impact

Cyber **Trust**

Annual Digest

November 2017

1. Introduction

Data and analytic tools enable principled science. In a domain like cybersecurity, where threats are dynamic and risks interrelated, effective research and development (R&D) demands a data sharing mechanism that addresses the operational, legal and administrative challenges to responsible innovation. Since an objective notion of truth is defined by the system of one's limitations, what is our collective truth about the security, stability, and resilience of our systems and communications infrastructures where cyber threats are more certain and prolific than their purported defenses? Vigorously stirring Alt-Facts and Post-Truth into a morass of Groundhog-Day discussions of problems and solutions provides few answers about how to move toward effective solutions in our information economy. IMPACT, the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (https://www.dhs.gov/csd-impact; https://www.impactcybertrust.org), is a DHS Science & Technology program aimed at improving our collective truths about cyber security risk to ultimately enhance trust in our solution paths.

The open secret of cyber security is that empirical data and effective analytics are fundamental to high quality R&D; reliable security decisions do not emanate from low-quality or missing data. Consider the cover of the Economist (May 6, 2017), which claims "The World's Most Valuable Resource is No Longer Oil, But Data." If scarcity is an indicator of value, we need look no further than the critical short supply of data for cyber risk R&D; data and analytics foundational elements are essential to develop advanced knowledge and to accelerate design, production, and evaluation of next-generation cybersecurity solutions. However, the value of having research infrastructure that enables real-world, large-scale, timely, and longitudinal data collection, sharing, and analysis is severely underestimated. Too often, such capabilities are assumed to exist without deliberate resource affordances. The result has been a scarcity of data from both industry and the government that is available to the open academic research community for innovation, reuse, and plain old truth-setting. IMPACT is an R&D resource uniquely championed with the support of the DHS Cyber Security Division.

IMPACT enables, sustains, and mediates the provisioning of freely available cybersecurity data and analytics between providers and seekers within the global industrial, academic, and government cybersecurity communities. It lowers the barrier to entry for cybersecurity R&D by addressing the operational, legal, and administrative costs that otherwise impede scalable, sustainable, and responsible data-sharing that underpins valuable and innovative cybersecurity R&D. IMPACT reduces the time and cost to find, acquire, curate, and store data in a manner that is mindful of the associated legal and ethical risks.

This IMPACT Digest is a sampling of recent outputs and outcomes enabled by IMPACT to expand awareness about and engagement with our current and future efforts to inform policy and the analysis of cyber-risk and trust. This Digest is organized along the core dimensions of the IMPACT mission: 2. Data and Analytics, 3. Tools, 4. Trust, and 5. Impact.

2. Data

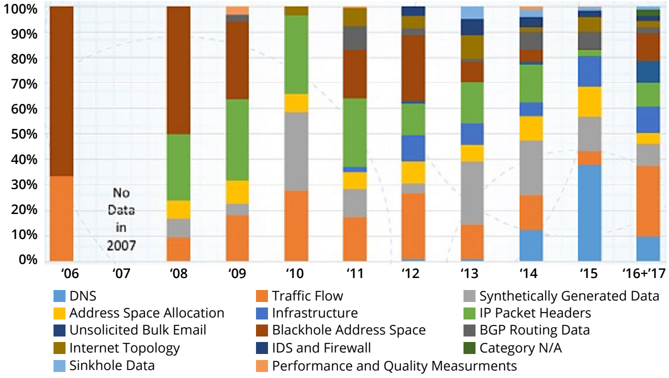


Figure 1: Data Trends; DHS IMPACT Program, SRI Analysis April 2017

2.1 Unsolicited Background Traffic Data

Needs Addressed: Researchers need data to study security and stability-related events including macroscopic connectivity disruptions, trends in malware propagation, and spoofed source-address denial-of-service attacks.

Approach Taken: CAIDA collected Blackhole address-space traffic data by monitoring routed but unused Internet Protocol (IP) address space that does not host any actual networked devices (e.g., hosts or routers). Systems that monitor unoccupied address space have a variety of names including darkspace, darknets, network telescopes, blackhole monitors, sinkholes, and background radiation monitors. Packets observed with such instrumentation can originate from a wide range of security-related events such as scanning for vulnerable targets, backscatter from spoofed denial-of-service attacks, automated spread of Internet worms or viruses, etc. Because unsolicited traffic may incidentally contain information about Internet hosts that are compromised or misconfigured, datasets in this category may be subject to specific disclosure control requirements.

Researchers at MERIT used a different approach; they analyzed traffic captured from MERIT's darknet to obtain insight into Internet-wide scanning activities in 2016. Their analysis defined a source IP address appearing in the darknet to be a scanner if that address contacted 25 unique destination IPs over a five-minute interval at the same destination port and protocol. (While the definition of scanning can often depend on context, this is the Bro default definition for scanners). Due to the dataset volume, MERIT adopted a sampling approach and analyzed only the first day of each month between Jan. 2016 and Oct. 2016.

