AFRL-RI-RS-TR-2021-176

DATASETS, METHODS AND TOOLS FOR INTERNET SECURITY DECISION ANALYTICS

UNIVERSITY OF WISCONSIN SYSTEM

OCTOBER 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2021-176 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
SCOTT ADAMS
Work Unit Manager

/ S /
JAMES PERRETTA
Deputy Chief, Information
Exploitation &Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
OCTOBER 2021	FINAL TECHNICAL REPORT	JUN 2017- MAY 2021

4. TITLE AND SUBTITLE

Datasets, Methods, and Tools for Internet Security Decision Analytics

5a. CONTRACT NUMBER
FA8750-18-2-0036

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)

Barford, Paul, Dr.

5d. PROJECT NUMBER
UWIS

5e. TASK NUMBER
20

5f. WORK UNIT NUMBER
18

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Wisconsin System
Suite 6401
21 North Park Street
Madison WI 53715-1218

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGB
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2021-176

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This grant focused on full-stack data distribution, service development and research. Over the performance period, the Internet Atlas portal received over 27K page views by over 16K unique visitors. 21 new user accounts for access to Internet Atlas were created, and 117 datasets provided. 15 research publications were published and 25 presentations made in academic and industrial forums. Research topics included network mapping, building a network outage detection system and a WWW panel system.

15. SUBJECT TERMS

Internet measurement, Internet physical layer topology, Internet maps, Internet risk analysis, Network Time Protocol, Internet outage analysis, web crawling, cross-layer analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**SCOTT F. ADAMS**
U	U	U	UU	19	19b. TELEPHONE NUMBER (*Include area code*) **NA**

Table of Contents

1. Summary

The University of Wisconsin (UW) was supported by the Department of Homeland Security (DHS) Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program during a period of performance from June 2017 to May 2021. Research and development was accomplished for Technical Topic Areas #1 Data Provider (DP) and #2 Decision-Analytics-as-a-Service (DASP). The approach to these activities was to collect and assemble data sets from different layers of the network protocol stack to offer unique and critical perspectives on cyber security risks, opportunities to reduce risk, user behavior, and forensic investigations of events of interest in cybersecurity research and for enterprise owners and operators.

Our data provider activities focused on extending and enhancing the Internet Atlas repository and portal, which were initially developed during the prior generation of the IMPACT / Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT) programs (see reference #22 in Section 8). In addition to Internet Atlas, the University of Wisconsin developed and distributed several other data sets including DShield Intrusion Detection System (IDS)/firewall logs, Border Gateway Protocol (BGP) update logs, Network Time Server logs and web crawl logs.

Our decision-analytics-as-a-service efforts focused on developing capabilities to identify Internet events (including outages, attacks, route changes, etc.) in real time based on Network Time Protocol (NTP) data collected from a set of 14 NTP servers that contributed data throughout the period of performance. We also worked on developing methods and tools to fuse data from different layers of the protocol stack to derive insights on performance, connectivity and risks that would not otherwise be possible. We also spent significant time on developing a system for collecting web browsing data from users who are willing to contribute data. We worked with the University of Wisconsin Institutional Review Board (IRB) and legal department to ensure that we had the proper authorization for these activities.

We distributed hundreds of data sets to the research community during the period of performance. For the period from June ‘17 through May ’21, the Internet Atlas portal received over 27K page views by over 16K unique visitors from all over the world. 21 accounts for detailed access to Atlas were provided. For the same period, 117 datasets were provided (primarily the Internet Long Haul infrastructure data) based on requests made through the IMPACT portal (impactcybertrust.org).

In addition to distributing data sets and developing decision-analytics-as-a-service capabilities our research efforts have resulted in 15 publications in high quality venues, 2 papers that were posted to the arxiv.org public archive and will be submitted for publication, and 4 additional manuscripts are in preparation and will soon be submitted for publication. The topics of these papers in preparation include an empirical analysis of the Domain Name System (DNS), a new method for identifying Internet connectivity based on end-to-end latency measurements, an analysis of the impact of power outages on Internet client availability in the United States (US) and a method for geolocation of the Internet routing hypergraph. A complete list of publications

that resulted from this grant along with arXiv papers and papers in preparation are provided in Section 8. We have given 25 technical presentations related to these research papers and our dataset. Finally, our Internet Atlas and associated maps and findings have been the subject of numerous articles in both technical and popular press.

