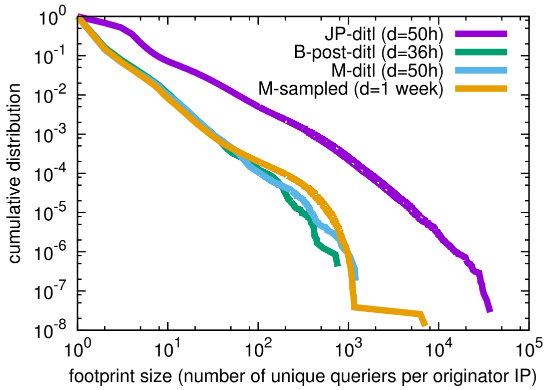


Figure 5: Distribution of originator footprint size.

5. RESULTS

We next study network-wide activities with our method, identifying large network events and trends in different applications and over time. Since our approach is based on feedback from targets, our results complement prior studies (such as darknets) and will observe targeted events will not appear in darknets.

5.1 Sizes of Originator Footprints

We estimate the footprint of each originator as the number of unique queriers per originator. Figure 5 shows the fraction of originators with each footprint size a log-log scale for each of our three datasets.

Our data suggests *there are hundreds of originators that touch large parts of the Internet*. Our controlled trials (§ 4.4) show high attenuation at root servers, yet hundreds of originators have footprints suggesting they scan most or all of the Internet (590 in M-ditl and 298 in B-post-ditl have footprints larger than 10^2).

The *distributions* of footprints is consistent across our datasets. (We cannot directly compare footprint sizes due to variation in duration and sampling.) As one would expect, they are a heavy-tailed, with some originators triggering queries from 10k queriers. We focus the remainder of our analysis of the originators with the largest footprints, typically the top-10000 (about 0.5% of each dataset), or the top-1000 or -100. Considering only large originators will miss those that are intentionally trying to be stealthy, but many scanners make no such attempt [17], and we expect commercial large services to also be open.

The largest footprints here are larger than those we observe in controlled trials at M-Root (Figure 3). Those scans were quite short (a few to a dozen hours), while here we aggregate data over one or two days. In addition, our trials used random targets, most of which are unoccupied (only 6–8% respond, as seen before [25]); many real-world scans are targeted, resulting in higher responses rates and thus greater backscatter.

5.2 Observability and Size of Application Classes

We next classify the top originators. Our goal is to understand what activity is taking place and approximately how aggressive they are. Our key observations are: there are *thousands* of originators causing network-wide activity, different *authorities see different applications*, and *we see evidence of team of coordinated scanners* even with *no direct information from originators.*

Size of application classes: There are *thousands* of unique originators that touch large parts of the Internet. Table 6 shows how many originators we see in each originator class for each dataset, with classes with counts within 10% of the largest count in bold. We use our preferred classifier (RF) with per-dataset training over the entire ground-truth. Classes that lack ground truth for some dataset have no matches (a "-").

Applications vary by authority: The classes of applications seen at *different authorities vary considerably.* For JP-ditl, *spam* is the most common class of originator. Although Japan hosts computers for most major CDNs, the size of the *cdn* class seen from backscatter is small because CDNs often use address space assigned by other registrars (we verify this statement for Akamai and Google with geolocation, whois and prior work [21]). The *update* class is exactly those in labeled ground-truth. We identified this class in examining the data (not from an external source), and lack of additional examples suggests either class has insufficient training data to avoid over-fitting, or update servers are rare.

Both unsampled root servers (B-post-ditl and M-ditl) show similar distributions of activity, with *mail* the most common and *spam* and *cdn* both close. The larger number of CDNs in at M-Root is due to 300 *cdn* originators located in two Chinese ISPs and interacting with queriers in China. Classification appears correct (they do not send traffic to darknets, nor appear in spam blacklists), but the originators lack domain names and we cannot identify them. Such originators appear only in M-ditl, suggesting that their queriers may be using DNS resolvers that prefer nearby authorities, since M-Root is well provisioned in Asia while B-Root is only based in North America.

Long-term, sampled root data (M-sampled) has some important differences from short term (M-ditl). Consider relative sizes of classes (since absolute counts vary due to dataset duration), we see many more scanner and spammers in long-term data. We believe the size of these categories reflect churn in the population carrying out the activity. We expect churn in spamming where computers known for spamming are less effective. We measure churn directly for scanners in § 5.3.

