Workshop Report: Darkspace and Unsolicited Traffic Analysis (DUST 2012)

Tanja Zseby
Fraunhofer FOKUS and CAIDA
tanja@caida.org

kc claffy
CAIDA
kc@caida.org

This article is an editorial note submitted to CCR. It has NOT been peer reviewed. The authors take full responsibility for this article's technical content.

ABSTRACT

On May 14-15, 2012, CAIDA hosted the first international Workshop on Darkspace and UnSolicited Traffic Analysis (DUST 2012) to provide a forum for discussion of the science, engineering, and policy challenges associated with darkspace and unsolicited traffic analysis. This report captures threads discussed at the workshop and lists resulting collaborations.

Categories and Subject Descriptors

C.2.3 [**Network operations**]: Network monitoring; C.2.5 [**Local and Wide-Area Networks**]: Internet; C.2.6 [**Internetworking**]: Standards; C.4.2 [**Performance of Systems**]: Measurement techniques—Passive

Keywords

passive measurement, Internet measurement techniques, IP darkspace, traffic analysis, data sharing

1. MOTIVATION

Internet Protocol (IP) "darkspaces" are globally routable address segments with no active hosts. All traffic to an IP darkspace comes unsolicited and unidirectional. Observing and analyzing darkspace traffic helps detect and analyze global incidents such as scanning, DDoS attacks, network outages and misconfigurations.

With the DUST 2012 workshop we aimed to bring together different groups that work on darkspace analysis and related fields. This included operators of darkspace monitors, researchers engaged in darkspace and unsolicited traffic analysis, scientists interested in working on the UCSD darkspace data, scientists or organizations interested in setting up a darkspace monitor as well as scientists working on related topics such as honeynets, intrusion detection, and data sharing methods.

The goals of the workshop included improving methods for darkspace traffic collection, curation, scientific analysis, correlation, and privacy-sensitive sharing of darkspace traffic data across research groups. As a broader goal we want to build a community of darkspace monitor operators and scientists seeking to share data and to explore the development of an integrated network of existing darkspace monitors for real-time comparative analysis. Materials presented at the workshop are available at http://www.caida.org/workshops/dust/.

2. DARKSPACE ANALYSIS

Xenofontas Dimitropoulos (ETH Zurich) presented a classification scheme for one-way traffic, which can work even on links with bidirectional traffic, i.e., filtering out all bidirectional traffic and only examining the unsolicited one-way traffic. His approach distinguishes six classes of one-way traffic according to host behaviors and flow features: unreachable services; P2P applications; scanning; backscatter; suspected benign and bogon. He applied his method to flow data from 2004-2011 collected on the Swiss academic backbone network (SWITCH), finding 34%-67% of the observed flows were only one-way, mostly from scanning, P2P protocols, and unreachable services. The number of total flows increased over the 8 years, but the number of one-way flows remained relatively consistent. Fontas' method can also be applied to service availability monitoring.

Shouhuai Xu (U. Texas, San Antonio) presented a statistical framework for modeling attack traffic from different source IP addresses as stochastic processes. He showed initial results from a analysis of attack processes from data collected using low-interaction honeypots in 2010 and 2012. He found attack rates and inter-arrival times exhibit long range dependencies, and then tried to fit his observations to fractional Gaussian noise (FGN) and fractional ARIMA (FARIMA) processes, finding the latter a closer fit. He acknowledged the need for more data to support confidence in his finding.

David Plonka (University of Wisconsin - Madison) presented a rendezvous-based traffic analysis scheme [9]. The rendezvous is the method a host uses to get the IP address for its communication peer, e.g., DNS, static configuration or application specific mechanisms.

Rendezvous-based traffic analysis has several advantages: more privacy-respecting than most traffic analysis approaches; low traffic volume; easy to separate the rendezvous traffic, e.g., DNS. Using sample data from DNS flows from office and residential networks at his university revealed quite different traffic profiles. Most office traffic used DNS to initiate a communication, unlike the residential traffic. More interestingly, half of residential traffic that used DNS ("named traffic") could be identified by analyzing just five domain names. Rendezvous-based methods are relevant to darkspace analysis because hosts sending traffic to darkspace use different methods from other hosts in selecting destination IP addresses (i.e., malicious and inadvertent actions have atypical characteristics). He also described his *treetop* tool which combines flow information with annotations from rendezvous-based analysis. He pondered to what extent we can trust rendezvous information for host profiling and reputation, and wondered if other rendezvous mechanisms (e.g. WWW, P2P) could as cleanly separate traffic as the DNS-based mechanisms.