Table 1: Top-20 Services Scanned

Port #/Protocol # (%)	Description (Service Name)
23/TCP (60.18%)	Telnet
53413/UDP (9.03%)	Vulnerability scan on Netis routers
No port /ICMP Ping (2.32%)	ICMP Ping
80/TCP (2.29%)	HTTP
3389/TCP (1.09%)	Microsoft RDP
2323/TCP (0.91%)	Mirai (Botnets scanning for IoT devices)
445/TCP (0.80%)	SMB-IP (Microsoft-DS Active Directory)
22/TCP (0.70%)	SSH
2222/6 (0.39%)	Ethernet/IP or DirectAdmin Remote Access
81/TCP (0.33%)	TorPark Onion Routing
8000/TCP (0.29%)	Radio streams such as iTunes Radio, DynamoDB Local
91/TCP (0.27%)	SG Security scan
21/TCP (0.20%)	FTP
443/TCP (0.16)	HTTPS
8123/TCP (0.14%)	Unknown (can be used for Web proxy)
8080/TCP (0.12%)	FilePhile Master/Relay over TCP
53/UDP (0.12%)	DNS
1080/TCP (0.11%)	SOCKS proxy
8888/TCP (0.10%)	HyperVM/ Freenet/ MAMP over TCP
3128/TCP (0.10%)	Squid Caching web proxy

The top twenty services scanned are tabulated in Table 1. Notice the volume of Telnet scanning.

The increase in Telnet scanning can be attributed to malicious activities at the time scanning for vulnerable IoT (Internet-of-Things) devices. Careful examination of the Mirai botnet source code[1] revealed that it was scanning for TCP port 23 (Telnet) and TCP 2323. The increase in darknet activities regarding these two ports is evident in Figure 2 and in the temporal summary contained in Table 2.

Table 2: Temporal Comparison of all Darknet Activities (not just scanning)

2004 [2]	2010 [3]	2014 [4]	2016 (this study)
HTTP (80)	SMB-IP (445)	SMB-IP (445)	Telnet (23)
NetBIOS (135)	NetBIOS (139)	ICMP Ping	UDP 53413 (Netis)
NetBIOS (139)	eMule (4662)	SSH (20)	HTTP (80)
DameWare (6129)	HTTP (80)	HTTP (80)	ICMP Ping
MyDoom (3127)	NetBIOS (135)	RDP (3389)	SSH (20)

Table 3: Top Countries Associated with Scanning

United States (21.68%)	Taiwan (3.63%)
China (10.16%)	Netherlands (1.91%)
Brazil (6.98%)	Turkey (1.77%)
Vietnam (5.25%)	Romania (0.57%)
South Korea (4.47%)	Russia (0.43%)

Table 4: The Top-20 Scanners Originate from These Providers (ASes)

Akamai	QUASINetworks	SingleHop, Inc
CariNet, Inc.	Steadfast	WELLPOWER-TW

Table 3 identifies countries associated with scanning activities (information obtained from the geolite2 MaxMind database). In 2016, Telnet ranks at the top, and activity for TCP 2323 made the top 10 (Table 2). Table 4 lists the providers associated with autonomous scanning systems (information obtained from "Whois" data).

Resulting Benefits: Darknet analysis provides valuable insights to network operators regarding trends in global scanning activities (associated with active cyber-threats), denial of service attacks, network outages, and other threats. The results of the analyses enable further activities for situational awareness and attribution, which enhance the security posture of both public and private organizations.

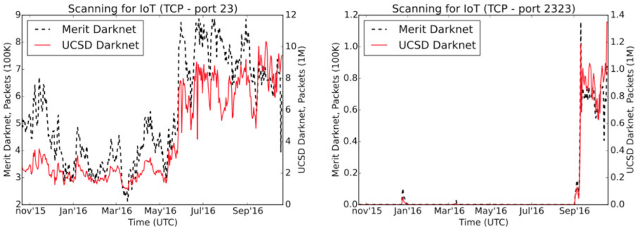


Figure 2. Darknet traffic to TCP ports 23 and 2323, as recorded by two large darknets, namely the Merit and the UCSD darknets. Data source: Merit and UCSD darknets. Graph data obtained from CAIDA's Charthouse.

2.2 Internet Addresses and Topology Data

Exchanging information over the Internet requires computers on both ends with IP addresses. Almost all servers on the Internet have their own addresses, although client computers sometimes share an address using network address translation (NAT). All commercial services today require addresses using IP version 4 (IPv4). IPv4 addresses are a limited resource; there are only 4 billion of them. In 2011, IANA (the organization managing global addresses) allocated the last open IPv4 addresses, and by 2015, all regional registries except one have exhausted their pools.

[1] Mirai was responsible for some of the largest DDoS attacks ever recorded, including the attacks against KrebsonSecurity and Dyn.

Marketplaces that trade IPv4 addresses are raising questions about imposing technical restrictions on what can be traded. Completely unfettered trading could fragment addressing and increase costs. IPv6 exists, and its use is increasing, but concerns include unanswered questions regarding how the cost of supporting and deploying IPv6 compares to more careful management of IPv4. IPv4 addresses affect factors besides cost, for example, address usage and density affect worms and port scanning. These issues open other unanswered questions:

- *How effective is scanning by botnets like Mirai?*
- *Can we estimate botnet size and capability by the probes we see?*
- *What about millions of new mobile phones today?*
- *What about millions more IoT devices tomorrow?*