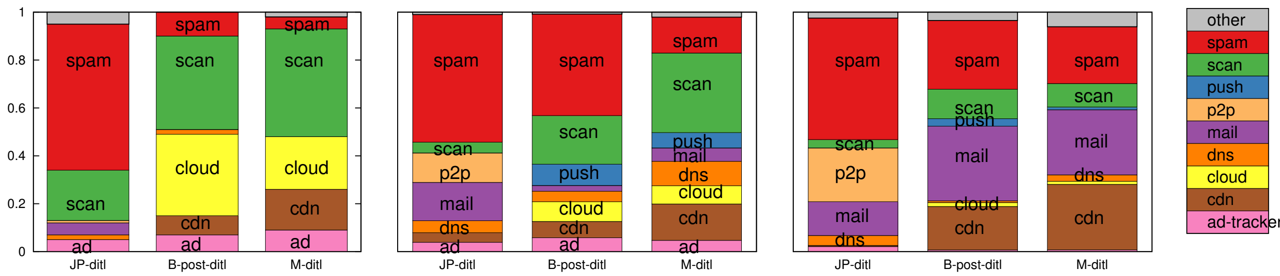
Big footprints can be unsavory: The mix of applications varies as we examine originators with smaller footprints, but we see that *big footprints are often unsavory activity*. Figure 6 shows how originator classes change as we look at more originators with smaller footprints (from Figure 6a to Figure 6c).

The largest footprints are often spammers (in JP-ditl) or scanners (for B and M). By contrast, we see that *mail* appears only in the top-1000 and top-10000, suggesting that legitimate mail servers may service large mailing lists (to many targets), but spammers touch many more targets. For B and M, more spammers rise in Figure 6c, suggesting spreading of traffic over many smaller originators to evade filtering.

By contrast, large but not top originators are often infrastructure: *cloud*, *mail*, *ad-tracker*, and *crawler*. In general, we find that application classes have a "natural" size, with some favoring origins with large footprints (prominent in

data	ad-track	cdn	cloud	crawl	dns	mail	ntp	p2p	push	scan	spam	update
JP-ditl	210	49	-	-	414	1412	237	2235	-	355	**5083**	6
B-post-ditl	72	1782	168	361	76	**3137**	8	-	318	1228	2849	-
M-ditl	76	**2692**	135	557	258	**2750**	67	-	119	983	2353	-
M-sampled	1329	17,708	2035	885	1202	14,752	-	-	3652	**47,201**	34,110	-

Table 6: Number of originators in each class for all datasets. (Classifier: RF.)



(a) Top 100

(b) Top 1000

(c) Top 10,000

Figure 6: Fraction of originator classes of top-N originators. (Dataset: JP-ditl, B-post-ditl, M-ditl; classifier: RF.)

Figure 6a), while others favor smaller footprints and so are more common in Figure 6c.

The *ad-tracker* class is most common in the prominent in top-most originators (a larger red *ad-tracker* fraction in Figure 6a compared to to Figure 6c). There are relatively a few originators (we see 5 companies as 22 unique originating addresses for top-100/JP-ditl). Unlike spam, they need not hide, and are likely prominent because tracking needs little traffic (a few originators can support a network-wide service), and because they use DNS records with short cache lifetimes (small TTLs). *Cloud* follows this pattern as well; 1 company across 21 distinct originating IPs for top-100 in M-ditl.

The *crawler* class shows the opposite behavior: most crawlers appear only in the top-10000, with few in top-1000 (554 vs. 3). This shift is consistent with web crawlers being data intensive, operating across many distributed IP addresses in parallel.

We also see that the physical location of the authority influences what they see. We earlier observed how differences in *cdn* for M-Root and B-Root are explained by their physical location to CDNs in China. B-Root's U.S.-only location may place it closer to more services in *cloud* (see Figure 6a) compared to M-Root's locations mainly in Asia and Europe.

New and old observations: A new observation in our data is *potential teams of scanners.* We have manually identified several /24 address blocks where many addresses are engaged in scanning, suggesting possible parallelized scanning. Without direct scan traffic, we cannot confirm coordination, but backscatter suggests networks for closer examination. To understand with scope of potential collaborative teams, we start with a a very simple model where a team is multiple originators in the same /24 IP address block. In M-sampled we see 5606 unique scan originators (by IP address), across 2227 unique originating /24 address blocks. Of these, 167 blocks have 4 or more originators, suggesting a potential team of collaborators. While 128 of these blocks have multiple application classes, suggesting against collaboration (or possibly mis-classification), we see 39 blocks with 4 or more originators all with the same application class. Such blocks warrant closer examination.

