated with DNS can complicate DNS-based measurements, and of course, not all bots even use DNS.[[40]](#footnote-40)

Some might also be tempted to try an RPZ-like approach (as implemented in BIND) to "take back" DNS and prevent bots from using DNS as part of their infrastructure. While this approach certainly has technical promise, any effort to instantiate policy via DNS is a potentially tricky one,[[41]](#footnote-41) and should only be considered if customers can opt-out of a filtered DNS view should they want or need to do so.

Directly Checking Systems 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 sources. That is, you want to go out and collect your own data, much as survey research groups survey entities about political viewpoints or consumer spending. 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 be able to individually track. If you don't have "hints" about who may be a botted user a priori, and you just need to discover them at random, you may be facing a daunting task, at least 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 what if just 0.0711% of all users are botted (recall that this was the CBL-reported rate for the US on December 9th, 2012)? On average you'd see just 0.711 botted users per thousand. To get 350 botted users to study, you'd likely need to check (350/.711)\*1,000=492,264 users. That's a LOT of data to collect.

Now assume that you want 350 users PER ISP, and assume you're interested in a dozen ISPs... 12\*492,264 users= 5,907,168 users. That's REALLY going to be a lot of work!

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 (a very 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 492,264 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 -- and that's just for one ISP, at one point in time.

Resistance to Government Scanning of Personal Computers: We 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/common 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 actually be infected.

Volunteers? Some users who think that their systems might be infected might welcome the opportunity to have their systems scanned. Unfortunately, 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.

Simulating 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." While conceptually intriguing, this might not be easy. For example, the fidelity of the results from such a simulation will depend directly on researchers ability to:  
  
 (a) 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?)

(b) Replicate the range of botnet malware seen on the Internet (constantly changing)

(c) Accurately model the ISP response to the malware threat  
  
Given these difficulties, we believe this is a fundamentally impractical approach.

**VIII. Recommendations**

The CSRIC Working Group 7 recognizes that summary metrics with which to determine the effectiveness of the U.S. Anti-Bot Code of Conduct are not yet available.

The working group recommends that the following course be undertaken in order to enable such metrics for the future:

• We recommend that specific Case Studies be supported to gain metrics around particular bot efforts. Summary metrics which may address and validate the overall effectiveness of the code can only be developed based upon component metrics derived from more specific efforts in combating specific bots. Further collaborative efforts will be required to arrive at a foundation of metrics. These efforts will need to involve not only the ISP community but the larger ecosystem as well.

• We further recommend that that specific Pilot Programs be supported to gain metrics around particular bot efforts. Such programs are required to test which metrics may be useful and which are not. For comparative purposes, the metrics definitions will have to be reasonably standardized between ISPs. We anticipate that some metrics methods used by ISPs will lend themselves to comparative analysis and some will not. Participation in pilot programs will indicate which are viable.

• We anticipate that the ongoing Georgia Tech sponsored study (Professor Wenke Lee, Director, Georgia Tech Information Security Center) be considered as an initial case study. This effort centers on the DNSchanger bot and related customer notification methods used by ISPs. This approach, although preliminary, may well show the cooperative, collaborative, steps required for the future development of overarching metrics involving multiple ISP approaches. These steps are a micro chasm of the recommendations in the code. Additionally, the test results may provide insight into the relative efficacy of different notification methods and insight into resulting best practice methods for notification of customers.

• It is recommended that the FCC recommend voluntary methods for standardization of metrics for the purpose of comparative analysis of methods and best practice development. Such voluntary methods can be applied ultimately toward education, detection, notification, and remediation of Bots as well as to the collaborative efforts required by the broadband ecosystem at large.

The participants in CSRIC III Working Group 7 would be happy to address any remaining questions that CSRIC or the FCC might have about this proposed program of work.