Acknowledgement: This material is based on research sponsored by DHS and Air Force Research Laboratory (AFRL) under agreement number FA8750-18-2-0036. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

2. Introduction

During the period of performance and through support from the DHS IMPACT program, the University of Wisconsin has developed and distributed unique data sets to the research community and developed a decision-analytics-as-a-service capability. The focus of our efforts over this period has been on: *(i)* enhancing and extending *Internet Atlas*, which includes a novel repository of maps of *physical Internet infrastructure* (defined as fiber conduits and buildings where conduits terminate) and a web portal for visualization and analysis of the maps and for real-time active measurement, *(ii)* developing a *web user panel* that would provide logs of web browsing activity, *(iii)* developing capabilities to fuse data across different layers of the protocol stack to enable new insights on performance, security and risk, and *(iv)* developing a real time Internet monitoring system for identifying events that include outages, attacks, route changes and congestion.

Prior funding from DHS (the first round of IMPACT and PREDICT) enabled development of Internet Atlas which includes a geographically anchored representation of the physical Internet including *(i)* nodes (*e.g.,* co-location centers, hosting facilities and data centers), *(ii)* fiber conduits/links that connect these nodes, and *(iii)* relevant meta data (*e.g.,* source provenance). The major activities conducted during this period of performance included a full-scale audit of all 1.4K maps in the Atlas repository, the addition of 150 new regional and national network maps and the addition of 275 maps of metro-area fiber infrastructure in the US. The audit was very time consuming (taking over a year to complete) since many of the network service providers had changed names or been acquired since their websites had originally been accessed for data. The metro area maps were an important addition to the repository and became an important focus for research. The openly available web portal (www.internetatlas.org) in which the maps are available to users with account access is based on the widely-used ArcGIS geographic information system (https://www.arcgis.com/index.html), which enables visualization and diverse spatial analyses of the data. Unfortunately, the Atlas web portal was written in Flash, which is no longer supported by Adobe or in browsers. Thus, the Atlas portal went offline on 1/12/2021. Of course, all of the data assets are preserved and continue to be distributed on request. In an effort to lower the security risk in distributing maps with fine-grained location information on network infrastructure, we developed a tool for fuzzing the maps

in a way that preserves certain key characteristics. This tool, which we call Bokeh, was applied to a subset of the metro fiber maps, which were subsequently made available as a separate data set in the IMPACT portal.

Our efforts on developing a web user panel were motivated by the observation that a great deal can be gained by understanding how users interact with websites. This work began by working with the University of Wisconsin Institutional Review Board and legal department to ensure that we had the proper authorization for these activities. Shortly thereafter, we began the effort of developing a system that would enable and facilitate user registration, client software distribution and data management. In parallel, we developed a plugin for Chrome that would capture browsing behavior (pages visited along with time of day, page characteristics and download time) and send it back to the data management system. (We later found out that another IMPACT performer, Prof. Nicolas Christin from Carnegie Mellon University (CMU), had developed a similar system for the purpose of dark web measurement.) Our system was thoroughly tested for functionality and security, and we had set up a mechanism through the UW financial services group to send payments to panelists. Our objective was to recruit users to the panel through advertisements in Reddit, Twitter and other social media forums and then to pay users at approximately the same level as the CMU panel (about $10/mo.). Unfortunately, due to funding delays in the middle of the IMPACT project (our work went without funding for nearly 1 year), we were never able to initiate recruiting of the panel and ultimately, this part of our effort was unsuccessful. The major lessons that we learned were that instead of trying to recruit the panel from scratch, we could have appealed to other crowd-sourcing pools such as Amazon's Mechanical Turk, which would have been a ready source of panelists. Despite these limitations, we were able to acquire a large panel data set from a third party that operates a very large web user panel (Comscore, Inc.). We used this in a research project focused on using machine learning methods to assess and project political candidate preferences. The success of this project highlighted the utility of web user panel data.

Our work on creating capabilities to fuse data across layers of the network protocol stack and thereby derive insights that might not otherwise be possible were focused in two areas. The first is related to our decision-analytics-as-a-service work which was aimed at developing an Internet event monitoring system based on end-to-end latency measurements (*i.e.,* a top-down approach), and the second is focused on annotating physical maps from Atlas with routing information to understand connectivity and where there might be vulnerabilities (*i.e.,* a bottom-up approach). In the top-down approach, we use measurements from NTP, which is an application layer service, to infer when there are changes in the network layer. This work required development of a scalable, high performance system for data gathering and analysis, along with a methodology for identifying meaningful changes in the data. We developed this system – which we named BigBen – and deployed it to collect logs on an hourly basis from 14 participating NTP servers in the US. We used this system to identify a wide range of events from outages to routing changes and made both NTP and event logs available in the IMPACT portal. In the bottom-up approach, we utilize data from Internet Atlas and other sources on Point of Presence (PoP) locations along with BGP peering information to generate an Autonomous System (AS)-level hypergraph representation of connectivity. This novel representation allows, for example, links between PoPs in two different countries (what we call an *Internet frontier*) to be readily identified. This capability required us to create a method for generating what we call a Geographic Footprint

Database (GFDB). In addition to the GFDB, we have written a manuscript on this work, which we intend to submit for publication in fall '21.