We also confirmed prior observations that clients linger on retired services. Originators we find include four retired root DNS servers (B, D, J, L), two prior cloud-based mail servers, and one prior NTP server. These cases show our methods can be used to systematically identify overly-sticky, outdated clients across many services, automating prior reports of clients that stick to retired servers in DNS [29] and NTP [39].

Classification on originator actions: An important benefit of our approach is that we classify on *indirect actions caused by the originator, with no direct information* from the originator. In fact, about a quarter of the originators in JP-ditl and half of those in the root datasets have no reverse domain names, but originator omissions have no affect on our approach because we do not observe *any* traffic or reverse names of originators. This separation makes it more difficult for adversarial originators to conceal their activities.

5.3 Trends in Network-Wide Activity

We next look for long-term trends in our data. We believe this is the first longitudinal study of network-wide activities such as scanning (prior work focused on specific events [17]). Our goal is to understand the ebb and flow of network-wide activity, so rather than examine the N largest originators, we count all originators with footprints of at least 20 queriers (see also Figure 5). While we see no growth in network-wide events, we see *peaks that respond to security events* and *a core of slow-and-steady scanners.*

Peaks in numbers of originators: The top *all* line in Figure 7 shows the absolute number of originators over time, each class (the lower, colored lines) and total (the top, black line). There are fairly large week-by-week changes, showing churn in the number of active originator activities,

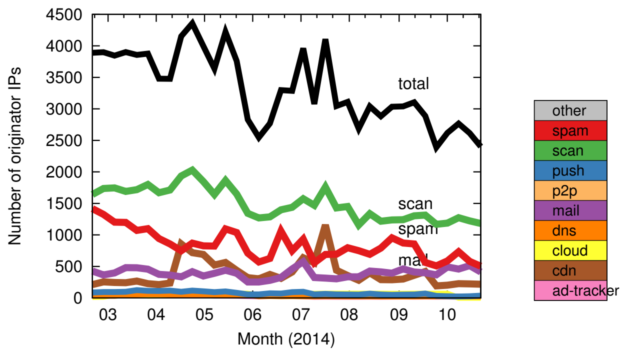


Figure 7: Number of originators over time. (Dataset: M-sampled; classifier: RF.)

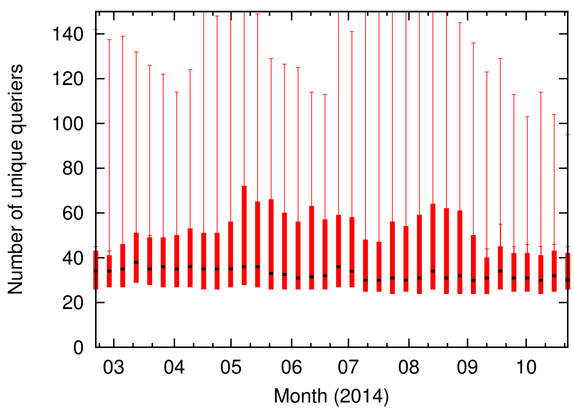


Figure 8: Box plot of originator footprint (queriers per scanner) over time; whiskers: 10%ile/90%ile. (Dataset: M-sampled.)

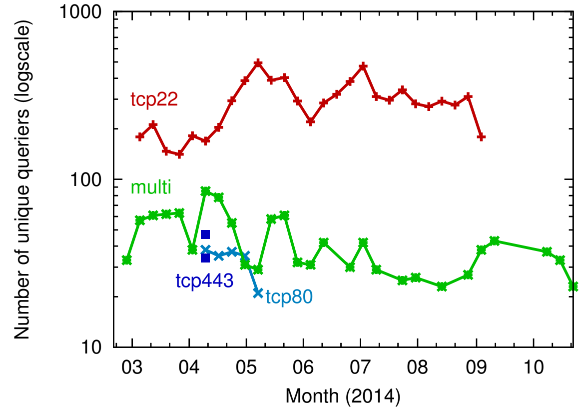


Figure 9: Three example originators with application class *scan*. (Dataset: M-sampled with darknet.)