Alberto Dainotti (CAIDA) presented an analysis of a botnet scan based on data from the UCSD /8 darkspace. The scan targeted Session Initiation Protocol (SIP) servers using specific UDP packets to port 5060 and TCP SYN packets to port 80. Over 12 days in February 2011 he observed more than 20 million probes from nearly 3 million unspoofed source IPs operating with a degree of coordination not previously reported. He found evidence of the identical botnet behavior in independent traffic sources such as the MAWI/WIDE traces. He presented a heatmap-based visualizations of the scanning pattern of the destination IP addresses, which increased sequentially in reverse byte order, and exhibited strongly delineated phase shifts. Alberto's team modified CAIDA's *cuttlefish* tool [5] to create a multi-window animation of the progression of the scan over time, in geographic, topological, and traffic-volume dimensions. The animation vividly revealed the high degree of coordination of the scan and the diurnal activity patterns of bots around the world.

Tanja Zseby (Fraunhofer FOKUS and CAIDA) described her work on finding efficient metrics that cannot only detect important phenomena in large darknet traffic samples, but also enable darknet operators to share quantitative indicators without having to share sensitive traffic data. In contrast to approaches that require complex packet classification or manual inspection of data to detect events of interest, she has studied a different approach based on the analysis of entropy in two distributions: IP addresses and port numbers. Entropy provides a compact metric to express the dispersion or concentration of feature distributions. Since darkspace traffic is generated by different software processes that use random addresses and port numbers in distinct ways, analyzing the entropy of these distributions can can reveal certain events of interest, which Tanja demonstrated using several months of traffic from the UCSD darkspace monitor. To validate her concept she used these entropy values to classify events, and then compared her results with a baseline analysis of the same data with the tool iatmon (see Nevil's talk, below). Her technique was able to recognize large scanning events, backscatter and probe traffic well, though suffered from the fact that independent events could have opposing effects on entropy, canceling out signals that would otherwise indicate an event. Although not as powerful as fine-grained packet classification approaches, entropy-based classification offers a lightweight alternative that enables rapid detection of some types of major incidents and can facilitate early warning capabilities and operational information exchange among network operators. Tanja plans to experiment with generalized entropy and different time intervals to address those challenges.

3. TOOLS AND METHODS

Eric Ziegast (Internet Systems Consortium ISC) described ISC's SIE infrastructure which collects, encapsulates and re-distributes traffic data. They share data with researchers as well as with commercial users for operational security use. He also discussed some of the challenges of sharing data, including anonymization (which they do not do yet since they rely on privacy agreements/NDAs) and timely processing and distribution of the traffic data.

Joanne Treurniet (Defence R&D Canada) presented a method to classify IP traffic into activity classes, based on IP addresses, port numbers, and expected protocol behavior. She distinguishes 4 TCP session classes (complete, incomplete, invalid, illegal), 2 UDP session classes (unidirectional, bidirectional) and 4 ICMP session classes (request, pair, lone reply, lone error). She then distinguishes four activity classes: productive, scanning, unproductive, ambiguous. The unproductive activity class contains traffic from DoS attacks and invalid TCP sessions caused by timeouts due to NAT and backscatter. Applying her scheme to a one-hour traffic trace from four class B address segments in August 2006, she found 85% of observed sessions were horizontal scans, many to known worm ports. Only 0.3% of the sessions were productive activity. But only 15 % of the byte traffic was scanning, and 82% of the bytes could be considered productive traffic. Her prototype implementation could also be used for real-time analysis, which would require expiring sessions not associated with activity for a certain time. She plans to investigate scalability challenges when applying her techniques to larger networks.

Nevil Brownlee (University of Auckland) presented the *iatmon* (Inter-Arrival Time Monitor) tool designed to classify sources of one-way traffic along two dimensions: inter-arrival time characteristics of packets from a given source IP, and protocol behavior. For each source it stores information from the packets such as protocols, port numbers, flags, and packet inter-arrival times. The *iatmon* analysis generates a 14x10 matrix of source types and source groups. The tool derives the *source type* from packet information and distinguishes 14 classes, such as horizontal or vertical scanning, or backscatter. *iatmon* also recognizes common darkspace traffic, such as Conficker C and μTorrent, by analyzing scanning patterns and payload. The tool derives the *source group* according to the inter-arrival times of packets from a given source IP, distinguishing among ten different source behaviors such as long-lived/stealth sources, sources with peaks at 3 seconds, etc. Nevil showed analysis results from three darkspace monitors, and plans to set up other monitors, further investigate the UDP sources he observed, and explore-data mining techniques to detect changes in types and groups.