**Appendix A.**

**Some Examples of Current or Historical Malware Families   
That Might Properly Be Considered to Be "Bots"**

Agobot/Phatbot[[42]](#footnote-42)

Bagle/Beagle[[43]](#footnote-43)

Bamital[[44]](#footnote-44)

Bredolab/Oficla[[45]](#footnote-45)

Citadel[[46]](#footnote-46)

Conficker[[47]](#footnote-47)

Coreflood[[48]](#footnote-48)

Cutwail/Pushdo/Pandex[[49]](#footnote-49)

DarkComet[[50]](#footnote-50)

Dirt Jumper/Russkill[[51]](#footnote-51)

Donbot[[52]](#footnote-52)

Festi/Spamnost[[53]](#footnote-53)

Flashback[[54]](#footnote-54)

Grum/Tedroo[[55]](#footnote-55)

Kelihos/Hlux[[56]](#footnote-56)

Koobface[[57]](#footnote-57)

Kraken/Bobax[[58]](#footnote-58)

Lethic[[59]](#footnote-59)

Maazben[[60]](#footnote-60)

Mariposa[[61]](#footnote-61)

Mega-D/Ozdok[[62]](#footnote-62)

Nitol[[63]](#footnote-63)

Ogee[[64]](#footnote-64)

Pikspam[[65]](#footnote-65)

Pushbot/Palevo[[66]](#footnote-66)

Ramnit[[67]](#footnote-67)

Rustock[[68]](#footnote-68)

Sality[[69]](#footnote-69)

SDBot[[70]](#footnote-70)

Spybot[[71]](#footnote-71)

SpyEye[[72]](#footnote-72)

Srizbi[[73]](#footnote-73)

Storm[[74]](#footnote-74)

TDSS/TDL-4[[75]](#footnote-75)

Waledac[[76]](#footnote-76)

Xpaj[[77]](#footnote-77)

ZeroAccess[[78]](#footnote-78)

Zeus/Zbot[[79]](#footnote-79)

**Appendix B.**

**Examples of Data Driven Cybersecurity Reports**

1. Composite Block List Statistics

http://cbl.abuseat.org/statistics.html

2. Kaspersky Security Bulletin/IT Threat Evolution

http://www.securelist.com/en/analysis/204792250/IT\_Threat\_Evolution\_Q3\_2012

3. McAfee's Quarterly Threats Report

http://www.mcafee.com/apps/view-all/publications.aspx?tf=mcafee\_labs&sz=10&region=us

4. Microsoft's Security Intelligence Report (SIR)

http://www.microsoft.com/security/sir/

5. Shadowserver Bot Counts

http://www.shadowserver.org/wiki/pmwiki.php/Stats/BotCounts

6. Symantec's Internet Security Threat Report (ISTR)

http://www.symantec.com/about/news/resources/press\_kits/detail.jsp?pkid=threat\_report\_17

**Appendix C.**

**Another Data Collection Alternative, If Botnets Are A**

**National Security Threat And Not Merely a Nuisance**

While botnets are often thought of purely as a nuisance, e.g., as 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 botnets, including per-ISP estimates.

Such information, once collected, might then be able to be shared with appropriately cleared government officials with 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 partner agencies.