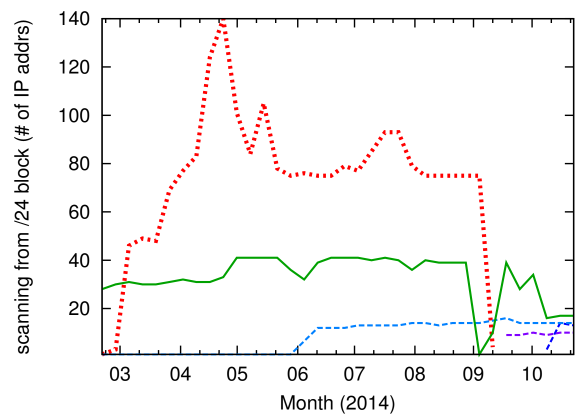


Figure 10: Five example blocks originating scanning activity. (Dataset: M-sampled.)

and peaks that can be explained by reactions to network security events.

To understand how network activity results from real-world events we next look the *scanner* application class. Our observation period includes public announcement of the Heartbleed vulnerability on 2014-04-07 [38], and we know that there were multiple research [1, 18], commercial, and presumably government scanning activities triggered by that announcement. The green scanner line in Figure 7 shows more than a 25% increase in scanning by mid-April, from 1400 originator IPs per week to 1800 at its peak. While this change is noticeable, it is smaller than we might expect. Instead, it shows that reaction to Heartbleed is small compared to *the large amount of scanning that happens at all times*—the 1200–1400 scanners we saw in March, and the 1000–1200 scanners that are present from June to October.

Churn: To understand long-term scanning, Figure 8 shows the distribution of footprint sizes over time for class *scan*. While the median and quartiles are both stable over these 36 weeks, but the 90th percentile varies considerably. This variation suggests a few very large scanners that come and go, while a core of slower scanners are always present.

We illustrate this observation with three different scanners that appear in both M-sampled and our darknet data (Figure 9). Two are long-lived (the top "tcp22" line scanning ssh, and the middle line scanning multiple ports), while the tcp80 scanner occurs in April and May. Furthermore, two tcp443 scans only appear in one week in April (shown as dark squares), suggesting they are Heartbleed-related. We also see that tcp22 has a bigger footprint than the others, and it looks a part of a big campaign whose 140 IP addresses belong to the same /24 block. Using our darknets, we confirm 164 scanners for TCP ports 22, 80, or 443, and while there is no "typical" scanner, these variations are common.

Our approach also *identifies networks supporting scanners*. For each /24 block, we count the number of IP addresses in class *scan* over time; Figure 10 shows five of these blocks. The top dotted line is a block with large scanning peaks corresponding with Heartbleed and Shellshock, ending in September. The solid line shows a block that scans continuously, while the three dotted lines are blocks that start scanning during our observation.

To understand if *who* scans changes over time, Figure 11 measures week-by-week change in scanner IP addresses. The bar above the origin shows the number of scanners each week, showing both new originators (top, dark) and con-

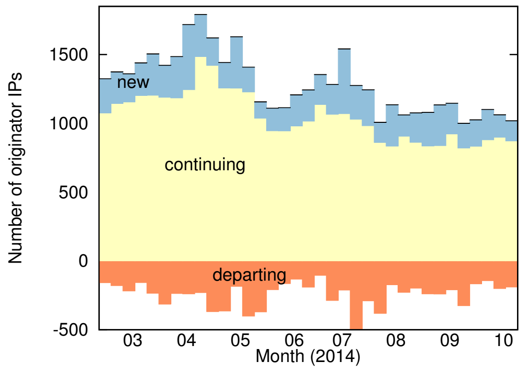


Figure 11: Week-by-week churn for originators of class *scan*. (Dataset: M-sampled.)

tinuing originators (middle, light) The red bar below the origin shows scanners that were lost from the prior week. While there are always scanners coming and going (about 20% turnover per week), this data confirms that there is a stable core of scanners that are consistently probing, week-after-week.

6. RELATED WORK

We next review prior work in both DNS- and non-DNS-based sensors and analysis. Overall, our contribution is to show that reverse DNS queries can identify network-wide behavior. Prior work instead considers forward DNS traffic and typically applies it to specific problems, or uses non-DNS sensors such as darknets and search engines.

DNS-specific sensors: Several groups use forward DNS queries to identify spam [56, 27], fast-flux [56, 26], automatically generated domain names [55], and cache poisoning [2]. Like our work, several of these approaches use machine learning to classify activity. However, this prior work focuses on *forward* DNS queries, while we consider *reverse* queries. Moreover, many use algorithms optimized to detect specific malicious activities, while we detect a range of network-wide behavior.