As indicated above, our efforts to create a decision-analytics-service for Internet event monitoring system based on NTP latency measurements was successful. Our BigBen system is in the process of transitioning from CloudLab (https://www.cloudlab.us/) to a dedicated system in the Computer Science Lab at UW-Madison. When the transition is complete, we plan to make the system openly available to the community for real time Internet event monitoring. We were successful in publishing several papers about the basic methodologies employed in BigBen. We also have a manuscript that describes the characteristics of events detected in the Internet Protocol version 6 (IPv6) address space that we plan to submit for publication in fall '21.

Distribution of data over the period of performance has focused on Internet Atlas and the other data sets mentioned above. The primary interest in terms of requests has continued be the map and associated data related to the US Long Haul Fiber Infrastructure identified in our Association for Computer Machinery (ACM) Special Interest Group on Data Communications (SIGCOMM) '15 paper (developed with support from our prior IMPACT grant). That map identifies for the first time, the main data carrying component of the Internet in the US. We have also provided accounts on the Atlas portal to 21 requesters over the performance period. The portal received regular, daily use until it was shut down due to Flash being deprecated in January '21. We have also provided the DShield intrusion detection and firewall logs many times over the performance period. The remaining data sets listed by the University of Wisconsin in the IMPACT portal have been requested only a small number of times.

Our research efforts have resulted 15 published papers that have appeared in prestigious venues such as the ACM Internet Measurement Conference (IMC), the ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Conference, International Federation for Information Processing (IFIP) Network Traffic Measurement and Analysis Conference, the ACM / Internet Research Task Force (IRTF) / Internet Society (ISOC) Applied Networking Research Workshop and others. We have also written two papers that have been posted in arXiv and will soon be submitted for publication. Details are listed in Section 8. Several other manuscripts are also underway and will be submitted in fall '21. Among these were two papers that assessed risks to Internet infrastructure due to climate change-related effects including sea level rise and wildfires (see papers #3 and 14 in Section 8). The sea-level rise paper received significant coverage in the popular press and led to many interviews and articles in the popular and technical press.

3. Methods

The novel methods developed during the period of performance fall into the following categories: *(i)* techniques for using Internet infrastructure maps to identify risk or assess performance or robustness, *(ii)* techniques for obfuscating network maps, *(iii)* techniques for utilizing web user panel data to assess political candidate preferences, *(iv)* techniques for generating an AS-hypergraph. Each of these are addressed below.

One of the most important steps that we took during this period of performance was to transition from spending a large amount of time adding new maps to the Atlas repository to using those maps to derive novel insights on Internet performance, robustness and vulnerability. In particular, we focused on developing techniques to merge data from different sources with Internet map data in Atlas to answer questions related to Internet vulnerability to climate change, performance in metro area networks and opportunities to improve performance in cellular networks. In the case of climate change risk, we developed methods to overlay sea level rise projection data from National Oceanic and Atmospheric Administration (NOAA), and Wildfire Hazard Potential projections from the US Forest Service with data in Atlas. To assess performance and robustness in cellular networks in the US and in metro fiber infrastructure in the US, we develop a tessellation-based sampling. This enabled us to summarize infrastructure deployment and project locations for deployment of new infrastructure (edge computing or data centers) that would reduce latency for users.

Our methodology for network map obfuscation is an amalgamation of techniques that are informed by prior work, and offers three tiers of obfuscation capabilities. Tier 1 does simple location fuzzing and produces maps with graph accuracy and minimal location obfuscation. Tier 2 also maintains graph accuracy but randomizes the locations of nodes/links thereby providing a higher level of location obfuscation. Tier 3 obfuscates both graph and location accuracy, and preserves only the distributional properties of the underlying graph. The algorithms at each tier provide configurable obfuscation of the target map. Details of each of the algorithms can be found in citation #11 in Section 8.

Our high-level approach to assess political candidate preferences based on web user panel data is to train a set of specially designed classifiers at the start of the prediction period and then apply the classifiers on a day-by-day and state-by-state basis to assess fine-grained shifts in candidate preference over the remainder of the prediction period. We had to address a number of challenges in developing this method. First, we had to address feature selection and identify which aspects of Web browsing are most informative for prediction of candidate preference. Second, we had to address the challenge of learning a model for individual user behavior given that training labels on a per-user basis are not available. Rather, the best training data available consists of polling done at the per-state or per-county level. Third, we had to address the problem of making predictions on a state level, since training per-state models requires subdividing data, with a large subsequent loss of training accuracy. This problem was compounded by the fact that the amount of data per state varies tremendously due to both population density differences and geographical state sizes. Finally, we had to address the challenge of varying population composition as a function of day of the week. It is well known that user activity on the Web varies qualitatively based on day of the week. In order to make accurate predictions on a day-to-

day basis, this effect must be corrected. For each of the above challenges, we had to clarify the nature of the problem and develop new solutions. Details are provided in citation #5 in Section 8.