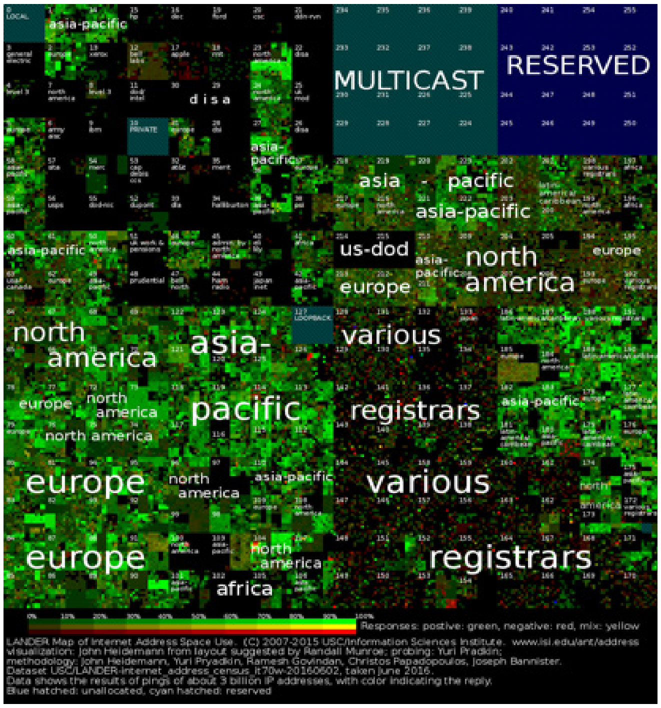

asia-pacific
MULTICAST
RESERVED
disa
asia-pacific
asia - pacific
asia-pacific
us-dod
north america
europe
europe
north america
asia-pacific
various
pacific
registrars
europe
north america
north america
asia-pacific
europe
north america
asia-pacific
various
europe
north america
africa
registrars
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Responses: postive: green, negative: red, mix: yellow
LANDER Map of Internet Address Space Use. (C) 2007-2015 USC/Information Sciences Institute. www.isi.edu/ant/address
visualization: John Heidemann from layout suggested by Randall Munroe; probing: Yuri Pradkin;
methodology: John Heidemann, Yuri Pryadkin, Ramesh Govindan, Christos Papadopoulos, Joseph Bannister.
Dataset USC/LANDER-internet_address_census_it70w-20160602, taken June 2016.
Data shows the results of pings of about 3 billion IP addresses, with color indicating the reply.
Blue hatched: unallocated, cyan hatched: reserved

Figure 3.A map of the IPv4 address space. Dataset: https://www.impactcybertrust.org/dataset_view?idDataset=621.

Clearly, effective planning of this important global resource requires high-quality, objective, and public insight into current allocation and activity trends.

Approach Taken: In 2003, researchers at ISI started collecting data about the Internet (IPv4) address space, and regularly probe all addresses in the allocated Internet address space, more than 3 billion. Figure 3 shows their most recent map of Internet addresses. Internet addresses are allocated in blocks of adjacent addresses, and these blocks turn into squares of different size in this map. Addresses usually are numbers like 192.0.2.0; there are about 4 billion addresses, from 0.0.0.0 to 255.255.255.255. Addresses follow the Hilbert fractal, so numerically close addresses are physically close on the map.

In the map, brighter areas indicate more replies, darker, fewer, with color indicating positive (green) or negative replies (red), and black showing areas that do not reply.

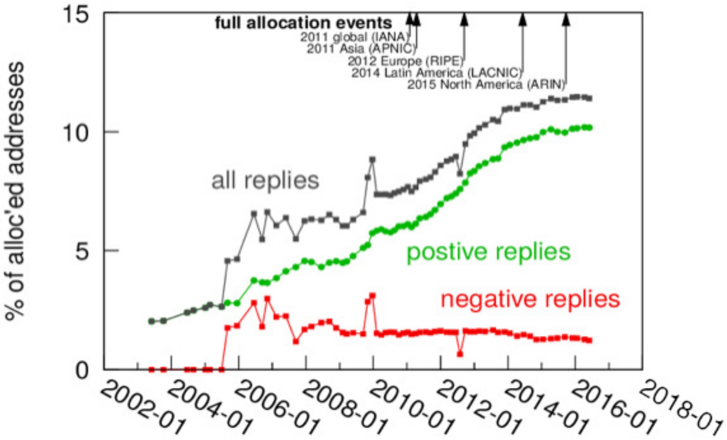


Figure 4. Address usage over time as seen by ISI's probing.

The cyan hatched areas are not studied, typically because they are reserved or private. With a consistent, careful methodology, our data provides a long-term longitudinal view of changes in IPv4 address use.

Our data (Figure 4) gives an estimate of the instantaneous size of the Internet. Scaling replies by our correction factor of 1.6 to 1.9. Our best estimate is that about 676 to 803 million addresses are active at any instant. We used the correction factor from "Census and Survey of the Visible Internet," by Heidemann et al., ACM IMC 2008.

No census of billions of addresses will be perfect, but we correct for addresses we cannot directly measure with statistical estimation. Our peer-reviewed study describes sources of under-counting: A few percent of probes and replies are lost due to congestion. Some addresses, such as those behind firewalls, chose not to receive or reply to our requests. Other computers use private addresses.

Table 5: Instantaneous size of the Internet (est.) Basic data from https://www.impactcybertrust.org/dataset_view?idDataset=621 in June 2016.

What	Addresses
IPv4 addresses:	4,294,967,296 (100% all)
..unprobed:	588,972,032 (13% all)
....special (multicast, etc.):	587,202,560 (13% all)
....unallocated:	1,769,472 (0% all)
..probed:	3,705,995,264 (86% all)
....replies:	422,662,606 (9% all)
......positive:	377,166,966 (8% all)
......negative:	45,488,476 (1% all)
....non-replies:	3,283,332,658 (76% all)

After we correct for under-counting, our measurements represent the most accurate current snapshot estimating IPv4 address usage.

Dynamic addresses are another source of change in the number of addresses that respond. Our census gives a virtual snapshot of how many addresses are active at any instant, on average. In some countries and regions, however, this number varies up or down by 20% over the course of a day. (For more information, see “When the Internet Sleeps: Correlating Diurnal Networks with External Factors,” by Quan et al., ACM IMC 2014).