Recent work has used DNS to infer the structure of CDN networks [4] or internal to DNS resolvers [44]. They infer specific services from DNS traffic, we search for network-wide events from reverse queries.

An earlier work uses targeted scan and DNS backscatter for detecting Tor exit routers peeking POP3 authentication information [33], an earlier use of DNS backscatter de-anonymization; we generalize this use to detect scanners.

Plonka and Barford use machine-learning-based clustering and visualization to identify undesirable activity from local DNS traffic [40]. They use DNS traffic from an organization's recursive resolver to infer activity about that organization. Overall, our approach provides larger coverage, both by using data from authoritative DNS servers that aggregate queries from many organizations, unlike their single organization, and by examining trends in ten months of data, unlike their week-long analysis.

Antispam software has long used reverse DNS lookups to directly classify sources of mail. We use the domain names of queriers to indirectly classify originators.

Non-DNS Passive sensors: Darknets (or network telescopes) are a commonly used passive technique to characterize large-scale network activity [37, 34, 54, 13, 14, 17]. By monitoring a large, unoccupied blocks of addresses, darknets see active probes from viruses and scanners, queries from misconfiguration, and backscatter from spoofed traffic; traffic that can predict global malware, and its absence, network outages. Our analysis of DNS backscatter shares the goal of understanding network-wide activity from a simple, passive observer, but we observe at DNS authorities rather than large ranges of addresses. Like Durumeric et al. [17], we seek to enumerate scanners, but our use of DNS backscatter will see targeted scans that miss their darknet, and our study considers eight months of activity, not just one.

Some security services use middleboxes with deep-packet inspection to passively monitor large ISPs [3]. They observe all traffic from multiple points, while we monitor DNS backscatter from a single provider only.

Staniford monitored network traffic for scanners [47], and Gates emphasized rapid detection with scanner modeling [23]. Rather than protecting a single network, we look for network-wide activity with a simple observer.

Honeypots (for example, [41]) are a form of application-level darknet. By interacting with originators they see attacks darknets miss, but they miss attacks that probe specific targets (such as Alexa top sites). Interactivity also makes them fewer because of deployment expense. DNS backscatter uses information from existing servers.

Unconstrained endpoint profiling [48] uses search engines to gather information on addresses that leak into the public web, possibly complementing network flow data. We both seek to understand network-wide activity, but we use different data sources and methods. They use largely unstructured information from the web, while we infer features from semi-structured domain names and also traffic patterns. Their work depends on the speed of search engine indexing, while our work can provide rapid feedback given data from a DNS authority.

General DNS traffic analysis and privacy: Finally, a wide body of work has explored DNS traffic in general (examples include [15, 52, 11, 22]). Their work seeks to understand DNS, while we instead study what reverse DNS tells us about network-wide activities.

Work in DNS privacy focuses on client-to-recursive resolvers for end-users (examples include [36, 57], and proposals in the IETF DPRIVE working group). Our use of reverse queries from automated systems triggered by originators should see almost no human-triggered, end-user queries (§ 2). Use of query minimization [5] at the queriers will constrain the signal to only the local authority (that immediately serving the originator's reverse address).

7. CONCLUSION

We identified DNS backscatter as a new source of information about benign and malicious network-wide activity, including originators of mailings list traffic, CDN infrastructure, spammers and scanners. Their activity triggers reverse DNS queries by or near their targets, and we show that classification of these queriers allows us to identify classes of activity with reasonable precision. We use our approach to identify trends in scanning across nine months of data from one data source, and we characterize several kinds of activity for two days over three data sources. Our work provides a new approach to evaluate classes of network-wide activity.

Acknowledgments: We thank Yuri Pradkin for B-Root data collection, Akira Kato for M-Root data, and Yoshiro Yoneya and Takeshi Mitamura for JP data. We thank Xun Fan for input about Google and Akamai sites in Japan. We thank Terry Benzel, Kenjiro Cho, Ethan Katz-Bassett, Abdul Qadeer, and John Wroclawski comments on this paper.

Kensuke Fukuda's work in this paper is partially funded by Young Researcher Overseas Visit Program by Sokendai, JSPS KAKENHI Grant Number 15H02699, and the Strategic International Collaborative R&D Promotion Project of the Ministry of Internal Affairs and Communication in Japan (MIC) and by the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement No. 608533 (NECOMA). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the MIC or of the European Commission.