Our methodology for generating an AS-hypergraph representation of the Internet is a combination of heuristics that focus on fusing data from a variety of sources. We utilize data from multiple layers of the protocol stack including physical layer data from Internet Atlas, PeeringDB (https://www.peeringdb.com/) and Internet eXchange points (IXPs), BGP data from Routeviews (http://www.routeviews.org/routeviews/) and Réseaux IP Européens (RIPE) Routing Information Service (RIS) (https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris), and active measurements from bdrmapIT (https://catalog.caida.org/details/paper/2018_pushing_boundaries_bdrmapit). We generate the AS-hypergraph representation of Internet connectivity by first inferring the components of the hypergraph that identify the physical presence of ASes at interconnection locations. Logical connectivity between ASes is then established through active measurement. This representation along with targeted traceroute measurements can be used in a variety of ways including identification, with different levels of confidence, of Internet frontiers *i.e.,* country-level Internet border crossings. Details of the methodology are described in a manuscript that is being prepared for submission for publication. See paper #21 in Section 8.

We would be remiss if we did not mention that most of the core methods for identifying Internet events using NTP data were developed in our prior round of funding from IMPACT. All of those methods were operationalized in BigBen and are described in the section on Procedures below.

4. Assumptions

Similar to our prior work, Internet Atlas is predicated on the assumption that detailed information on network infrastructure can be found on publicly available webpages and that this information – to the extent of its detail is accurate at the time of publication. Further, since physical infrastructure such as fiber conduits and Points of Presence do not typically move, we assume that the maps are stable over long periods of time, although they can expand due to new deployments. (We found out during our audit of the maps in Internet Atlas during the period of performance, that meta data associated with the maps such as network name and ownership, can change relatively frequently.)

Our work on projecting candidate preferences based on web panel data assumes that the data provided by the panelists is a complete summary of their browsing activity during the measurement period and that they self-reported their demographic information accurately.

5. Procedures

Data in the Internet Atlas repository is the starting point for much of the research and data distribution by the University of Wisconsin over the period of performance. During that time, we added about 150 new regional and national networks to the Atlas repository (most were outside of the US), and 275 metro fiber maps from the largest Metropolitan Statistical Areas (MSAs) in the US. Beyond Atlas, other data sets that have been assembled and distributed relate to research activities on the *(i)* NTP logs provider by NTP server operators *(ii)* Network event data from the BigBen system, *(iii)* a set of 42 obfuscated network maps based on the Bokeh tool and *(iv)* a set of 50 metro fiber maps.

Data related to Atlas was distributed in two ways. First, password protected access to the portal is granted on request for the purpose of research. Second, the long-haul fiber infrastructure data, obfuscated network maps and metro fiber maps were assembled into separate distributions which were provided to researchers on request. The NTP and network event data was also package into compressed distributions, which have been provided on request via the IMPACT portal.

Research over the period of performance was conducted in the areas associated with the overall objectives of the grant, which were to provide datasets and services that offer unique perspectives across all layers of the protocol stack. The specific research procedures developed during this grant include *(i)* using physical layer maps to gain insight on Internet connectivity, risk and performance, *(ii)* developing the BigBen system for real time Internet event detection, and *(iii)* using web panel data for assessing political candidate preferences. Each of these are addressed below.

Research procedures for using physical layer maps to gain insights on Internet connectivity, risk and performance have focused specifically on *(i)* understanding the details of the characteristics of mobile edge and metro fiber connectivity, *(ii)* understanding risks due to climate change including sea level rise and wildfires. To assess and characterize connectivity, our procedures begin with Internet map data from Atlas – shapefiles with specific locations of cell tower or metro fiber respectively – and focus on developing novel methods for analysis. We assess characteristics that include correspondence with other critical infrastructure (*e.g.,* roads, rails, data centers, etc.), population, auto traffic, etc. (all of these are available in ArcGIS, which is our tool of choice) to understand demand and opportunities for new deployments. To assess and characterize risks due to climate change related phenomena, our procedures also begin with Internet map data from Atlas and include importing risk data *e.g.,* from NOAA or the US Forest Service. We then conduct an overlap analysis, which reveals which infrastructure is at risk, the severity of the risk and suggests ways in which the risk might be mitigated.

Research procedures for using NTP data to identify events in the Internet have focused on *(i)* operationalizing the NTP data analysis process and *(ii)* understanding details of the global NTP infrastructure, assessment of risk and opportunities to improve performance. Our research on utilizing NTP as the basis for real time event detection in the Internet are significant because this service to use passive measurements instead of active probing. The benefits of passive measurement are that it does not add load to the Internet and it can operate much more efficiently in terms of scan rate. Our efforts have focused on the fact that embedded in NTP packets are

details on one-way-delays between hosts that can provide a unique perspective on path properties. Our design for harnessing NTP in BigBen is shown in Figure 1.

