

## **NSF Workshop on Emerging Research Opportunities at the Intersection of Statistics and Internet Measurement Final Report**

Paul Barford, University of Wisconsin – Madison  
Tony Ng, Bentley University

November 27, 2023

### **Abstract**

The Workshop on Emerging Research Opportunities at the Intersection of Statistics and Internet Measurement was held in January 2023 at Boston University. The goal of the workshop was to bring together Internet measurement researchers and statisticians/mathematicians to identify important and promising future research directions and exchanging research ideas that could lead to future collaborations. Forty-two scholars from academia, industry, and the NSF participated. The workshop agenda consisted of a series of 17 keynote talks interspersed by small group discussions. Talks and discussions addressed a wide range of topics including Internet traffic analysis, connectivity and topology analysis, application and user behavior, challenges and opportunities in Internet data collection and data processing, cutting-edge models, and methodologies in different areas of statistics, applied mathematics and data science. More specifically, talks discussed selection biases and qualifying biases in estimates from crowdsourced and other non-random samples, process monitoring, anomaly detection, adversarial risk analysis, data protection, models and methods for network and Internet traffic data collection and processing, and advanced and innovative statistical models, methods, algorithms, and tools that can be applied directly in Internet measurement research. The major outcomes from the workshop are twofold. First, the workshop was successful in facilitating introductions between participants from the Internet measurement and statistics/mathematics communities. Conversations in small group sessions were cordial and carried on beyond allotted times. Second, opportunities for future collaborations were clearly identified over the range of topics that were discussed from experimental design and data gathering to data analysis and modeling to data privacy and visualization. The possibility of holding this workshop again in the future to toward the goal of fostering new research and new collaborations was also discussed.

### **Motivation and Objectives**

The enormous size, complexity, and dynamics of the Internet make it impossible to understand or reason about simply by reading specifications or examining it from the perspective of any single network or set of network protocols. Hence, it is crucial to develop and utilize empirical techniques for Internet research that can systematically assess its organization, configuration, operation, performance, and security. Such techniques can lead to insights that form the foundation for improving and expanding Internet technologies and systems and inventing new methods of networking and communication. The ability to effectively collect, process, analyze, and visualize Internet measurement data is essential in Internet measurement research. The huge volume and diversity of Internet data speaks directly to the need for solutions based on modern statistical techniques and tools.

This workshop was motivated in part by the new NSF Internet Measurement Research: Methodologies, Tools, and Infrastructure (IMR) program, which is a partnership between the Directorate for Computer and Information Science and Engineering (CISE) and the Directorate of Mathematical and Physical Sciences (MPS). The IMR program, launched in 2022, supports innovative research focused on methodologies, tools, and research infrastructure for Internet measurement for access networks (both wireless and fixed broadband) and core Internet. The stated goal of IMR is to “encourage, coordinate, and connect research in Internet measurement in a comprehensive manner” and to accomplish this through collaborations between members of the CISE and MPS communities.

The goal of this workshop was to bring together Internet measurement researchers and statisticians/mathematicians to identify important and promising future research directions, including creating novel methods for collecting, anonymizing, modeling, and analyzing Internet measurement data. The intention was to provide a platform for researchers who work on different aspects of Internet measurement, including investigation of Internet design, structure, and performance, applications and end-user behavior, and methods and tools for network instrumentation and visualization; and statisticians/mathematicians who develop statistical models and methodologists for complex data collection, processing, and analysis to communicate and exchanging research ideas. In addition, the workshop aimed to facilitate introductions and interactions between participants to pave the way for collaborative research projects in the future and ultimately to significant impacts on the theory and practice of Internet measurement research.

## **Overview and Organization**

The workshop on Emerging Research Opportunities at the Intersection of Statistics and Internet Measurement was held at Boston University in Boston, MA, on January 11 and 12, 2023. Forty-two scholars from academia, industry, and the NSF participated (see list of participants below). The workshop agenda consisted of a series of keynote talks followed by small group discussions (see agenda and titles/abstracts of talks below).

There were 8 talks from Internet measurement experts, 11 talks from statistics/mathematics experts, and one talk from program managers from the NSF. Through the keynote talks, statisticians/mathematicians had opportunities to learn about the challenges, approaches, and practical issues related to Internet measurement research from experts in the field. Likewise, Internet measurement researchers had opportunities to learn about the recent developments and advances in statistical models, algorithms, and methodologies for collecting and analyzing large and complex datasets. The talk from NSF program managers highlighted details of the IMR program as well as other programs that align with the interests of workshop participants.

Each of the four small group discussions was organized to enable different groups of workshop participants to meet and interact with each other. Each session featured a set of discussion topics that were offered to facilitate conversations about foundational issues, current challenges, and opportunities for future collaboration.