John Heidemann's work in this paper is partially sponsored by the Department of Homeland Security (DHS) Science and Technology Directorate, HSARPA, Cyber Security Division, via SPAWAR Systems Center Pacific under Contract No. N66001-13-C-3001, and via BAA 11-01-RIKA and Air Force Research Laboratory, Information Directorate under agreement number FA8750-12-2-0344. The U.S. Government is authorized to make reprints for Governmental purposes notwithstanding any copyright.The views contained herein are those of the authors and do not necessarily represent those of DHS or the U.S. Government.

8. REFERENCES

[1] Mustafa Al-Bassam. Top Alexa 10,000 Heartbleed scan. https://github.com/musalbas/heartbleed-masstest/blob/94cd9b6426311f0d20539e696496ed3d7bdd2a94/top1000.txt, April 14 2014.

[2] Manos Antonakakis, David Dagon, Xiapu Luo, Roberto Perdisci, and Wenke Lee. A centralized monitoring infrastructure for improving DNS security. In *Proc. of the 13th International Symposium on Recent Advances in Intrusion Detection*, pages 18–37, Ottawa, Ontario, Canada, September 2010. Springer.

[3] Arbor Networks. Worldwide infrastructure security report. Technical Report Volume IX, Arbor Networks, January 2014.

[4] Ignacio Bermudez, Marco Mellia, Maurizio M. Munafo, Ram Keralapura, and Antonio Nucci. DNS to the rescue: Discerning content and services in a tangled web. In *Proc. of the ACM Internet Measurement Conference*, pages 413–426, Boston, MA, November 2012.

[5] S. Bortzmeyer. DNS query name minimisation to improve privacy. Work in progress (Internet draft draft-bortzmeyer-dns-qname-minimisation-02), May 2014.

[6] Carna Botnet. Internet census 2012: Port scanning /0 using insecure embedded devices. web page http://census2012.sourceforge.net/paper.html, March 2013.

[7] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, October 2001.

[8] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees.* Chapman and Hall, 1984.

[9] Nevil Brownlee. One-way traffic monitoring with iatmon. In *Proc. of the Passive and Active Measurement Workshop*, pages 179–188, Vienna, Austria, March 2012.

[10] Matt Calder, Xun Fan, Zi Hu, Ethan Katz-Bassett, John Heidemann, and Ramesh Govindan. Mapping the expansion of Google's serving infrastructure. In *Proc. of the ACM Internet Measurement Conference*, pages 313–326, Barcelona, Spain, October 2013. ACM.

[11] Sebastian Castro, Duane Wessles, Marina Fomenkov, and kc Claffy. A day at the root of the Internet. *ACM SIGCOMM Computer Communication Review*, 38(5):41–46, October 2008.

[12] Jakub Czyz, Michael Kallitsis, Manaf Gharaibeh, Christos Papadopoulos, Michael Bailey, and Manish Karir. Taming the 800 pound gorilla: The rise and decline of NTP DDoS attacks. In *Proc. of the ACM Internet Measurement Conference*, pages 435–448, Vancouver, BC, Canada, November 2014. ACM.

[13] Jakub Czyz, Kyle Lady, Sam G. Miller, Michael Bailey, Michael Kallitsis, and Manish Karir. Understanding IPv6 Internet background radiation. In *IMC'13*, pages 105–118, Barcelona, Spain, 2013.

[14] Alberto Dainotti, Claudio Squarcell, Emile Aben, Kimberly C. Claffy, Marco Chiesa, Michele Russo, and Antonio Pescape. Analysis of country-wide internet outages caused by censorship. In *Proc. of the ACM Internet Measurement Conference*, pages 1–18, Berlin, Germany, November 2011.

[15] Peter B. Danzig, Katia Obraczka, and Anant Kumar. An analysis of wide-area name server traffic: A study of the Domain Name System. In *Proc. of the ACM SIGCOMM Conference*, pages 281–292, January 1992.

[16] DNS-OARC. Day in the life of the internet (DITL) 2014. https://www.dns-oarc.net/oarc/data/ditl, April 2014.

[17] Zakir Durumeric, Michael Bailey, and J. Alex Halderman. An Internet-wide view of Internet-wide scanning. In *Proc. of the 23rd USENIX Security Symposium*, pages 65–78, San Diego, CA, August 2014. USENIX.