While the previous numbers capture instantaneous snapshots, far more addresses are used when considering an entire day or multiple days. CAIDA added data to ISI’s IPv4 2013 Census data set that elucidates the IPv4 address-space usage. CAIDA cross-correlated results of analysis of passive and active measurements to taxonomize 24 address blocks as IETF-reserved, used, routed-unused, unrouted-assigned, and available. These datasets include both raw and curated forms of topology data gathered from across the global Internet. Typically, this data is obtained by deploying traceroute-like probes from monitoring points around the network. Raw IP topology data can include IP addresses on machines that a packet traverses along the forward path to a target destination, allowing heuristic-based inference of router-level and AS-level topologies. Some topology datasets are already curated into router-level or autonomous system-level topologies for ease of researcher use. Datasets in this category support modeling and simulation of malware outbreaks, spread, distribution, containment, and countermeasures; macroscopic vulnerability assessments; longitudinal analysis; and modeling of Internet evolution.

Resulting Benefits: Our data provides a unique, long-term view of Internet use with several applications:

- *Work with census and aggregated datasets helps inform address usage. Our long-term study shows that address usage is growing, but overall use is still low (less than 10%). We could improve the management of the IPv4 address space, but better management comes with a cost.*
- *Censuses help improve studies of Internet topology. We use censuses to build hitlists, a comprehensive list of addresses to probe for good coverage of the Internet. Multiple research groups (71 different researchers) use these hitlists to conduct new research studies about Internet topology.*
- *Internet census-taking has inspired new approaches. Based on census findings, USC/ISS have developed techniques to estimate network outages around the globe, and have shown how diurnal shifts in Internet address usage vary from country to country, reflecting the maturity and policies in each country’s part of the Internet. CAIDA has aggregated additional datasets, showing that passive and active measurement approaches can produce a fuller picture of long term Internet address use.*
- *The USC/ISI censuses have inspired faster versions. Multiple research groups (Michigan, MassScan by Graham et al.) have focused on conducting censuses as fast as possible. While our censuses are intentionally paced at a slow rate to minimize concerns, fast probing has a place in security studies.*
- *Data from the USC/ISI censuses have been used by others. Since 2006, 71 different researchers have acquired copies of some or all our censuses and related datasets, over 53TB of data. This data has been used in dozens of papers and follow-on studies.*

2.3 Malware Data

Needs Addressed: Malicious software is a centerpiece among current threats on the Internet. It is used to create botnets that, for example, generate unsolicited email, conduct DDoS attacks, and steal sensitive financial information and intellectual property, fueling the intentions of criminals and nation-states alike. Efforts to understand and defend against malicious software are therefore critical to cyber-security research and applied defense. These efforts, which include cyber-threat discovery, compromise detection, and asset remediation, require the ability to observe and study current malware behavior.

Major information security organizations collect over 100,000 new malware samples each day from spam traps, web crawlers, user submissions, and malware exchanges. Each sample holds actionable network information that, once derived via a dynamic analysis sandbox, has both research and operational utility. However, the sensitive nature of malware, the resources required to process the substantial volume of new samples that appear each day, and the commoditization of anti-analysis techniques leave many researchers and practitioners with an ongoing, unmet need for this data.

Figure 5. GTISC 2016 Malware DNS Wordle.

Approach Taken: Through DHS IMPACT, the Georgia Tech Information Security Center (GTISC) leverages its extensive malware collection and analysis experience to facilitate the availability of real-world, large-scale malware network datasets. In providing such data to approved requestors vetted by the IMPACT Coordinating

Center, GTISC and DHS fill a research and operational gap for those who could not otherwise access this data.

GTISC malware network datasets enable the study's core Internet-facing sample behaviors, including the use of email (for propagation, or to advertise the sale of illicit goods and services) and interactions with the Domain Name System (DNS) (for rendezvous with attacker command and control). Malware's use of the DNS has proven especially useful in cyber security research and defense. Figure 5 provides a wordle (word histogram) intended to serve as a visual explanation of this resource.

In Figure 5, domain names associated with higher counts of samples that queried for a domain are indicated by larger fonts. Benign domain names feature prominently because malicious software abuses the services hosted at them (e.g., malware accesses a search engine to perform clickfraud). Some malicious domain names also feature prominently because the malware instances that use them are both prolific and highly polymorphic, meaning that many programs, distinct by hash, are serial variants generated to make detection more difficult. Finally, some malicious domains (e.g., ns1.player1532.com) are named to masquerade as legitimate infrastructure (e.g., a nameserver).

Resulting Benefits: Large-scale malware network datasets offered by GTISC enable a host of cybersecurity activities ranging from academic research that examines network abuse and cyber threat evolution to improvements in the operational cyber defense of commercial entities such as major financial institutions and organizations that operate critical infrastructure. Over 150 entities spanning academia, industry, and government have made use of GTISC malware network datasets.[2]

2.4 Internet Latency Measurement

Needs Addressed: The performance of Internet services is intrinsically tied to propagation delays between end points (i.e., network latency). Standard active probe-based or passive host-based methods for measuring end-to-end latency are difficult to deploy at scale and typically offer limited precision and accuracy.

Approach Taken: Figure 6 shows the client locations for one day from the network time protocol (NTP) server located at UW-Madison. Our paper, "Times Forgotten: Using NTP to Measure Internet Latency," which appeared in ACM HotNets 2015, describes an investigation of a novel but non-obvious source of latency measurement: logs from NTP servers. Using NTP-derived data for studying latency is compelling due to NTP's pervasive use in the Internet and its inherent focus on accurate end-to-end delay estimation. We considered the efficacy of an NTP-based approach for studying propagation delays by analyzing logs collected from 10 NTP servers distributed across the US. These logs included over 73M latency measurements to 7.4M worldwide clients (indicated by unique IP addresses) collected for one day. Our initial analysis of the general characteristics of propagation delays derived from the log data reveals that delay measurements from NTP must be carefully filtered to extract accurate results. We developed a filtering process that removes measurements that are likely to be inaccurate[3]. After applying our filter to NTP measurements, we analyzed the scope and reach for US-based clients and the characteristics of the end-to-end latency for those clients.