## **Key Issues from the Technical Talks**

The technical talks from Internet measurement researchers addressed a wide range of topics, including measurement in quantum networks, Internet delay-space characterization, measuring and characterizing online service dependencies, crowdsourced techniques for assessing net neutrality, Internet outage detection and characterization, large-scale infrastructure for Internet measurement, the use of AI/ML in Internet measurement research, and key challenges associated with Internet data analysis. These topics highlight the diversity of issues that are of interest to the Internet measurement community and areas of potential future collaboration.

The technical talks from statisticians and mathematicians also highlighted a variety of topics including cutting-edge models, and methodologies in different areas of statistics, applied mathematics and data science. For instance, these talks discussed selection biases and qualifying biases in estimates from crowdsourced and other non-random samples, process monitoring, anomaly detection, adversarial risk analysis, data protection, models and methods for network and Internet traffic data collection and

processing. The talks included details on some advanced and innovative statistical models, methods, algorithms, and tools that can be applied directly in Internet measurement research.

The common themes and opportunities for collaboration that were conveyed in these talks included:

- Diverse techniques and infrastructure are employed for Internet data collection. These include active probe-based methods that send packets into the Internet (e.g., from applications, standard IP-level tools such as Ping or custom-built tools) and then measure the response by either network elements or, end hosts, or passive methods that utilize a device or specialized software placed somewhere within the network (including on end hosts) to collect data. In many cases, data collection requires large, complex, costly infrastructure to measure phenomena at scale and/or in a comprehensive fashion. Several such infrastructures, including M-Lab and RIPE Atlas, were described during talks. In other cases, collaboration with industry partners is required to get access to otherwise private data to assess certain aspects of Internet phenomena.

Collecting reliable and high-quality data is often a complex and time-consuming task. Statistical techniques in data collection through sampling and well-designed experiments can be applied to simplify and streamline Internet data collection without the loss of the required fidelity in the resulting data corpus. Survey sampling and experimental design are well-developed areas in statistics, however, due to the irregular nature, the scale/dimension, and the complexity of Internet data, modifications of the existing statistical techniques developed for survey sampling and experiment design would require the collaboration of statisticians and Internet measurement researchers.

- Care is required in gathering, organizing, transforming, and utilizing Internet data in research studies. Data collected from the Internet can never be assumed to be sound when it arrives at a collection point. Internet data is often contaminated with records that are malformed, inaccurate, or incomplete in some ways. This has important implications for how Internet data processing pipelines and repositories are designed and implemented. It also has important implications for how data from open repositories must be evaluated before it can be used in research studies. In some cases, data may be private or include personally identifiable information (PII). In such cases, appropriate steps must be taken to secure appropriate permissions and ensure compliance e.g., with IRB requirements. Finally, there are community norms and ethics for gathering data e.g., compliance with Terms of Use statements that must be observed. Unfortunately, these may not always be clear or obvious to new researchers.

Data privacy, data security, and data protection are research areas studied by statisticians that are directly relevant to Internet measurement research. To protect sensitive Internet data and the data transmission process, classification methods, data masking, and deidentification techniques can be developed and employed. In addition, investigation of how the imperfect data collected from the Internet affect the outcome of different analyses and development of adjustment methods are of great interest to statisticians and potential utility in Internet measurement research. For instance, it is common that Internet measurement data are obtained from non-probability sampling methods where the samples are not representative of the whole population of interest. In such a situation, qualifying the biases of the estimates of a parameter of interest and performing suitable adjustments for Internet measurement research would require statistical techniques from statisticians and domain knowledge from Internet measurement researchers. This is clearly an area for potential future collaborations between the two communities.

- Specialized techniques for analysis and visualization are often required to assess and derive insights from Internet measurement datasets. Internet measurement infrastructures can potentially collect

very large quantities of data from diverse locations and at different layers of the protocol stack. Such datasets present significant challenges for exploratory, qualitative, and quantitative assessment. Traditional methods for analyzing these data sets, including statistics, time series, signal processing, and learning-based, are often unable to identify or fully describe phenomena of interest.

Modern statistical process monitoring methods, anomaly detection techniques, and high-dimensional signal identification methods provide feasible approaches to detect Internet network events. Scalable statistical models and methods that account for the dynamic nature and spatial and temporal structure of Internet data are desired. Statistical methodologies developed for detecting structural breaks and anomalies, and analyzing different kinds of Internet data (*e.g.*, Internet traffic data, Internet connection data, and network transmission data) have been discussed in the talks by statisticians/mathematicians.