1. http://transition.fcc.gov/pshs/advisory/csric3/wg-descriptions.pdf at PDF page 7. [↑](#footnote-ref-1)
2. "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 [↑](#footnote-ref-2)
3. http://www.maawg.org/abcs-for-ISP-code [↑](#footnote-ref-3)
4. ASNs, or Autonomous System Numbers, are a convenient way of referring to a particular ISP, or perhaps part of an ISP. For example, Google uses AS15169, Sprint uses AS1239, Intel uses AS4983, the University of California at Berkeley uses AS25 and so on. For more on ASNs, see http://pages.uoregon.edu/joe/one-pager-asn.pdf [↑](#footnote-ref-4)
5. ISPs and other entities receive "blocks" or "ranges" of IP addresses for their use. For example, the University of Oregon has 128.223.0.0/16 (the IP addresses 128.223.0.0 through 128.223.255.255) for its use among other netblocks. These network blocks represent another way of referring to an entity online, albeit one that is less convenient than ASNs because a large ISP may have accumulated hundreds of netblocks over time as their requirements evolved, or as a result of mergers and acquisitions with other ISPs. Ownership of IP netblocks is documented in a distributed online database known as whois. Unfortunately whois servers often rate limit (or otherwise control) the number of queries that a given user can make against it during any given period of time, making it frustrating to use in conjunction with large datasets. [↑](#footnote-ref-5)
6. Inverse addresses (also known as "in-addrs" or "PTRs"), are the domain names that get returned when you lookup an IP address. They can provide hints about who controls a given IP address (although they are often not present, and can be subject to spoofing). [↑](#footnote-ref-6)
7. There's often a tendency to treat "North America" as if it is just comprised of the United States and Canada (with everything else in the Western Hemisphere being part of Latin/South America and the Caribbean), but in fact there are actually 29 countries that are serviced by ARIN, the "North American" Internet number allocation organization. If a researcher determines what's a "North American" household by checking to see if the IP address associated with each infection came from ARIN (rather than some other entity, such as RIPE, APNIC, LACNIC or AFRINIC), the anti-botnet efforts of U.S. ISPs will be potentially conflated with the botnet experiences of 28 other countries or territories. That means that even if U.S. botnet numbers improved, domestic improvements (if any) may end up marginalized or eliminated by a hypothetical worsening of Canadian/Mexican/Caribbean/other "North American" countries bot numbers. [↑](#footnote-ref-7)
8. To understand this distinction, imagine a hypothetical media report about the impact of the flu on 150 area businesses employing 30,000 people. If each business had exactly one employee sick with the flu (a total of 150 sick people among all area businesses), we could either report that *"100% of businesses had been hit by the flu"* (since each business does in fact have exactly one employee sick with the flu), or that *"just 1/2 of one percent of all employees have the flu"* (e.g., since 150/30,000\*100=0.5%). These two different metrics convey radically different stories about the hypothetical flu problem in area businesses, right? [↑](#footnote-ref-8)
9. http://www.maawg.org/system/files/20120322%20WG7%20Final%20Report%20for%20CSRIC%20III\_5.pdf [↑](#footnote-ref-9)
10. Microsoft Security Intelligence Report, Volume 13, http://www.microsoft.com/security/sir/ [↑](#footnote-ref-10)
11. While the primary language of that site is French, content is also available in English via the selector in the left hand column. [↑](#footnote-ref-11)
12. A "command and control host" is a system that a botmaster uses to run his botnet. [↑](#footnote-ref-12)
13. A DDoS attack is a Distributed Denial of Service attack, often conducted by flooding a site with so much bogus traffic that the site's network connection or servers can't keep up, thereby preventing legitimate users from being able to use that site. [↑](#footnote-ref-13)
14. A single system might have multiple simultaneous infections. [↑](#footnote-ref-14)
15. One user might have two or three systems, or one system might be shared by multiple users. [↑](#footnote-ref-15)
16. If we talk to three different people from the same household, and they all used the same botted computer, we might potentially get three reports that are all about that single botted system.

    This scenario also runs afoul of multiple other issues, including things as basic as the fact that users may not know what a bot is, or they may forget having been botted. [↑](#footnote-ref-16)
17. Virtually All New Mobile Malware is Aimed at Android,