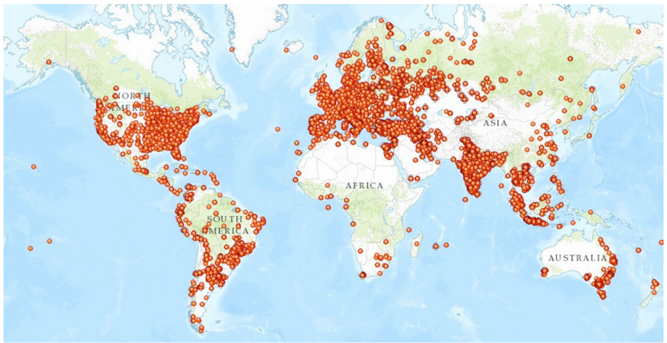


Figure 6. Client locations for one day from the NTP server located at UW-Madison

Resulting Benefits: Our findings show a range of behaviors that include latency to mobile hosts (vs. desktop systems), highlight the asymmetry of one-ways delays, and indicate opportunities to apply NTP-based latency measurements to a variety of additional problems.

3.Tools

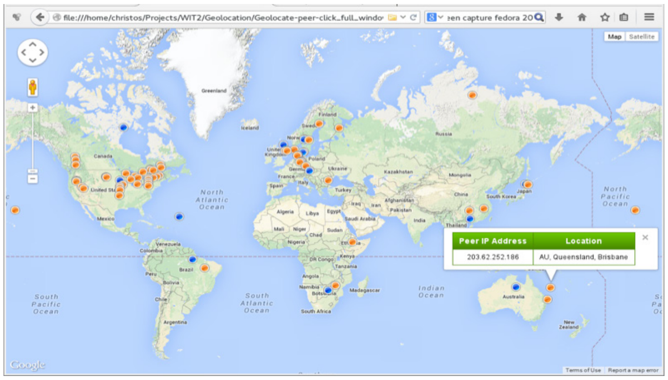
3.1 Monitoring the Internet Highways–BGPMon

Needs Addressed: The Internet today is so vast and dynamic that there is virtually no hope of monitoring it in its entirety. While this may be good for privacy advocates, it becomes a nightmare for network operators who want to fix things when they break. It also offers rich opportunities for malicious entities, who can inject faults and hijack traffic without being detected, weakening privacy and capturing sensitive information.

[2] A small set of example organizations is as follows: Academia- George Washington University, New York University, Dalhousie University (Canada), Edinburgh Napier University (UK); Industry- Symantec, Intel Security (McAfee), Trend, Bank of America, Wells Fargo, Capital Group Companies, Lloyds Banking Group (UK), ANZ Bank (Australia), Pfizer, Athena Health, Emdeon, DoD Cyber Crime Center, DARPA, DHS, RCMP (Canada), CSE (Canada), CCIRC (Canada), CERT Australia, DST (Australia).

[3] See, "Times Forgotten: Using NTP to Understand Internet Latency", In Proceedings of ACM Workshop on Hot Topics in Networks (HotNets), November, 2015 (https://dl.acm.org/citation.cfm?id=2834108)

Approach Taken: Fortunately, one approach appeases privacy advocates while foiling malicious plans to capture Internet traffic: we can monitor the entire Internet at the routing level. We can lay down a coarse map of the major streets and highways on the Internet without actually looking at the traffic on them, every few minutes. We can quickly detect changes in the because the Internet uses a routing protocol called Border Gateway Protocol (BGP) that tells every router how to direct traffic anywhere on the Internet. BGP is a verbose protocol, telling each router the next hop to send traffic and the series of networks that traffic will follow to get to that destination. This allows routers to implement policies and make local decisions about which paths they prefer.


file:///home/christos/Projects/WIT2/Geolocation/Geolocate-peer-click_full_windo

Peer IP Address	Location
203.62.252.186	AU, Queensland, Brisbane

Figure 7: The BGPmon eyes: IPv4 (orange) and IPv6 (blue)

Resulting Benefits: The BGPmon project (http://www.bgpmon.io) is one of the many projects contributing data to IMPACT. By leveraging data collection started by RouteViews about 15 years ago, BGPmon offers a continuous live feed of BGP information from about 400 points around the world. BGPmon is a quiet listener, it never talks back, so it is safe to deploy. The information it collects is time-stamped, geolocated, and collected in a fast database at Colorado State University, where it is made available in many convenient formats to the entire research community.

An important operational benefit of BGPmon, is its ability to detect route hijacks. A route hijack occurs when a router, accidentally or maliciously, announces a path to an unauthorized destination network. BGP cannot detect lying routers; the only defenses are the instincts, skills, and expertise of network operators. Sometimes these qualities fail, especially in networks with less experienced operators or when a trusted network makes a mistake. Such a mistake results in the diversion of traffic, which can cause a routing outage or a man-in-the-middle attack. Only the operators can remedy this situation; however, BGPmon can detect such incidents and notify the operators immediately so they can respond.