Similarly, visualizing very large, multi-dimensional datasets often goes beyond the capabilities of traditional techniques and tools. These issues become even more difficult when one considers potential applications in network operations that may require real-time analyses, concise representations of data and visualizations. Finally, the recent advent of large language models and other AI/ML-based techniques will certainly have an impact on Internet measurement research. While these are powerful tools that hold great promise, they must be carefully considered within the context of the problem space and datasets to which they will be applied – the naive application will almost certainly lead to erroneous results. The incorporation of statistical learning and deep learning methods, as well as high-dimensional data visualization and analytical techniques to analyze Internet measurement data are important research topics. These issues speak directly to potential opportunities for collaboration between Internet measurement researchers, and analysis and visualization domain experts.

- Care must be taken when analyzing and drawing conclusions from Internet data analysis. Data that has been carefully processed to remove erroneous records can still be biased or skewed (*e.g.*, due to measurement vantage point selection or Internet infrastructure changes over time) in ways that can lead to incorrect conclusions about phenomena of interest. Best practices must be applied to identify such issues. A related issue is how the identification and characterizations of specific phenomena such as attacks, outages, anomalies, and other network events that are important in network operations are validated. A natural approach is to appeal to sources of ground truth for such events, which can sometimes be provided by network operators. Unfortunately, sources of accurate ground truth are typically few and far between, leading to uncertainty in the correctness of results.

## Conclusion

The workshop was a success in terms of meeting the goal of bringing together Internet measurement researchers and statisticians/mathematicians to discuss pressing challenges and identify future research directions and opportunities for collaboration. Conversation during small group sessions was lively and often drilled down into detail on specific issues that could bear fruit in future research.

While the talk topics and messages conveyed therein speak to a broad research agenda and opportunities for collaboration, they are by no means an exhaustive list. Further, while the immediate goal of bringing together members of the two communities was achieved, the longer-term impact of the workshop in terms of successful collaborations on grant proposals and research projects remains to be seen. A natural means of increasing the probability of such outcomes is to hold similar workshops in the future. One potential option would be to organize either before or after the ACM Internet Measurement Conference (IMC). This

is a venue that is well-attended by Internet measurement researchers and may be of interest to researchers in stats/math. Given the number of participants and level of enthusiasm at the First Workshop on Emerging Research Opportunities at the Intersection of Statistics and Internet Measurement, one would expect that future workshops – especially co-located with IMC – would be similarly successful.

### **Recommendations for Next Steps**

Based on discussions during the workshop and follow-up conversations with participants, we make the following recommendations to the community and the NSF for next steps:

- 1) While the workshop was a success in terms of meeting the goal of bringing together Internet measurement researchers and statisticians/mathematicians, fostering collaborations between these communities that result in new research contributions will take time. Toward the goal of building meaningful and impactful research collaborations, we recommend future – perhaps annual - Workshops on Emerging Research Opportunities at the Intersection of Statistics and Internet Measurement.
- 2) The NSF Internet Measurement Research program (the most recent call can be found at <https://www.nsf.gov/pubs/2022/nsf22519/nsf22519.htm>) is designed to support “methodologies, tools, and research infrastructure for Internet measurement spanning access (both wireless and fixed broadband) and core Internet.” It is often difficult to acquire funding for collaborative research efforts that span communities. Yet, such support is vital in Internet measurement since addressing many of the research challenges requires diverse skills and experience. We recommend that the NSF continue the IMR program to enable the nascent collaborations between Internet measurement researchers and statisticians/mathematicians to mature and expand.
- 3) Visualizations play a key role in conveying what is interesting and important in Internet measurement research. Images of time series, scatter plots, bar charts and graph connectivity are standard fare in Internet measurement research papers. However, in some cases, a visualization can be the highlight in a paper that captures and summarizes the major point of the work. In future workshops and/or funding programs, we recommend a visualization track that concentrates on new and innovative methods for visualizing network data and network-related information (*e.g.*, for contextualizing), including novel data representation methods, animation tools, AR-based use cases, etc. should be considered.
- 4) A critical touchpoint between stats/math researchers and Internet measurement researchers is data. Unfortunately, in most cases, data is collected for a project and not preserved in a fashion that supports reproduction of results or future research. While there are groups that collect, archive and make data available to the community (*e.g.*, CAIDA), there is a broader need to identify and support longitudinal data collection, archival and distribution. We recommend working toward community consensus around what longitudinal datasets should be valuable to the community and funding/institutional methods to sustain support for collection, archival and distribution of those datasets. One potential mechanism to encourage the follow-up that is required to curate and document datasets collected over the course of a grant would be a “dataset packaging supplement” that could be requested for any CISE grant (analogous to an REU supplement).
- 5) The Workshop on Emerging Research Opportunities at the Intersection of Statistics and Internet Measurement as well as the NSF IMR program specifically encourages collaborations between stats/math and Internet measurement researchers. We recommend encouraging broader cross-disciplinary research that could include social scientists, economists, etc. While such research collaborations take time and effort to build, they offer an opportunity to bring principled data gathering and analysis methodologies to bear on diverse problems, and to develop solutions that have important broader impacts. Fostering collaborations between Internet measurement, stats/math and researchers from other communities could be part of a future, expanded IMR call.