[18] Zakir Durumeric, Frank Li, James Kasten, Johanna Amann, Jethro Beekman, Mathias Payer, Nicolas Weaver, David Adrian, Vern Paxson, Michael Bailey, and J. Alex Halderman. The matter of Heartbleed. In *Proc. of the ACM Internet Measurement Conference*, pages 475–488, Vancouver, BC, Canada, November 2014. ACM.

[19] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. ZMap: Fast Internet-wide scanning and its security applications. In *Proc. of the USENIX Security Symposium*, pages 605–620, Washington, DC, USA, August 2013. USENIX.

[20] Robert Edmonds. ISC passive DNS architecture. Technical report, Internet Systems Consortium, Inc., March 2012.

[21] Xun Fan, Ethan Katz-Bassett, and John Heidemann. Assessing affinity between users and CDN sites. In *Proc. of the 7th Workshop on Traffic Monitoring and Analysis (TMA)*, pages 95–110, Barcelona, Spain, April 2015. Springer.

[22] Hongyu Gao, Vinod Yegneswaran, Yan Chen, Phillip Porras, Shalini Ghosh, and Jian Jiang Haixing Duan. An empirical reexamination of global DNS behavior. In *Proc. of the ACM SIGCOMM Conference*, pages 267–278, Hong Kong, China, 2013.

[23] Carrie Gates. Coordinated scan detection. In *Proc. of the ISOC Network and Distributed System Security Symposium*, San Diego, CA, February 2009. The Internet Society.

[24] Kenneth Geers, Darien Kindlund, Ned Moran, and Rob Rachwald. World War C: Understanding nation-state motives behind today's advanced cyber attacks. Technical report, FireEye, September 2014.

[25] John Heidemann, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Genevieve Bartlett, and Joseph Bannister. Census and survey of the visible Internet. In *Proc. of the ACM Internet Measurement Conference*, pages 169–182, Vouliagmeni, Greece, October 2008. ACM.

[26] Thorsten Holz, Christian Gorecki, Konrad Rieck, and Felix Freiling. Measuring and detecting fast-flux service networks. In *Proc. of the ISOC Network and Distributed System Security Symposium*, San Diego, CA, USA, February 2008. The Internet Society.

[27] Keisuke Ishibashi, Tsuyoshi Toyono, Katsuyasu Toyama, Masahiro Ishino, Haruhiko Ohshima, and Ichiro Mizukoshi. Detecting mass-mailing worm infected hosts by mining DNS traffic data. In *Proc. of the ACM SIGCOMM MineNet Workshop*, pages 159–164, Philadelphia, PA, August 2005.

[28] Julian Kirsch, Christian Grothoff, Monika Ermert, Jacob Appelbaum, Laura Poitras, and Henrik Moltke.

NSA/GCHQ: The HACIENDA program for internet colonization. *C'T Magazine*, Aug.15 2014.

[29] Matthew Lentz, Dave Levin, Jason Castonguay, Neil Spring, and Bobby Bhattacharjee. D-mystifying the D-root address change. In *Proc. of the ACM Internet Measurement Conference*, pages 57–62, Barcelona, Spain, October 2013. ACM.

[30] Kirill Levchenko, Andreas Pitsillidis, Neha Chachra, Brandon Enright, Márk Félegyházi, Chris Grier, Tristan Halvorson, Chris Kanich, Christian Kreibich, He Liu, Damon McCoy, Nicholas Weaver, Vern Paxson, Geoffrey M. Voelker, and Stefan Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Proc. of the IEEE Symposium on Security and Privacy*, pages 431–446, Oakland, CA, USA, May 2011. IEEE.

[31] Zhichun Li, Anup Goyal, Yan Chen, and Aleksandar Kuzmanovic. Measurement and diagnosis of address misconfigured P2P traffic. In *INFOCOM'10*, pages 1–9, San Diego, CA, March 2010.

[32] MaxMind LLC. GeoIP. http://www.maxmind.com/geoip.

[33] Damon McCoy, Kevin Bauer, Dirk Grunwald, Tadayoshi Kohno, and Douglas Sicker. Shining light in dark places: Understanding the tor network. In *Proc. of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 63–76, Leuven, Belgium, July 2008.

[34] Jon Oberheide, Manish Karir, Z. Morley Mao, and Farnam Jahanian. Characterizing dark DNS behavior. In *Proc. of the 4th International Conference on Detection of Intrusions & Malware, and Vulnerability Assessment (DIMVA)*, pages 140–156, Lucerne, Switzerland, July 2007. Springer.