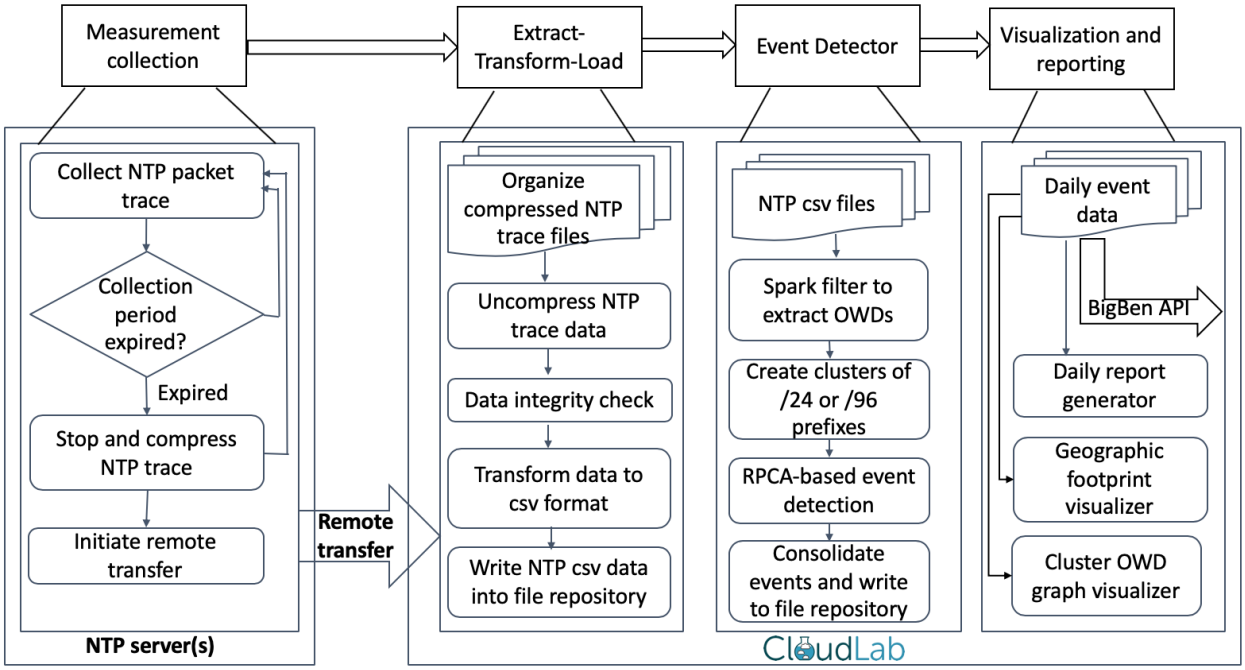

Figure 1: BigBen's system architecture. The measurement component resides on each remote NTP server that contributes data, while the remaining components operate in a cloud infrastructure. (csv – comma-separated values, OWD – one way delay, RPCA - Robust Principal Component Analysis, API – Application Programming Interface)

Details of this procedure can be found in citation #16 in Section 6. Research procedures related to understanding how network time synchronization can be improved were focused on Internet of Things (IoT) devices. In that work, we utilized both Raspberry Pi and Arduino devices to synchronize using standard protocols and with our new protocol called Synchronization Protocol for ioT (SPoT).

Research procedures on using web panel data for assessing political candidate preferences begin with a corpus of web panel data. In the case of our work, we partnered with Comscore, Inc. which operates a web panel with ~2 million (M) users worldwide (roughly 300 thousand (K) of these are in the US). As noted above, we developed a sophisticated set of Machine Language (ML)-based methods to classify the preferences of users. While this could be done in a general fashion, our focus was on presidential candidates in the 2016 election. Since this was a retrospective study, we were able to use the voting details from the 2016 election to train our models (using county-level voting results) and then to apply them in a n-fold holdout fashion to assess their efficacy. A similar procedure could be used to build and maintain models for future election cycles alongside or potentially in place of standard polling, which has been shown to be increasingly unreliable.

6. Results and Discussion

Key results from our studies of Internet risk and performance include the following:

1) In our study of how climate change related sea level rise poses a risk to Internet fiber in the US we find that 4,067 miles of conduit will be under water and 1,101 nodes (*e.g.,* points of presence and colocation centers) will be surrounded by water in the next 15 years. We further quantify the risks of sea level rise by defining a metric that considers the combination of geographic scope and Internet infrastructure density. We use this metric to examine different regions and find that the New York, Miami, and Seattle metropolitan areas are at highest risk. We also quantify the risks to individual service provider infrastructures and find that CenturyLink, Inteliquent, and AT&T are at highest risk. While it is difficult to project the impact of countermeasures such as sea walls, our results suggest the urgency of developing mitigation strategies and alternative infrastructure deployments. See paper #3 in Section 8.