## **Acknowledgement**

This workshop was supported by National Science Foundation (NSF) grants CNS-2234288. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

## Additional Materials

### Participants:

#### Organizers

Paul Barford	University of Wisconsin	<a href="mailto:pb@cs.wisc.edu">pb@cs.wisc.edu</a>
Tony Ng	Bentley University	<a href="mailto:tng@bentley.edu">tng@bentley.edu</a>

#### Measurement

Mark Allman	ICSI	<a href="mailto:mallman@icir.org">mallman@icir.org</a>
Fabian Bustamante	Northwestern University	<a href="mailto:fabianb@cs.northwestern.edu">fabianb@cs.northwestern.edu</a>
John Byers	Boston University	<a href="mailto:byers@bu.edu">byers@bu.edu</a>
David Choffnes	Northeastern University	<a href="mailto:choffnes@ccs.neu.edu">choffnes@ccs.neu.edu</a>
Mark Crovella	Boston University	<a href="mailto:crovella@bu.edu">crovella@bu.edu</a>
Alberto Dainotti	Georgia Tech	<a href="mailto:dainotti@gatech.edu">dainotti@gatech.edu</a>
Ram Durairajan	University of Oregon	<a href="mailto:ram@cs.uoregon.edu">ram@cs.uoregon.edu</a>
Zakir Durumeric	Stanford University	<a href="mailto:zakir@cs.stanford.edu">zakir@cs.stanford.edu</a>
Phillipa Gill	Google	<a href="mailto:phillipagill@google.com">phillipagill@google.com</a>
John Heidemann	USC/ISI	<a href="mailto:johnh@isi.edu">johnh@isi.edu</a>
Anita Nikolich	University of Illinois University of California San Diego	<a href="mailto:anitan@illinois.edu">anitan@illinois.edu</a>
Alex Snoeren		<a href="mailto:snoeren@cs.ucsd.edu">snoeren@cs.ucsd.edu</a>
Joel Sommers	Colgate University	<a href="mailto:jsommers@colgate.edu">jsommers@colgate.edu</a>
Neal Spring	Meta	<a href="mailto:nspring@gmail.com">nspring@gmail.com</a>
Don Towsley	University of Massachusetts Amherst	<a href="mailto:towsley@cs.umass.edu">towsley@cs.umass.edu</a>
Walter Willinger	Niksun	<a href="mailto:wwillinger@niksun.com">wwillinger@niksun.com</a>
Abraham Matta	Boston University	<a href="mailto:matta@bu.edu">matta@bu.edu</a>
Azer Bestavros	Boston University	<a href="mailto:best@bu.edu">best@bu.edu</a>
Ricky Mok	CAIDA	<a href="mailto:cskpmok@caida.org">cskpmok@caida.org</a>

#### Stats/Math

Yves Atchadé	Boston University	<a href="mailto:atchade@bu.edu">atchade@bu.edu</a>
Trent D. Buskirk	Bowling Green State University	<a href="mailto:buskirk@bgsu.edu">buskirk@bgsu.edu</a>
Jie Chen	Augusta University	<a href="mailto:jiechen@augusta.edu">jiechen@augusta.edu</a>
Yuzhou Chen	Temple University	<a href="mailto:yuzhou.chen@temple.edu">yuzhou.chen@temple.edu</a>
Yili Hong	Virginia Tech	<a href="mailto:yilihong@vt.edu">yilihong@vt.edu</a>
Eric Kolaczyk	McGill University	<a href="mailto:eric.kolaczyk@mcgill.ca">eric.kolaczyk@mcgill.ca</a>
Ming Li	Amazon	<a href="mailto:mli@alumni.iastate.edu">mli@alumni.iastate.edu</a>
Dennis K. J. Lin	Purdue University	<a href="mailto:dkjlin@purdue.edu">dkjlin@purdue.edu</a>
Regina Liu	Rutgers University	<a href="mailto:rliu@stat.rutgers.edu">rliu@stat.rutgers.edu</a>

George Michailidis	University of Florida	gmichail@ufl.edu
Rong Pan	Arizona State University	rong.pan@asu.edu
Fabrizio Ruggeri	CNR-IMATI, Milan	<a href="mailto:fabrizio@mi.imati.cnr.it">fabrizio@mi.imati.cnr.it</a>
Jonathan Stewart	Florida State University	jrstewart@fsu.edu
Stilian Stoev	University of Michigan	sstoev@umich.edu
S. Lynne Stokes	Southern Methodist University	slstokes@mail.smu.edu
Kwok-Leung Tsui	Virginia Tech	klttsui@vt.edu
Bowei Xi	Purdue University	<a href="mailto:xbw@purdue.edu">xbw@purdue.edu</a>
Emmanuel Yashchin	IBM	yashchi@us.ibm.com
Kevin Rath	Tempo, Inc.	kevin@manoshomecare.com