Alistair King (CAIDA) introduced the *Corsaro* architecture for supporting collection, curation, and modular extensible plug-ins for analysis of darkspace traffic at the UCSD Network Telescope monitor. Design goals include a high compression rate and fast packet processing. The basic *Corsaro* plug-in creates eight-tuple keys based on source IP, destination IP, source port, destination port, protocol, TCP flags, TTL, IP length and then reports the packet count per eight-tuple for each time interval. The eight-tuple provides enough detail to allow much of the analysis of interest to researchers, but achieves a compression of more than 80% compared to the original pcap file. Further plug-ins, such as a DoS detection method, have also been developed. The tool is in use since February 2012 and compiles on FreeBSD, Linux, Mac OSX and Solaris X. CAIDA plans to extend *Corsaro* to provide real-time analysis, reporting, visualization and archival of darkspace data. Additional plug-ins are planned for geolocation and AS-mapping.

4. IPV6

Geoff Huston (APNIC) presented results from analysis of a /12 APNIC IPv6 darkspace in which 97.25% of the address block is unadvertised and unallocated, i.e., the /12 is a covering prefix for a thin slice of more specific IPv6 prefixes announced underneath it. Since random IP scanning is infeasible in the astronomically large IPv6 address space, it seems unlikely we will observe as much IPv6 darkspace traffic as IPv4 darkspace traffic. Previous work on a /48 IPv6 darkspace by Matt Ford in 2006 reported only one packet per month [4]. Geoff used a much larger address space and analyzed eight days in 2010 and 107 days in 2011. He mostly observed ICMP packets, and most of these were from Teredo connection attempts to hosts in a Japanese network. The cause was a Japanese network using public IPv6 addresses in a private context and the addresses leaked onto the public Internet. In Geoff's data, this explains much of the IPv6 darkspace traffic. Traffic actually destined for the dark unallocated IPv6 address space was small, caused by misconfigurations, DNS typos and oddities.

Casey Deccio (Sandia National Labs), in collaboration with Geoff Huston (APNIC), described lessons learned while setting up an IPv6 darkspace monitor. Sandia hosted a collector and announced an IPv6 route for a mostly (but not completely) unallocated APNIC address range. Casey stressed that the administrative effort for announcing the address space can be immense and requires coordination between originating AS, ISP, and ISP peers. He collected six weeks of data starting April 2012, for four weeks before and two weeks after announcing the address space. Before the announcement he observed 600-1300 packets per day of mostly DNS requests from clients attempting to pull information from servers in zones with improperly advertised prefixes. After the announcement the traffic grew to 6-7 million packets per day of mostly ICMPv6 with some DNS and TCP packets. A breakdown of the DNS requests by originating addresses showed a heavy-tailed distribution, i.e. a few hosts make a large number of requests to the darkspace. He plans to further investigate this effect.

5. SHARING AND COMBINING

Manish Karir (DHS) gave an overview of the PREDICT project, which supports data sharing with researchers. The project coordinates a large data repository that links to a variety of data sets including BGP routing data, netflow traces, topology data and several darkspace data sets. One PREDICT data provider worked with a regional address registry (APNIC) to announce 16 of the last /8 address segments as darknets for a week in 2011 and collected all the traffic for later comparison. Both Merit and CAIDA share their darknet data with researchers though PREDICT. As of May 2011, PREDICT indexes and sponsors over 200 TB of darkspace data, which has been used to analyze scanning and worm propagation, and general pollution of an address space to assess its value for future use. Analysis of the same time period at five different darkspaces showed some synchronized activities, e.g., scanners targeting hosts in all five darkspaces.

John McHugh (RedJack LLC) reported results from a small /22 address segment in Halifax, Nova Scotia, Canada which contained 899 dark addresses. He recorded 14 months of Netflow V5 data between February 2005 and March 2006, during which he observed 2.5GB of

traffic, of which 90 MB was directed to dark addresses. The traffic to the dark addresses was a mix of TCP (4M flows), UDP (1M flows) and ICMP (500K flows). Traffic was destined to many TCP port numbers, only some of which were associated with known services or vulnerabilities. All the unsolicited UDP traffic he observed was directed to hosts that existed at some point in the past, consistent with this network not being entirely dark. He described a "contact surface" of the observed traffic that shows the number of sources and destinations in a given time interval. He explained what seemed to be a heavy-tailed distribution as an interactive effect of three processes: low frequency traffic (many sources to few destinations); normal traffic; and scanners (few sources sending to many destinations). He and co-workers examined traffic to 20 dark /16 address ranges from several /8 networks and found that the number of unique source IP addresses (sending TCP SYNs) was quite similar across the ranges, regardless of whether the segment was dark or had some active hosts. John stressed that more research should be dedicated to darkspace traffic analysis, especially long-term studies.