[35] Robert O'Harrow, Jr. Cyber search engine Shodan exposes industrial control systems to new risks. *The Washington Post*, June 3 2012.

[36] OpenDNS. DNSCrypt: Introducing DNSCrypt. web page http://www.opendns.com/about/innovations/dnscrypt/, January 2014.

[37] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of Internet background radiation. In *Proc. of the ACM Internet Measurement Conference*, pages 27–40, Sicily, Italy, 2004.

[38] Nicole Perlroth. Thought safe, websites find the door ajar. *New York Times*, page A1, Apr. 9 2014.

[39] Dave Plonka. Flawed routers flood university of wisconsin internet time server. http://pages.cs.wisc.edu/~plonka/netgear-sntp, 2003.

[40] David Plonka and Paul Barford. Context-aware clustering of DNS query traffic. In *Proc. of the ACM Internet Measurement Conference*, pages 217–229, Vouliagmeni, Greece, October 2008.

[41] Niels Provos. A virutal honeypot framework. In *Usenix Security Symposium 2004*, pages 1–14, San Diego, CA, August 2004.

[42] Lin Quan, John Heidemann, and Yuri Pradkin. Trinocular: understanding Internet reliability through adaptive probing. In *Proc. of the ACM SIGCOMM Conference*, pages 255–266, Hong Kong, China, August 2013.

[43] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[44] Kyle Schomp, Tom Callahan, Michael Rabinovich, and Mark Allman. On measuring the client-side DNS infrastructure. In *Proc. of the ACM Internet Measurement Conference*, pages 77–90, Barcelona, Spain, October 2013.

[45] Farsight Security. SIE (Security Information Exchange). https://www.farsightsecurity.com/Services/SIE/, 2013.

[46] Shadow server foundation. http://www.shadowserver.org/.

[47] Stuart Staniford, James A. Hoagland, and Joseph M. McAlerney. Practical automated detection of stealthy portscans. *Journal of Comptuer Security*, 10(1):105–136, 2002.

[48] Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovi, and Antonio Nucci. Unconstrained endpoint profiling (Googling the Internet). In *Proc. of the ACM SIGCOMM Conference*, pages 279–290, Seattle, WA, Aug 2008.

[49] USC/LANDER project. Internet address census, datasets internet_address_census it63w, it63c, it63j, it63g, it64w, it64c, it64j, it64g. web page http://www.isi.edu/ant/lander, January (it63) and April (it64) 2015.

[50] Paul Vixie. Passive DNS collection and analysis, the 'dnstap' approach. Keynote talk at FloCon, January 2014.

[51] Florian Weimer. Passive DNS replication. In *Proc. of the 17th Forum of Incident Response and Security Teams (FIRST)*, Singapore, April 2005.

[52] Duane Wessels and Marina Fomenkov. Wow, that's a lot of packets. In *Proc. of the Passive and Active Measurement Workshop*, La Jolla, CA, April 2003.

[53] Wikipedia. Gini coefficient. http://en.wikipedia.org/wiki/Gini_coefficient, 2015.

[54] Eric Wustrow, Manish Karir, Michael Bailey, Farnam Jahanian, and Geoff Houston. Internet background radiation revisited. In *Proc. of the 10th ACM Internet Measurement Conference*, pages 62–73, Melbourne, Australia, November 2010. ACM.

[55] Sandeep Yadav, Ashwath Kumar, Krishna Reddy, A.L. Narasimha Reddy, and Supranamaya Ranjan. Detecting algorithmically generated malicious domain names. In *Proc. of the ACM Internet Measurement Conference*, pages 48–61, Melbourne, Australia, November 2010.

[56] Bojan Zdrnja, Nevil Brownlee, and Duane Wessels. Passive monitoring of DNS anomalies. In *Proc. of the 4th International Conference on Detection of Intrusions & Malware, and Vulnerability Assessment (DIMVA)*, pages 129–139, Lucerne, Switzerland, 2007.

[57] Liang Zhu, Zi Hu, John Heidemann, Duane Wessels, Allison Mankin, and Nikita Somaiya. Connection-oriented DNS to improve privacy and security. In *Proc. of the 36th IEEE Symposium on Security and Privacy*, pages 171–186, San Jose, Californa, USA, May 2015. IEEE.