### NSF

Deep Medhi	NSF	dmedhi@nsf.gov
Rob Beverly	NSF	<a href="mailto:rbeverly@nsf.gov">rbeverly@nsf.gov</a>

### Agenda:

#### Day 1 (January 11, 2023):

8:00am - 8:45am - light breakfast. (Building will be open at 7:30am)

8:45am - 9:00am - Welcome and remarks from organizers

9:00am - 11:00am – talks

#### **Don Towsley** (University of Massachusetts - Amherst): "Quantum Network Tomography"

Abstract: The vision of a quantum Internet is to enable quantum communication between any two nodes on earth. This will allow for distributed quantum computing, quantum sensing, and quantum key distribution achieving capabilities not possible through classical means. One of the big challenges facing the development and operation of such a network is noise. This will require the development of new technologies and statistical inference techniques to enable the measurement and characterization of the performance of different network components such as links and nodes. In this talk I will describe one possible approach based on quantum network tomography. I will describe parallels to approaches taken for classical network tomography along with the new challenges posed by the quantum nature of a quantum network.

#### **Mark Crovella** (Boston University): "Manifold Visualizations of the Internet, or, the Power of Ping"

Abstract: In this talk I'll describe why traditional methods for discovering Internet topology (ie, traceroute) need updating. I'll describe our team's work on developing new visualization methods that avoid the limitations of traceroute and aspire to give a better overall understanding of network structure. The methods combine graph (Ricci) curvature, continuous manifolds, and optimization and rely only on measured latency (as provided by ping). Our methods are well suited for understanding network topologies in general, but work well even in networks (eg cloud providers) where traditional topology measurement is impossible.

**Fabian Bustamante** (Northwestern University): “Dependency and Centralization Around the World”

Abstract: We present results from a large-scale study of third-party dependencies around the world based on regional top-500 popular websites accessed from vantage points in 50 countries, together covering all inhabited continents. This broad perspective shows that dependencies on a third-party DNS, CDN or CA provider vary widely around the world, ranging from 19% to as much as 76% of websites, across all countries. The critical dependencies of websites -- where the site depends on a single third-party provider -- are equally spread ranging from 5% to 60% (CDN in Costa Rica and DNS in China, respectively). Interestingly, despite this high variability, our analysis suggests a highly concentrated market of third-party providers: three providers across all countries serve an average of 92% and Google, by itself, serves an average of 70% of the surveyed websites. Even more concerning, these differences persist a year later with increasing dependencies, particularly for DNS and CDNs. We briefly explore various factors that may help explain the differences and similarities in degrees of third-party dependency across countries, including economic conditions and Internet development, and discuss future directions for this work.

**Fabrizio Ruggeri** (Italian National Research Council in Milano): “Adversarial Risk Analysis and Data Protection”

Abstract: Adversarial Risk Analysis (ARA) is an emergent paradigm supporting decision makers who confront adversaries in problems with random consequences that depend on the actions of all participants. ARA provides one-sided prescriptive support to decision makers maximizing their subjective expected utility by treating the adversaries’ decisions as random variables. We will illustrate the approach presenting results about adversarial classification and detection of attacks aimed at inducing acceptance of spam messages as legitimate, with various negative consequences like access to private data. The approach can also be extended to the transmission of other types of data subject to possible attacks.

**Kwok Leung Tsui** (Virginia Polytechnic Institute and State University): “Healthcare and Public Health Monitoring and Management”

Abstract: Due to the advancement of computation power, sensor technologies, and data collection tools, the research on healthcare and public health monitoring and management has been evolved over the past several decades under different names among various application domains, such as statistical process control (SPC), process monitoring, health surveillance, prognostics and health management (PHM), personalized medicine, etc. There are tremendous opportunities in interdisciplinary research on health monitoring and management through integration of SPC, system informatics, data analytics, PHM, and personalized health management. In this talk we will present our views and experiences in the related research. In particular, we will focus on research on healthcare and public health surveillance and forecasting.

**Dennis KJ Lin** (Purdue University): “Data is Power--Some personal views on Statistics for Modern Data Science”

Abstract: If Francis Bacon were born today, I could well imagine that he might have said “Data is Power” (instead of “Knowledge is power” in his original saying). Data is everywhere. When data speak, do you understand what data is trying to tell you? This talk attempts to explain (1) What exactly does Data Science really mean, and (2) What is the impact of Data Science to the real world. Statisticians have much more to contribute in both the intellectual vitality and the practical utility of Data Science. At the same time, Data Science challenges statisticians to move out of some familiar habits to engage less structured problems, to become more comfortable with ambiguity,

and to engage less scientists in a more fruitful discussion on what the various parties can bring to this new mode of investigation. Some potential directions for future research along this direction will be proposed.