    http://www.androidauthority.com/mobile-malware-aimed-android-112403/ [↑](#footnote-ref-17)
18. http://en.wikipedia.org/wiki/Nessus\_%28software%29 [↑](#footnote-ref-18)
19. This raises an interesting methodological question: if we scan and can't reach a potentially botted host, how often should we attempt to rescan it? Once? Twice? Time after time after time? Never? [↑](#footnote-ref-19)
20. A related potentially important methodological question (at least from the point of view of ISPs actively mitigating botted hosts): if a botted system has been detected as being botted and successfully put into a so-called "walled garden" where it can't cause problems for other Internet users, should that host still be counted as "botted"? Or do we need an additional category to capture systems in this status, "botted but not online and able to cause problems," perhaps? [↑](#footnote-ref-20)
21. This may be a material problem, kin to giving cough syrup to lung cancer patients. You may stop the externally visible symptoms with symptomatic treatment, but you're not curing the underlying disease. Sometimes having bad symptoms can be a good thing when it comes to forcing attention to be paid to a serious underlying problem. [↑](#footnote-ref-21)
22. *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

    [↑](#footnote-ref-22)
23. http://www.ietf.org/rfc/rfc2131.txt [↑](#footnote-ref-23)
24. DHCP address assignments are normally record in DHCP logs, and can be correlated with customer name and contact information when necessary, at least until DHCP logs get discarded. [↑](#footnote-ref-24)
25. IPv6 Stateless Address Autoconfiguration, http://tools.ietf.org/html/rfc4862 [↑](#footnote-ref-25)
26. Privacy Extensions for Stateless Address Autoconfiguration in IPv6, http://tools.ietf.org/html/rfc4941 [↑](#footnote-ref-26)
27. http://www.circleid.com/posts/20120326\_household\_botnet\_infections/ [↑](#footnote-ref-27)
28. http://www.census.gov/compendia/statab/cats/information\_communications/internet\_publishing\_and\_broadca

    sting\_and\_internet\_usage.html (table 1155) [↑](#footnote-ref-28)
29. http://cbl.abuseat.org/countrypercapita.html [↑](#footnote-ref-29)
30. 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%). [↑](#footnote-ref-30)
31. http://rarediseases.info.nih.gov/RareDiseaseList.aspx [↑](#footnote-ref-31)
32. McAfee Threat Report, Third Quarter 2012, PDF page 30,

    http://www.mcafee.com/ca/resources/reports/rp-quarterly-threat-q3-2012.pdf [↑](#footnote-ref-32)
33. *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 [↑](#footnote-ref-33)
34. http://www.eleven.de/botnet-timeline-en.html [↑](#footnote-ref-34)
35. http://cbl.abuseat.org/country.html as of Sunday, December 9th, 2012. [↑](#footnote-ref-35)
36. See for example Zeus Tracker, https://zeustracker.abuse.ch/ [↑](#footnote-ref-36)
37. Has the FCC explicitly indicated that they'd be interested in receiving data of this sort, and told ISPs about an address/department to which such data might be sent? Are there clear terms and conditions around how such data would be used or potentially redisclosed? [↑](#footnote-ref-37)
38. Yes, there are statutory limits to what data can be shared by the government with ISPs, and there was legislation proposed to deal with this problem, but that legislation hasn't passed to-date. [↑](#footnote-ref-38)
39. One noteworthy exception: customers who are security researchers! [↑](#footnote-ref-39)
40. Hypothetically, a botnet might choose to use raw IP addresses, or to use a peer-to-peer alternative to traditional DNS, such as distributed hash tables.