2) In our study of how climate change related wildfires poses a risk to cellular infrastructure in the US we find wide variability in the number of cell transceivers that were within wildfire perimeters over the past 18 years. In a focused analysis of the California wildfires of 2019, we find that the primary risk to cellular communication is power outage rather than cellular equipment damage. Our analysis of future risk based on wildfire hazard potential identifies California, Florida and Texas as the three states with the largest number of cell transceivers at risk. Importantly, we find that many of the areas at high risk are quite close to urban population centers, thus outages could have serious impacts on a large number of cell users. We believe that our study has important implications for governmental communication assurance efforts and for risk planning by cell infrastructure owners and service providers. See paper #14 in Section 8.

3) In our study of cellular network infrastructure in the US, we find that cell tower deployments are relatively consistent in large urban areas and along highways near those areas, but that deployments are more sparse and inconsistent in rural areas. We also find that cell towers are typically within 10 miles of data centers in large urban areas, but distances can be much further in rural areas. We also find that placement of micro data at cell towers in both urban and especially rural settings can significantly improve response times for users. See paper #4 in Section 8.

4) In our study of metro area fiber deployments in the US we find a close correspondence between fiber deployments and roadways in all of the metro areas. We show that a handful of metro areas have a relatively large number of fiber providers (25 out of 204 metro areas we analyzed have 10 or more different fiber providers), while others are much more limited. We also find that some metro areas with relatively low population density have a 3 to 4 times more deployed fiber than metro areas with similar population density. These characteristics and others included in our analysis highlight the diversity of current metro fiber deployments and suggest different strategies for future deployments in each area. We also assess new connections to data centers over a range

of scenarios in the New York, Newark, Jersey City metro area. Our results show that even a small number of new connections to data centers can significantly reduce physical distances to users, which has important implications for service performance and moving traffic closer to the end users. See paper #12 in Section 8.

Key results from our studies of the Network Time Protocol and how it can be used for Internet event detection include the following:

1) In our study that describes the implementation of BigBen, we use the system to collect and process large NTP data sets and provide daily event reporting over a period of 6 months. We find that our implementation is efficient and could support hourly event reporting. We find that BigBen identifies a wide range of Internet events characterized by their location, scope and duration. We compare the events detected by BigBen vs. events detected by a large active probe-based detection system. We find only modest overlap and show how BigBen provides details on events that are not available from active measurements such as University of Southern California's Trinocular (also a performer on this grant). We report on the perspective that BigBen provides on Internet events that were reported by third parties such as the Center for Applied Internet Data Analysis (CAIDA) and Internet Service Providers (ISPs) that experienced large outages. We find that in each case, BigBen confirms the event and provides details that were not available in prior reports, highlighting the utility of the passive, NTP-based approach. See papers #2, 6 and16 in Section 8.

2) We use BigBen to collect a 70 gigabyte (GB) corpus of NTP data over a period of 5 months. We find that the events identified in IPv6 have a wide range of characteristics in terms of their location, scope and duration. We compare events detected in IPv6 space vs. those detected in Internet Protocol version 4 (IPv4) space and find that while there is some overlap, majority of what we find in the IPv6 space is unique. We also examine the IPv6 perspective of Internet events that have been reported by third parties. In each case we find that BigBen confirms the reported IPv4 event and provides details on impacted IPv6 prefixes that were not available in prior reports, highlighting the utility of the passive, NTP-based approach to Internet event monitoring. See paper #16 in Section 8.

3) We investigated how current local time is reported accurately by devices connected to the Internet based on the Time Zone Database (TZDB). Our longitudinal analysis of the TZDB highlights how Internet time has been managed by a loose confederation of contributors over the past 25 years. We drill down on details of the update process, update types and frequency, and anomalies related to TZDB updates. We find that 76% of TZDB updates include changes to the Daylight Saving Time (DST) rules, indicating that DST has a significant influence on internet-based time keeping. We also find that about 20% of updates were published within 15 days or less from the date of effect, indicating the potential for instability in the system. We also consider the security aspects of time management and identify potential vulnerabilities and propose several enhancements for TZDB management that reduce vulnerabilities in the system. See citation #10 in Section 8 for details.

4) We investigated how IoT devices synchronize their clocks. We began by examining clock drift on two standard IoT prototyping platforms. We find clock drift on the order of seconds over relatively short time periods, as well as poor clock rate stability, each of which make standard synchronization protocols ineffective. To address this problem, we developed a synchronization system, which includes a lightweight client, a new packet exchange protocol called SPoT and a scalable reference server. We evaluated the efficacy of our system over a range of configurations, operating conditions and target platforms. We find that SPoT performs synchronization 22x and 17x more accurately than Message Queue Telemetry Transport (MQTT) and Simple Network Time Protocol (SNTP), respectively, at high noise levels, and maintains a clock accuracy of within ~15ms at various noise levels. See paper #7 in Section 8.