11:00am - 11:30am – coffee break

11:30am - 12:30pm - small groups

12:30pm - 1:30pm - lunch

1:30pm - 3:30pm – talks

**Lynne Stokes** (Southern Methodist University): “Overcoming Selection Bias in Crowd-sourced and other Non-Random Samples”

Abstract: Large datasets containing information about a variety of scientific, business, and social questions are increasingly available. These types of samples include crowdsourced or other voluntarily offered information, or administrative records that only partially cover a population of interest. Such data are known as nonprobability samples, to distinguish them from samples collected using random selection, called probability samples. Because these datasets are often so large and inexpensive, it seems reasonable to assume they must be useful. However, statistics calculated from these samples can result in selection bias because they may not be representative of the population of interest. The promise of such data has inspired sampling researchers to develop new techniques to reduce this bias. In this presentation, I will provide an overview of several of these techniques. We will see that all methods require access to auxiliary data sources. These sources can be either a complete census or a probability sample from the entire population, either of which must contain data for some attributes common to the nonprobability sample. When the common attributes are predictive of the probability of inclusion in the nonprobability sample, then the selection bias for may be reduced by weighting adjustments based on the modeled inclusion probability. These methods are cost effective in cases where large and cheap (volunteer or administrative) datasets can be paired with existing census data or small probability samples.

**Trent D. Buskirk** (Bowling Green State University): “Connecting the Bits: Defining the Internet Population in the U.S. and Discovering Potential Models for Mitigating Biases in Estimates Derived from ‘Connected Population’ Samples”

Abstract: With rises in data collection costs using traditional methods like sample surveys, researchers have been turning more of their focus onto using the internet as a primary mode of data collection. Consequently, measured audiences share the common trait that they are all “connected” to the internet. While the percentage of non-internet households has declined, it persists at about one in ten households. This raises the question of coverage error and bias, and whether there is an approach to reduce possible biases in internet-only samples. In this paper we sought out to determine factors that were associated with internet connectivity status and leveraged machine learning methods and survey methods to identify a small handful of variables that could be used to adjust for internet only household status in resulting estimates derived from samples taken exclusively from the internet population. Specifically, we processed nearly 5,000 variables from over a dozen major public and private datasets to assess properties that could define those who were connected versus those who were not. Finding substantive differences, we then applied a series of data-reducing techniques to arrive at 38 variables that independently skew across internet and non-internet populations. We then developed and fielded a survey of these metrics to assess which dozen or less could be used to construct an efficient propensity model to reduce bias in

internet-only samples. Our analyses revealed that many variables noted in prior research are important predictors of non-internet use, but also identified others. Our final propensity model of 10 variables was highly effective, reducing bias significantly. Many variables tested had biases reduced fourfold. In this talk we will describe an overview of the approach to measure and potentially mitigate biases that can result in estimates derived from internet populations in the U.S. and we will also provide a detailed description of a taxonomy of variables that are related to internet access from a sociological perspective. This work can inform the future of internet measurement research by describing succinctly the so called “connected” population of those who have access to the internet.

**Yuzhou Chen** (Temple University): “Towards Understanding the Internet Measurement through Higher-order Structures and Statistical Learning”

Abstract: Modern computer infrastructure for commerce and communications such as the Internet, electronic payment systems, and file-sharing systems can be represented as complex networks. Cybersecurity analysis is one of the possible ways to manage risk exposure for these complex networks. To understand the robustness of complex networks, equipped with higher-order (sub)structures, we propose a modified Wiener process model for the degeneration of the network functionality upon the removal of nodes due to attacks or malfunctions. We further propose three statistical testing procedures based on the Wiener process model to compare the risk and resilience of two different networks. The proposed methodologies can be applied to any topological measures of network robustness or risk. Practical data analysis for the peer-to-peer networks are presented to illustrate the proposed model and methods.

**David Choffnes** (Northeastern University): "Crowdsourced Net Neutrality Measurements"

Abstract: Through the Wehe mobile app for iOS and Android devices, we have collected millions of measurements that can indicate whether a network provider is giving different performance to different applications. However, the crowdsourced measurements from mobile devices are subject to numerous confounding factors that can make sound inference difficult, including issues such as variability in signal strength, competing flows, and congestion. In this talk, I discuss the data we gather (and practical trade-offs compared to optimal conditions), the methods we use to analyze the data, and how we draw (hopefully) sound conclusions about non neutrality and implementations of bandwidth limits. We are interested in an open discussion about how to better collect and analyze such noisy---but incredibly useful---data.