Markus De Shon (Google) introduced flow collection activities and darkspace analysis at Google, noting that any darkspace at Google is likely only temporarily dark. He used Tanja's entropy method (see above) to detect some backscatter, scanning behavior and misconfigurations. In some cases entropy was not sufficient to precisely identify the events and he recommended complementary information such as a summarization of TCP flags. He is considering integrating darkspace analysis into other near-real-time flow processing and also analyzing IPv6 darkspace traffic. Markus briefly discussed data sharing challenges at Google, where even source anonymization may not protect all user data. Sharing highly aggregated forms of data is more likely at Google.

Brian Trammell (ETH Zurich) presented a data sharing architecture developed in the EU project DEMONS [1 7]. The DEMONS project uses a decentralized approach for collecting and analyzing data, where aggregation and analysis occurs close to data capture. Such early aggregation increases the scalability and privacy-sensitivity of the analysis infrastructure, but reduces the utility of the data, since only intermediate or final results get shared. Brian introduced *Blockmon* [8], an implementation of a modular data analysis approach in the DEMONS framework. *Blockmon* provides a platform to compose measurement and analysis functions using well-defined modules for filtering, metric calculation, correlation and other computations. Blocks exchange messages (such as packet or flow data) and use the IPFIX [2] format to share data. Challenges abound, since at the moment researchers lack not only a common set of tools and formats but also a common vocabulary to describe analysis functions. Brian also introduced the SEPIA software [3] developed at ETHZ for secure multiparty computation. The SEPIA framework allows one to aggregate operations on data from multiple sources without using a trusted third party. It is useful for limited types of interdomain data exchange.

Claude Fachkha (NCFTA Canada & Concordia University) presented ongoing work on profiling darkspace traffic and correlating threats. They analyzed packet traces from several /16 networks that received approximately 1200 packets/second. They found the largest amount of traffic originated from China, Russia and Korea, and 44% of the traffic originated from Microsoft Windows ^{TM}machines. To analyze correlations among threats they ran the *snort* intrusion detection software on the darkspace data, which mainly detected scanning activity [10]. They also used other data-rule mining to find correlations among threats.

6. DARKSPACE SHARING

Erin Kenneally (CAIDA) led a discussion on darkspace data sharing. She presented challenges to defining practical guidelines to share data, noting the difficulty in establishing attack risks to seemingly de-sensitized data. Furthermore, the heterogeneity of the data and interactions between policy and technology make it difficult to define suitable guidelines. She described a reference framework for data sharing that provides guidance on technical controls and enables risk-sensitive data sharing for data producers and consumers [6]. She posed the following discussion questions:

1. What major factors drive your decision to collect and share network data?
2. Do you feel you understand of the risks (legal, contractual, etc.) of sharing network data?
3. Do you feel you have a strong understanding of the available controls for mitigating those risks (both technical and policy)?
4. What (if anything) would motivate you to collect and share more network data with the research and operational community?
5. What should a Data Disclosure Best Practices Guide include to improve data sharing?

Participants pointed out that incentives for sharing differ for researchers and companies. Researchers need to share data to meet publication goals and to ensure reproducibility of results. In contrast, companies often fear that data sharing will result in competitive disadvantages or negative legal consequences. The discussion highlighted that often the establishment of bilateral, individual trust relations are needed to enable data sharing between organizations. The varying legal contexts across the world further complicate data

sharing with researchers who span legal regimes. For example, the E.U. and the U.S. have different models for understanding what kind of data must be protected. Workshop participants acknowledged that in some ways sharing darkspace data is less risky because it only contains unsolicited traffic, but it also contains IP addresses of vulnerable attack victims and possibly payload, both sensitive data that need some type of disclosure control.

7. COLLABORATIONS

Workshop participants used meal and break times to initiate collaborations, and requested to keep the DUST mailing list active to share information and coordinate cooperative darkspace activities. Developers of darkspace analysis tools have already begun to collaborate on analysis methods and tools, and are exchanging lessons learned from implementations. We include below those collaborations reported to us shortly after the workshop. There was consensus on the utility of a follow-up workshop next year.