    [↑](#footnote-ref-40)
41. Remember the Internet's negative reaction to SOPA/PIPA. [↑](#footnote-ref-41)
42. http://en.wikipedia.org/wiki/Agobot [↑](#footnote-ref-42)
43. http://en.wikipedia.org/wiki/Bagle\_%28computer\_worm%29 [↑](#footnote-ref-43)
44. http://www.symantec.com/connect/blogs/anatomy-bamital-prevalent-click-fraud-trojan [↑](#footnote-ref-44)
45. http://en.wikipedia.org/wiki/BredoLab [↑](#footnote-ref-45)
46. http://nakedsecurity.sophos.com/2012/12/05/the-citadel-crimeware-kit-under-the-microscope/ [↑](#footnote-ref-46)
47. http://en.wikipedia.org/wiki/Conficker [↑](#footnote-ref-47)
48. http://en.wikipedia.org/wiki/Coreflood [↑](#footnote-ref-48)
49. http://en.wikipedia.org/wiki/Cutwail [↑](#footnote-ref-49)
50. http://blog.malwarebytes.org/intelligence/2012/06/you-dirty-rat-part-1-darkcomet/ [↑](#footnote-ref-50)
51. http://ddos.arbornetworks.com/2011/08/dirt-jumper-caught/ [↑](#footnote-ref-51)
52. http://en.wikipedia.org/wiki/Donbot\_botnet [↑](#footnote-ref-52)
53. http://blog.eset.com/wp-content/media\_files/king-of-spam-festi-botnet-analysis.pdf [↑](#footnote-ref-53)
54. http://en.wikipedia.org/wiki/Trojan\_BackDoor.Flashback [↑](#footnote-ref-54)
55. http://www.theverge.com/2012/8/5/3220834/grum-spam-botnet-attack-fireeye-atif-mushtaq [↑](#footnote-ref-55)
56. http://en.wikipedia.org/wiki/Kelihos\_botnet [↑](#footnote-ref-56)
57. http://en.wikipedia.org/wiki/Koobface [↑](#footnote-ref-57)
58. http://en.wikipedia.org/wiki/Kraken\_botnet [↑](#footnote-ref-58)
59. http://en.wikipedia.org/wiki/Lethic\_botnet [↑](#footnote-ref-59)
60. http://labs.m86security.com/2009/10/maazben-best-of-both-worlds/ [↑](#footnote-ref-60)
61. http://en.wikipedia.org/wiki/Mariposa\_botnet [↑](#footnote-ref-61)
62. http://en.wikipedia.org/wiki/Mega-D\_botnet [↑](#footnote-ref-62)
63. http://en.wikipedia.org/wiki/Nitol\_botnet [↑](#footnote-ref-63)
64. http://riskman.typepad.com/perilocity/2012/03/

    what-other-asns-were-affected-by-botnet-ogee-in-february-2012.html [↑](#footnote-ref-64)
65. http://www.symantec.com/connect/blogs/pikspam-sms-spam-botnet [↑](#footnote-ref-65)
66. http://www.microsoft.com/security/portal/threat/encyclopedia/entry.aspx?Name=Win32%2fPushbot [↑](#footnote-ref-66)
67. http://en.wikipedia.org/wiki/Ramnit [↑](#footnote-ref-67)
68. http://en.wikipedia.org/wiki/Rustock [↑](#footnote-ref-68)
69. http://en.wikipedia.org/wiki/Sality [↑](#footnote-ref-69)
70. http://www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/

    white-papers/wp\_sdbot\_irc\_botnet\_continues\_to\_make\_waves\_pub.pdf [↑](#footnote-ref-70)
71. http://en.wikipedia.org/wiki/Spybot\_worm [↑](#footnote-ref-71)
72. https://spyeyetracker.abuse.ch/ [↑](#footnote-ref-72)
73. http://en.wikipedia.org/wiki/Srizbi\_botnet [↑](#footnote-ref-73)
74. http://en.wikipedia.org/wiki/Storm\_botnet [↑](#footnote-ref-74)
75. http://en.wikipedia.org/wiki/TDL4\_botnet [↑](#footnote-ref-75)
76. http://en.wikipedia.org/wiki/Waledac\_botnet [↑](#footnote-ref-76)
77. http://about-threats.trendmicro.com/us/webattack/118/XPAJ+Back%20with%20a%20Vengeance [↑](#footnote-ref-77)
78. http://en.wikipedia.org/wiki/ZeroAccess\_botnet [↑](#footnote-ref-78)
79. http://en.wikipedia.org/wiki/Zeus\_botnet [↑](#footnote-ref-79)