**Alberto Dainotti** (Georgia Tech): “Understanding the Spatio-Temporal Dynamics of Internet Outages”

Abstract: As our reliance on the Internet increases, so does the need for a reliable Internet. With services continuously migrating to the Internet and as increasingly many people work from home, the Internet is a critical infrastructure that we use and depend upon. Given the potential impact of residential Internet outages, there is a pressing need for techniques that can accurately detect and characterize them and disseminate findings publicly so that various stakeholders can benefit. Despite the significant research advances in Internet outage detection in the last decade, science still lacks an understanding of when, for how long, and how frequently residential Internet outages happen, as well as of their scope in terms of geography, address space, operators, and access technologies. This knowledge gap represents a major obstacle to the progress of science on and operations for Internet reliability. In the proposed work, by harnessing the full potential of active probing for accurate outage detection, we seek to measure and understand residential Internet reliability. Specifically, we develop methods to detect and characterize connectivity outages along various dimensions, including their frequency, duration, and scope.

**Walter Willinger** (Niksun): "NetAI: The dumbing down of networking research!?"

Abstract: Several recent research efforts have proposed Machine Learning (ML)-based solutions that can detect complex patterns in network traffic for a wide range of network security problems. However, without understanding how these black-box models are making their decisions, network operators are reluctant to trust and deploy them in their production settings. One key reason for this reluctance is that these models are prone to the problem of underspecification, defined here as the failure to specify a model in adequate detail. Not unique to the network security domain, this problem manifests itself in ML models that exhibit unexpectedly poor behavior when deployed in real-world settings and has prompted growing interest in developing interpretable ML solutions (e.g., decision trees) for explaining to humans how a given black-box model makes its decisions. However, synthesizing such explainable models that capture a given black-box model's decisions with high fidelity while also being practical (i.e., small enough in size for humans to comprehend) is challenging. In this talk, we describe our recent study that focus on synthesizing high-fidelity and low complexity decision trees to help network operators determine if their ML models suffer from the problem of underspecification. To this end, we present Trustee, a framework that takes an existing ML model and training dataset as input and generates a high-fidelity, easy-to-interpret decision tree and associated trust report as output. Using published ML models that are fully reproducible, we show how practitioners can use Trustee to identify three common instances of model underspecification; i.e., evidence of shortcut learning, presence of spurious correlations, and vulnerability to out-of-distribution samples.

3:30pm - 4:00pm - coffee

4:00pm - 5:00pm - small groups

6:00pm - 9:00pm – dinner at Yardhouse, 126 Brookline Ave, Boston, MA (<https://www.yardhouse.com/>)

**Day 2 (January 12, 2023)::**

8:00am - 9:00am - light breakfast (Building will be open at 7:30am)

9:00am - 11:00am – talks

**Phillipa Gill** (Google) "M-Lab: "User initiated Internet data for the research community"

Abstract: Measurement Lab (M-Lab) is an open, distributed server platform on which researchers have deployed measurement tools. Its mission is to measure the Internet, save the data and make it universally accessible and useful. This talk details the current state of the M-Lab distributed platform, highlights existing measurements/data available on the platform, and describes opportunities for collaboration between M-Lab and network measurement researchers.

**John Heidemann** (University of Southern California/ISI) "Accuracy in Internet Measurement: Correctness, Completeness, and Ground Truth"

Abstract: Measurement of network characteristics such as size (number of end systems, users, routers, and links) and reliability (network outages, uptime) are important, but pose serious technical challenges. How can we judge accuracy of new methods when even operators do not know current ground truth? This talk will discuss the role of correctness and completeness as measures of accuracy when ground truth is uncertain.

**George Michailidis** (University of Florida Informatics Institute): “Fast methods for detection of structural breaks in data with network structure”

Abstract: Detecting structural breaks and anomalies in the data generation mechanism of complex data with network structure represents a critical task, due to numerous applications in Internet monitoring, as well as other engineering, public health and social science applications. We present a simple to implement, yet powerful algorithmic framework, for this detection task. We also briefly discuss theoretical guarantees for key quantities of interest, in the form of consistency, finite sample bounds, and asymptotic distributions for the points where structural breaks occur and other model parameters. We illustrate the framework with examples from different application domains.

**Bowei Xi** (Purdue University): “Multifractal and Gaussian Fractional-Sum-Difference Models for Internet Traffic”

Abstract: A multifractal fractional sum-difference model (MFSD) is a monotone transformation of a Gaussian fractional sum-difference model (GFSD), the Gaussian image of the MFSD. The GFSD is a mixture of two components: a moving two-sum of discrete fractional Brownian motion (fBm), and white noise. Internet packet traffic inter-arrival times are very well modeled by an MFSD; this is validated by extensive model checking for 715,665,213 measured arrival times on 3 Internet links. Mathematical investigations of many traffic statistics, enabled by the mathematical tractability of the model, provide new insight for traffic phenomena; this includes fundamental explanations of a number of phenomena based on how the relative weights of the fBm and white noise components change with changing factors such as the traffic rate and time aggregation. The MFSD can be used to generate synthetic traffic for network simulation.