Key results from our studies of the web and candidate preferences based on web panels include the following:

1) In our study of assessing political candidate preferences through web browsing behavior, we find that "social referrals", *i.e.,* visits to sites originated from social media, are more important to infer candidate preference than those originated from other sources, such as search engines and Uniform Resource Locators (URLs) directly typed into the browser. We also find that Web browsing behavior for assessing candidate preference, can be particularly useful in terms of day-to-day and state-to-state level predictions that elucidate the impact of exogenous effects such as rallies or announcements such as the "Comey letter." Our results suggest that access to browsing data gives considerable power to assess the preferences of the electorate. With respect to understanding candidate preference, we believe that further use of Web browsing data is likely to uncover additional insights about the impact of campaign strategies, candidate speeches and visits, and political ad campaigns. See paper #5 in Section 8.

2) We investigated Web structure and dynamics by analyzing over 1 trillion Uniform Resource Locators (URLs) requested during Web browsing by the 2 million-person Comscore user panel over a period of 12 months. We examined the lifetime of URLs and find that in contrast to early studies, the set of URLs visited is highly dynamic and well-modeled by a gamma distribution. Next, we analyzed URL-traversal patterns and find that browsing behaviors differ substantially from hyperlink connectivity. One consequence of this is that the structure of the Web that is derived from hyperlink connectivity does not extend directly to actual user behavior. Finally, we considered the commonly used path and query portions of URLs and highlight their characteristics when used by different website genres. We find that these semantic differences suggest that URL structure can broadly classify the kind of resource that a URL references. Our analyses lead to a set of proposed enhancements to the URL standard that would improve Web manageability and transparency and make a step toward the semantic web. Details of our proposals can be found in papers #8 and 9 in Section 8.

7. Conclusions

Over the course of the past four years, our efforts on this grant have focused on data distribution, developing decision-analytics-as-a-service capabilities and research. For the period of performance, the Internet Atlas portal received over 27K page views by over 16K unique visitors, we provided 21 accounts for access to Internet Atlas, and 117 datasets were provided through the IMPACT portal (primarily the Internet Long Haul infrastructure data and DShield logs).

In addition to distributing data sets, we published 15 research papers in high quality venues, posted 2 additional manuscripts on arXiv and are in the process of writing 4 more manuscripts. Our research activities have focused on *(i)* using maps in our Internet Atlas repository to understand details of the Internet's connectivity, risks to Internet infrastructure due to climate change-related phenomena, (ii) opportunities to improve performance understanding of network time synchronization and utilizing NTP measurements as the input for an Internet event detection service, and *(iii)* using web panel data to gain insights on web structure and dynamics and as input for identifying political candidate preferences. Despite significant efforts to build a software infrastructure to support a web user panel of our own, we were ultimately unsuccessful at setting up the panel due to a long delay in funding at a key point in the program. Our team has given 25 talks in various academic and industry forums. Our results on the risk of sea level rise to Internet infrastructure were the subject of numerous articles in both technical and popular press.

8. References (research papers produced during the period of performance)

1) J. Sommers, R. Durairajan and P. Barford. "Automatic Metadata Generation for Active Measurement", In Proceedings of the ACM Internet Measurement Conference, November, 2017.

2) M. Syamkumar, S. Mani, R. Durairajan, P. Barford and J. Sommers. "Wrinkles in Time: Detecting Internet-wide Events via NTP", In Proceedings of IFIP Networking Conference, May, 2018.

3) R. Durairajan, C. Barford and P. Barford. "Lights Out: Climate Change Risk to Internet Infrastructure", In Proceedings of the ACM/IRTF/ISOC Applied Networking Research Workshop, July, 2018.

4) M. Syamkumar, P. Barford and R. Durairajan. "Deployment Characteristics of The Edge in Mobile Edge Computing", In Proceedings of the ACM SIGCOMM Workshop on Mobile Edge Communications, August, 2018.

5) G. Comarela, R. Durairajan, P. Barford, D. Christenson and M. Crovella. "Assessing Candidate Preference through Web Browsing History", In Proceedings of ACM SIGKDD, Conference on Knowledge Discovery and Data Mining (KDD), August, 2018.

6) R. Durairajan, S. Mani, P. Barford, R. Nowak and J. Sommers. "TimeWeaver: Opportunistic One Way Delay Measurement via NTP", In Proceedings of ITC30 - Teletraffic in a Smart World, September, 2018.