Xenofontas Dimitropoulos, Eduard Glatz, and Brian Trammell (ETH Zurich) with Alberto Dainotti and Alistair King (CAIDA) started discussions of parallels between Corsaro and Blockmon and of the comparison of one-way unsolicited traffic observed on a darknet with the same on a live network. The group plans to collaborate on identifying and understanding potential differences and on devising approaches for sharing and correlating data of unsolicited traffic observed at multiple monitoring locations.

Geoff Huston (APNIC) and Markus De Shon (Google) discussed an existing collaboration between the two organizations to analyze 1.0.0.0/8 traffic. Geoff also spent time with Casey Deccio (Sandia National Laboratories) to discuss the IPv6 darkspace monitor on which they collaborate.

John McHugh (RedJack) and k claffy (CAIDA) discussed long term analysis of realtime darkspace traffic feeds with possible (lagged) comparisons with other darkspace collections.

8. WORKSHOP PARTICIPANT LIST

Workshop participants included: Emile Aben (RIPE NCC), Nevil Brownlee (University of Auckland), Jeffrey Cubbal (DHS), Casey Deccio (Sandia National Laboratories), Xenofontas Dimitropoulos (ETH Zürich), Claude Fachkha (NCFTA Canada & Concordia University), Eduard Glatz (ETH Zrich, Communication Systems Group), Geoff Huston (APNIC), Manish Karir (DHS), Douglas Maughan (DHS S&T), John McHugh (RedJack, LLC), David Plonka (University of Wisconsin - Madison), Markus De Shon (Google), Darren Shou (Symantec Research Labs), Brian Trammell (ETH Zürich), Joanne Treurniet (Defence R&D Canada - Ottawa) - remote participant, Shouhuai Xu (University of Texas at San Antonio), and Eric Ziegast (ISC).

CAIDA participants included: kc claffy, Tanja Zseby, Alberto Dainotti, Erin Kenneally, and Alistair King.

Acknowledgments

The workshop was sponsored by NSF CRI CNS-1059439 "A Real-time Lens into Dark Address Space of the Internet" and DHS S&T contract D07PC75579 "Supporting Research and Development of Security Technologies through Network and Security Data Collection."

9. REFERENCES

[1] "FP7-DEMONS.eu: DEcentralized, coopertive and privacy-preserving MONitoring for trustworthinesS", 2012. `http://www.fp7-demons.eu`.

[2] B. Claise et al. "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information". RFC 5101 (Proposed Standard), January 2008. `http://www.ietf.org/rfc/rfc5101.txt`.

[3] Martin Burkhart, Mario Strasser, Dilip Many, and Xenofontas Dimitropoulos. "SEPIA: Privacy-Preserving Aggregation of Multi-Domain Network Events and Statistics". In *19th USENIX Security Symposium*, August 2010.

[4] Matthew Ford, Jonathan Stevens, and John Ronan. "Initial Results from an IPv6 Darknet". In *Proceedings of the International Conference on Internet Surveillance and Protection*, ICISP, 2006.

[5] Bradley Huffaker. "Cuttlefish: Geographic Visualization Tool", 2006. `http://www.caida.org/tools/visualization/cuttlefish`.

[6] Erin Kenneally and Kimberly Claffy. "Dialing Privacy and Utility: A Proposed Data-sharing Framework to Advance Internet Research". *IEEE Security and Privacy (S&P)*, July 2010.

[7] S. Niccolini, F. Huici, B. Trammell, G. Bianchi, and F. Ricciato. "Building a Decentralized, Cooperative, and Privacy-Preserving Monitoring System for Trustworthiness: the Approach of the EU FP7 DEMONS Project [Very Large Projects]". *Communications Magazine, IEEE*, 49(11):16 –18, november 2011.

[8] Andrea Di Pietro and Nicola Bonelli. "BlockMON: A Modular System for Flexible, High-Performance Traffic Monitoring and Analysis", 2012. `https://github.com/blockmon/blockmon`.

[9] D. Plonka and P. Barford. "Flexible Traffic and Host Profiling via DNS Rendezvous". In *Proceedings of the Securing and Trusting Internet Names Workshop (SATIN 2011)*, April 2011.

[10] Martin Roesch, Chris Green, and Inc. Sourcefire. snort, 1998-2012. `http://www.snort.org/`.