BGPmon offers more than just a stream of BGP messages to help detect hijacks. The BGPmon team has geolocated all the monitors, so when a hijack is reported, operators will know the prefix, time, and duration of the hijack and the geographical regions that were affected. The BGPmon team provides an AS geolocation service, so users can issue queries about the prefixes an AS advertises at a particular time, and where those prefixes geolocate. While geolocating Internet hosts is hardly an exact science, BGPmon provides a good approximation.

3.2 Mapping Internet Physical Infrastructure—Internet Atlas and BigFoot

Needs Addressed: Mapping and understanding the physical underpinnings of our Internet infrastructure helps identify shared risk and defend against adversarial and natural threats to our communications infrastructure. A large body of economic research has shown a strong correlation between broadband connectivity and economic productivity. These findings motivate government agencies, such as the FCC in the US, to provide incentives to Internet Services Providers to deploy broadband infrastructure in unserved or underserved areas.

Approach Taken: We developed a framework for automating the identification of target areas for network infrastructure deployment. Our approach considers infrastructure availability, user demographics, and deployment costs. We used multi-objective optimization to identify geographic areas that have the highest concentrations of unserved and underserved users and that can be upgraded at the lowest cost. To demonstrate the efficacy of our framework, we considered physical infrastructure and demographic data from the US and two deployment cost models. Our results identified a list of counties that would be attractive targets for broadband deployment from cost and impact perspectives. We validated our findings by comparing the results with the FCC's Connect America report. We believe that there are a variety of broader applications of our framework, which is now incorporated into the Internet Atlas portal.

The Internet Atlas project is focused on building and maintaining a repository of geocoded maps of Internet physical infrastructure. We define physical infrastructure as nodes (PoPs, co-location centers, IXPs, etc.) and links (fiber optic cables) that carry Internet traffic. The repository currently includes maps of over 1,100 networks from around the world, which have been carefully audited over the past year. The process of adding new maps to the repository is on-going. The repository is complemented by a GIS-based web portal, which enables data to be visualized and analyzed. The portal also enables a wide variety of Internet and related data to be imported and visualized.

Resulting Benefits: The Atlas repository enabled the generation of a first-of-its-kind map of the US long-haul fiber infrastructure[4]. The map is a composite of 20 US fiber providers and comprises 273 nodes and 542 conduits. Importantly, the details of the connectivity and shared use of conduits has been verified using public records of rights-of-way.

3.3 BigFoot: Visualizing BGP Update Anomalies

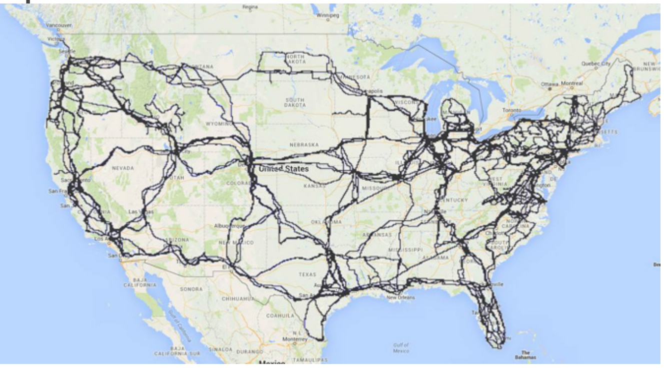


Figure 8. Map of the Internet's long-haul infrastructure in the US[5].

Needs Addressed: Studies of inter-domain routing in the Internet have highlighted the complex and dynamic nature of connectivity changes that take place daily on a global scale. The ability to assess and identify normal, malicious, irregular, and unexpected behaviors in routing update streams is important in daily network and security operations.

Approach Taken: BigFoot is a new tool for BGP update visualization that is designed to highlight and assess a wide variety of behaviors in update streams. BigFoot's core notion is visualizing the announcements of network prefixes via IP geolocation. We investigated different representations of polygons for network footprints and showed how a straightforward application of IP geolocation can lead to representations that are difficult to interpret. BigFoot includes techniques to filter, organize, analyze, and visualize BGP updates that enable the effective identification of characteristics and behaviors of interest. To assess BigFoot's capabilities, we considered 1.79B BGP updates collected over a period of one year, and identified 139 candidate events in this data. We investigated a subset of these events in detail, along with ground truth from existing literature, to understand how network footprint visualizations can be used in operational deployments. Figure 9 highlights that hijacks and other unwanted BGP behaviors, which can be challenging to detect using other methods, are easily distinguished using BigFoot.

Resulting Benefits[6]: BigFoot is deployed and available in the Internet Atlas portal (Figure 9). The Internet Atlas maintains a 7-day rolling window of BGP updates from the BGPmon project (Colorado State University). BigFoot visualization and analysis of these updates are available through the Internet Atlas portal.

4. Operationalizing Trust

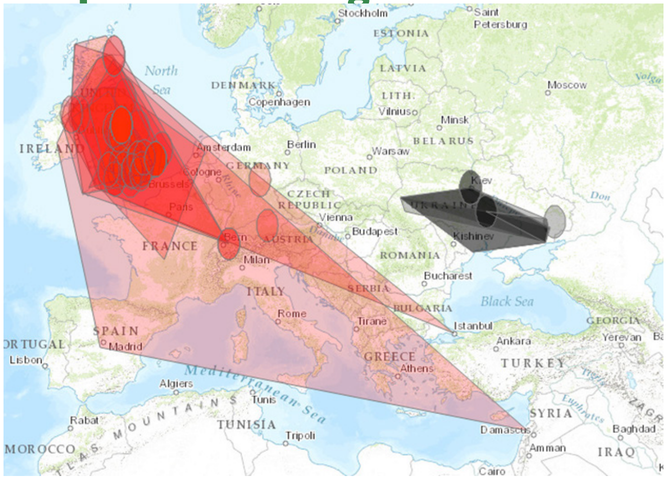


Figure 9. Example of BigFoot highlighting a BGP highjacking event from 2015

4.1 IMPACT Core Components

The IMPACT ecosystem consists of four components supporting core functional requirements for data sharing to support R&D: metadata discovery (FIND & CONNECT), data and tool matchmaking (GET & USE), and a social forum (FEEDBACK LOOP). The fourth component is the business rules broker that is the policy middleware enabling the other three components.