**Eric Kolaczyk** (McGill University): “Accounting for Noise in Network Analysis”

Abstract: While the use of network analysis has now permeated most domains, an overwhelming proportion of network analysis methods still operate as if the networks we observe are noise free. In many settings, such an assumption could not be further from the truth. Examples include most biological networks, contact networks in epidemiology, and various IoT networks. In this talk, I will present a simple methods-of-moments approach to network analysis that allow users to obtain unbiased inferences of network-related parameters under 'noisy networks'. These estimators are accompanied by confidence intervals deriving from a novel bootstrap algorithm. The primary application will be to counting subgraphs, but I'll comment briefly on how we have extended the same ideas to problems of controlling epidemic spread, and quantifying treatment effects in network experiments.

11:00am - 11:30am - coffee

11:30am - 12:30pm - small groups

12:30pm - 1:30pm - lunch

1:30pm - 2:30pm – talks

**Jonathan Stewart** (Florida State University): “Statistical network analysis methods for internet measurement research”

Abstract: Internet data can often be relational in nature, meaning data correspond to pairwise observations which describe interactions or relationships between entities in populations of interest.

Represented as a network, such data can be analyzed through statistical network analysis methodology. Key statistical challenges present in internet data include significant data heterogeneity and scale in quantity of data. Moreover, a complete observation of the network or relevant data may be infeasible for a variety of reasons, which include privacy, availability, or feasibility constraints due to the scale of data collected. Modern research in statistical network analysis has focused on providing novel solutions to these challenges and more in network data settings. In this talk, I will review some existing approaches to these challenges in the literature, and will highlight important directions for future research, focusing on pertinence to modern problems in internet data and measurement.

**Stilian Stoev** (University of Michigan): “Concentration of Maxima and Sparse Anomaly Identification in High Dimensions: A phase transition for the exact support recovery problem”

Abstract: Anomaly detection and identification in high-dimensional signals is an ubiquitous statistical problem. We study the theoretical limits for the exact recovery of all non-zero entries in a sparse signal observed with additive light-tailed noise. In the regime when the dimension of the signal tends to infinity, we identify the fundamental statistical limits for exact support recovery in the large family of thresholding estimators. This boundary determines when exact support recovery is possible and when it is not as a function of the signal magnitude and sparsity. This result emerges from a concentration of maxima phenomenon for light-tailed errors, which holds under remarkably strong error-dependence. The exact support recovery boundary is universal and applies to all estimators in the class of log-concave error models. Many more results on finite sample Bayes optimality, asymptotic minimax optimality, as well as applications to statistical genetics, can be found in the recent monograph:

Gao, Zheng and Stoev, Stilian, 2021. Concentration of Maxima and Fundamental Limits in High-Dimensional Testing and Inference, SpringerBriefs in Probability and Mathematical Statistics. <https://link.springer.com/book/10.1007/978-3-030-80964-5>

**Robert Beverly and Deep Medhi** (NSF): “NSF CISE Updates”

2:30pm - 3:00pm - coffee

3:00pm - 4:00pm - small groups

4:00pm - final remarks

### **Group Session #1: discussion topics**

- Taxonomies for empirical study of the Internet
  - Historical?
  - Current?
- What are the major challenges in Internet measurement research?
  - Data
  - Analytic techniques
  - Privacy
- What are the major successes in Internet measurement?
- Where are future opportunities?

### **Group Session #2: discussion topics**

- Internet measurement data sources and tools
  - Where and how is internet data collected?
- Active vs. passive measurement - challenges and opportunities for each
- Data collection processes: sampling design, optimal experimental schemes
- Commercial sources of data - how can they be analyzed without revealing private information?
- What data sources are openly available?

### **Group Session #3: discussion topics**

- Data processing and data quality control.
- How can data pipelines be generated to ensure that results are accurate and not contaminated by bad data?
- Managing and analyzing very large data sets - are databases worth the effort?
- Organizing data for specific types of analyses - how can cloud infrastructures be used?
- What tools are available/useful for visualizing large, diverse data?

### **Small Group Session #4: discussion topics**

- Data analysis and modeling: What are the best ways to address research questions with empirical datasets?
  - Basic statistics, time series analysis, signal processing, etc.
- How can AI/ML be useful to improve understanding of Internet data?
  - Feature engineering: the process of selecting, manipulating, and transforming data into features
  - Feature selection and dimension reduction
- Techniques for anomaly detection
- Predictive analysis
- Validation of results