7) S. Mani, R. Durairajan, P. Barford and Joel Sommers. "An Architecture for IoT Clock Synchronization", In Proceedings of the 8th International Conference on Internet of Things (IoT 2018), October, 2018.

8) E. Oakes, J. Kline, A. Cahn, K. Funkhouser and P. Barford. "A Residential Client-side Perspective on SSL Certificates", In Proceedings of the IFIP Network Traffic Measurement and Analysis Conference, June, 2019.

9) J. Kline, E. Oakes, and P. Barford. "A URL-based Analysis of WWW Structure and Dynamics", In Proceedings of the IFIP Network Traffic Measurement and Analysis Conference, June, 2019.

10) S. Mani, P. Barford, R. Durairajan and J. Sommers. "What time is it? Managing Time in the Internet", In Proceedings of the ACM/IRTF/ISOC Applied Networking Research Workshop, July, 2019.

11) Y. Gullapalli, J. Koritzinsky, M. Syamkumar, P. Barford, R. Durairajan and J. Sommers. "Bokeh: Obfuscating Physical Infrastructure Maps", In Proceedings of the ACM SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, November, 2019.

12) S. Mani, M. Hall, R. Durairajan and P. Barford. "Characteristics of Metro Fiber Deployments in the US", In Proceedings of the IFIP Network Traffic Measurement and Analysis Conference, Berlin, Germany, June, 2020.

13) J. Kline, A. Aelony, B. Carpenter and P. Barford. "Triangulated Rank-ordering of Web Domains", In Proceedings of the ITC32 Conference, Osaka, Japan, September, 2020.

14) S. Anderson, C. Barford and P. Barford. "Five Alarms: Assessing the Vulnerability of US Cellular Communication Infrastructure to Wildfires", In Proceedings of the ACM Internet Measurement Conference, Pittsburgh, PA, October, 2020.

15) S. Mani, Y. Cao, P. Barford and D. Veitch "iHorology: Lowering the Barrier to Microsecond-level Internet Time", arXiv:2011.06713, November, 2020.

16) M. Syamkumar, Y. Gullapalli, W. Tang, P. Barford and J. Sommers. "BigBen: Telemetry Processing for Internet-wide Event Monitoring", arXiv:2011.10911, November, 2020.

17) S. Patnaik, P. Barford, D. Fratta, B. Jensen, N. Lord, M. Malloy and H. Wang. "Internet Photonic Sensing: Using Internet Fiber Optics for Vibration Measurement and Monitoring", In Proceedings of the ACM SIGCOMM Workshop on Optical Systems (OptSys), August, 2021.

18) A. Anderson, P. Barford, M. Crovella and J. Sommers, "An Elemental Decomposition of the Domain Name System", Paper in preparation, 2021.

19) L. Salamatian, S. Anderson, J. Mathews, P. Barford, M. Crovella and W. Willinger, "On the Geometry of Measured Internet Latencies: A Curvature-based Analysis and its Applications", Paper in preparation, 2021.

20) S. Anderson, T. Bell, P. Eagen, N. Weinshenker and P. Barford, "PowerPing: Measuring the Impact of Power Outages on Internet Hosts in the US", Paper in preparation, 2021.

21) L. Salamatian, S. Anderson, P. Barford and A. Dianotti, "Geolocating the Internet Routing Hypergraph", Paper in preparation, 2021.

22) P. Barford, "DIVERSE DATA SETS FOR IMPACT, Report # AFRL-RI-RS-TR-2018-030, February, 2018.

9. Acronym List

ACM	Association for Computer Machinery
API	Application Programming Interface
arXiv	an open-access electronic publication archive
AS	Autonomous System
BGP	Border Gateway Protocol
CMU	Carnegie Mellon University
csv	comma-separated values
DHS	Department of Homeland Security
DNS	Domain Name System
DShield	a community-based collaborative firewall log correlation system
DST	Daylight Savings Time
GB	gigabyte
GFDB	Geographic Footprint Database
IDS	Intrusion Detection System
IMPACT	Information Marketplace for Policy and Analysis of Cyber-risk & Trust
IoT, ioT	Internet of Things
IPv4, IPv6	Internet Protocol version 4, version 6
IRB	Institutional Review Board
IXP	Internet eXchange Point
K	thousand
M	million
MQTT	Message Queue Telemetry Transport
MSA	Metropolitan Statistical Area
NOAA	National Oceanic and Atmospheric Administration
NTP	Network Time Protocol
OWD	one way delay
PoP	Point of Presence
PREDICT	Protected Repository for the Defense of Infrastructure Against Cyber Threats
RPCA	Robust Principal Component Analysis
SNTP	Simple Network Time Protocol
TTA	Technical Topic Area
TZDB	Time Zone Database
URL	Uniform Resource Locator
US, U.S.	United States
UW	University of Wisconsin