The metadata catalog lowers the barriers to finding data by providing a centralized index of available data and a profile of the data to enhance research efficiency. The data match-making component lowers the barrier to acquiring data by connecting the two-sided market, data and analytics supply and demand, that is, the providers and seekers. The tool match-making component lowers barriers to using data by offering a federated network of resource providers, who act as a virtual laboratory by provisioning data and analytics to enable research needs.

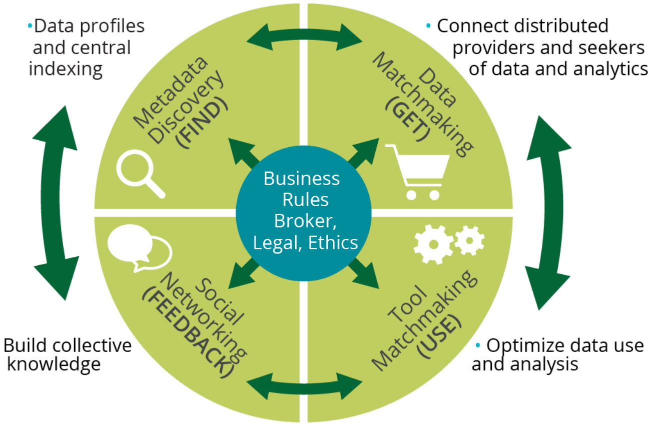
[4] An example of the utility of the Atlas repository was highlighted in the paper, "A Techno-Economic Framework for Broadband Deployment in Underserved Areas," which appeared in Proceedings of the ACM SIGCOMM Global Access to the Internet for All (GAIA) Workshop," in August, 2016.

[5] Details of the generation and analysis of this map can be found in the paper, "Intertubes: A Study of US Long Haul Infrastructure," which was published in proceedings of ACM SIGCOMM, August, 2015.

[6] The utility of this approach to research and network operations can be found in [cite BigFoot paper and hyperlink - Proceedings of IEEE VizSec '16]

To capitalize on the collective knowledge of the marketplace, the forum enables social networking and creates a critical feedback loop between the two-sided market. This feature allows users to engage with the collective IMPACT community to determine the data that will assist specific R&D, suggest new data to suit dynamic R&D needs, and take part in collective analyses and knowledge building.

Finally, the business rules component mediates the other dimensions by streamlining request processes with standardized legal agreements and risk assessments. The result is an ecosystem that addresses the major challenges to data sharing (operational costs, administrative overhead, and legal risk) to enable effective cyber security R&D:



•Data profiles and central indexing
• Connect distributed providers and seekers of data and analytics
Metadata Discovery (FIND)
Data Matchmaking (GET)
Business Rules Broker, Legal, Ethics
Social Networking (FEEDBACK)
Tool Matchmaking (USE)
Build collective knowledge
• Optimize data use and analysis

Figure 11. IMPACT Core Components

4.2 IMPACT Return on Investment (ROI)

Ground truth data about the security, stability, and resilience of our 16 critical infrastructures is an obvious prerequisite to understanding deficiencies and developing solutions. Just as measuring ROI for cyber security is challenging (given that its value lies in cost reduction rather than revenue produced), capturing the value of IMPACT requires thinking beyond first-order financial calculations. However, the benefits produced by IMPACT are manifold for all HSE stakeholders: DHS, other government entities, industry, and academia. These benefits include:

Parity. IMPACT tackles the 90:10 percent, data-rich:data-poor problem that permeates R&D by lowering the barrier to entry on the demand side and for producers of valuable data who have neither the ability nor the resources to provision data directly. As a result, IMPACT extends the lifespan of data so that its value can be exploited by the collective of stakeholders in a readily usable manner.

Scale and Sustainability. To date, successful data sharing models have largely been ad hoc, relying on interpersonal relationships, often resulting in trading packets under the table to avoid legal tangles. By creating a federation of data providers mediated by standardized processes and business rules, IMPACT has enabled a uniform, repeatable, and trusted interface between seekers and providers. IMPACT has gone beyond the precedent and network-effects requirements, on which attorneys rely to approve data sharing. By commoditizing this middleware, IMPACT can engage new and different endpoints to avoid data sharing start-up costs that can interfere between entities with no prior relationship.

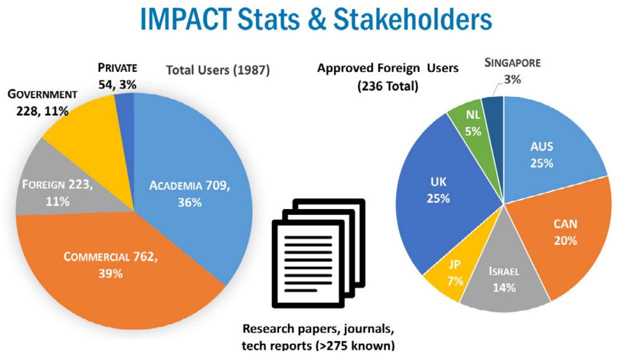


Figure 12. IMPACT Users

Utility. Rather than starting from risk aversion, IMPACT leads by questioning the value the data will provide, and then mitigating any associated risk via disclosure-control approaches. Just as cyber security is anchored in risk management, so too is data sharing to support it.

Scope. IMPACT's R&D-enabling infrastructure supports the broad DHS/S&T CSD research agenda by providing data, tools, and analytical capabilities to other CSD researchers and their HSE customers. IMPACT has a footprint in seven international locations and is poised to expand to new countries.

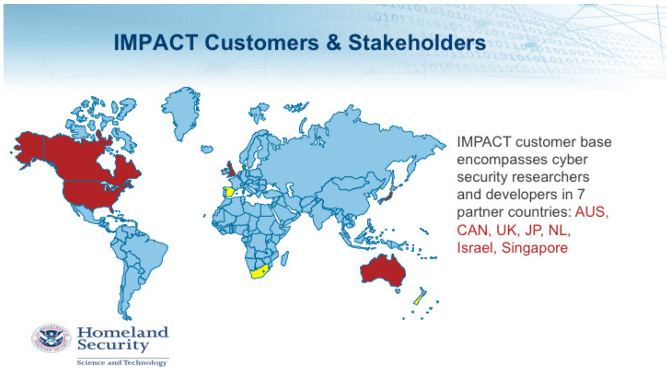
Leadership. Because of the unprecedented and sustained support for this communal resource, the high-level benefit of the IMPACT project to DHS/S&T, CSD is recognized within government, industry, and academia as a leader in providing research infrastructure to the cyber-security R&D community. Despite widespread talk about the need for data sharing to support cyber security, IMPACT is one of the few models to successfully operationalize data sharing for R&D.

Trust. Perhaps the most valuable, but intractable asset that IMPACT introduces is a trusted ecosystem to exploit the value of federated data. This has been accomplished in several dimensions:

- *Vetted data, researchers, and providers to assure legitimacy*
- *Balancing the efficiency sought by researchers against the certainty sought by risk overseers*
- *Accountability via a stable and time-tested legal infrastructure with ethical oversight.*

Regarding ethics, a satellite aspect of IMPACT has focused on responsible innovation by setting ICT research ethics standards. IMPACT supports cyber-risk R&D by enabling data and tool sharing, and addresses the socio-technical layer that can impede R&D. IMPACT has been a leader in community-informed ethics and sensitive data-disclosure guidance that address the principles, controls, and responsible implementation of solutions to issues that impede cyber security research. The Menlo Report, Ethical Principles Guiding Information and Communication Technology Research and its Companion Report were the flagship ethics outputs. They were inspired to pre-empt some of the socio-technical issues coming down the pike, and to embrace a principals-in-context approach by applying the framework from the Belmont Report to modern information and communication technology research.

Current efforts are focused on Cyber-risk Ethics Decision Support (CREDS). The CREDS Tool is an applied research and development project intended to operationalize a decision support methodology, conceptual framework, and an interactive online tool to identify, reason, and manage ethical and legal issues related to cyber-based research (e.g., network and system security). It operationalizes the Ethical Impact Assessment (EIA) framework that was coined and commenced under the IMPACT ethics project. The objectives of the tool are to facilitate research that minimizes potential harm while enabling innovation, and to advance the collective dialogue between and among researchers, oversight entities and policymakers about research ethics principles and practices.


IMPACT Customers & Stakeholders

IMPACT customer base encompasses cyber security researchers and developers in 7 partner countries: AUS, CAN, UK, JP, NL, Israel, Singapore

Figure 13. IMPACT Footprint

The functional goals include estimating and communicating ethical risk; identifying potential impacts of technology; and measuring and improving judgment and reasoning. The methodology involves deriving principles and practices from established law, ethics and best practices, and then using that output to drive the underlying logic of the tool. The CREDS tool is intended to be a resource for the entire community to engage in repeatable and transparent decision-making to prevent diminished public trust and reputational blowback caused by association with undifferentiated comparisons to public or private surveillance and cyber opportunism. CREDS is currently in prototype development and is scheduled for release to the community in Fall 2017.

Stakeholder Engagement. In February 2016, CAIDA organized and hosted a two-day BGP Hackathon event for teams of researchers and students with interests in the development of tools to measure, monitor, and model the routing infrastructure of the Internet[7]. Sponsored by industry, professional organizations, and government agencies, the teams competed to develop practical solutions to live BGP measurements and monitoring challenges. All fifteen teams made significant progress in critical BGP areas including investigating anycast routing, automating detection of BGP anomalies (e.g., hijacking events), improving the BGPStream framework by adding customized filters, and developing informative visualizations of live BGP data.

5. Conclusion

Data are critical to R&D capabilities, and despite interpretation challenges with statistics, we can confidently conclude that impactful R&D is impossible without empirical data. Cybersecurity needs real-world data to develop, test, and evaluate knowledge and technology solutions to counter cyber threats. Big data may grow on trees, but it must be picked, sorted, and trucked. IMPACT addresses these needs and moves forward with the IMPACT vision. Decision analytics are critical to HSE capabilities: cybersecurity needs an integrated, holistic understanding of the risk environment for strategic interventions. IMPACT is addressing the noticeable gap between data and decisions by providing the multi-dimensional data, complex associations and fusion, and high-context presentation elements that will close that gap. Finally, data sharing and analytics are not easy; high value data often results in high legal risk and/or cost. IMPACT addresses the technologists' need to optimize for efficiency and the lawyers' need to optimize for certainty. Help us help you make an IMPACT. https://ImpactCyberTrust.org

[7] A report is available at http://caida.org/publications/papers/2016/bgp_hackathon_2016_report.