Measuring the Internet

Michael Rabinovich
Case Western Reserve University

Mark Allman
International Computer Science Institute (ICSI)

At the heart of the Internet's unquestionable success is simplicity and flexibility that not only facilitates easy communication, but also fosters innovative applications. However, as innovation drives the Internet into ever-deeper corners of our everyday lives, the technological ecosystem underlying our Internet use becomes increasingly complex. To continue the evolution of the Internet requires a sound and accurate understanding of how the network works, making the topic of this special issue — Internet measurement — a crucial component of advancing the state of networking.

Research and operational communities have made many advances in our understanding of the Internet and networking through a multitude of measurement efforts over the years. While advances will continue, we identify three key challenges that empiricalists are increasingly facing: scale, opacity, and ethical issues. These obstacles represent key areas where new methodologies and approaches are crucially needed.

Scale

The Internet is rapidly expanding along many axes, including users, businesses, devices, houses, criminals, applications, connection technologies, protocols, and threats. The immense scale means that the system's behavior is highly variable, and therefore even a relatively large (in an everyday intuitive sense) number of observations might not accurately characterize the system.

While statistics teach us how to choose sample sizes to represent a population, certain assumptions about the underlying population (for example, the normality of a distribution) or the sampling process (such as the randomness) must hold to use these techniques. These assumptions do not necessarily hold for Internet measurements. Thus, often we are stuck between gathering too little data — which leaves us with a biased view — and expending a great deal of effort to gather a massive amount of data that "looks big enough" and therefore is seemingly beyond reproach.

The reality is that in both cases, we often have little understanding of a dataset's representativeness. Small datasets might be perfectly fine in some cases, while seemingly massive datasets might be biased in some fashion. The worst part is that we often lack the tools or methodology to answer these "how much data is enough" questions.

Opacity

The Internet's growth has also fueled ever-increasing complexity. This, in

Published by the IEEE Computer Society 1089-7801/16/$33.00 © 2016 IEEE

turn, makes designing measurement experiments and interpreting results challenging.

Originally, the Internet simply forwarded packets from a source to a destination. This made measurement a relatively straightforward task. However, as we have introduced complexity into the forwarding of traffic — for example, in terms of proxies, firewalls, caches, replicas, NATs, ad injectors, and performance enhancers — an observation at one point might bear little resemblance to an observation of the same traffic at a different point. For instance, there is little about a data stream that a recipient can directly ascribe to the presumed source, because some of the data could have been altered in transit. Furthermore, various players on the Internet intentionally try to make the situation more opaque. For instance, applications camouflage themselves to avoid being blocked or throttled and encrypt communication to avoid external observation (whether malicious or for innocuous research purposes), while ISPs block Internet Control Message Protocol (ICMP) messages to avoid exposing their infrastructures to external scans.

Whether it rises from complexity or intentional obfuscation, the Internet's opacity makes the process of soundly measuring the system immensely difficult for two reasons. First, we always need increasingly clever methodologies to infer the network's true operation. Second, inevitably these methodologies are not simple and straightforward, so they raise the logistical burden of conducting measurements (by, for example, requiring many measurements to ascertain some particular behavior and ascribe it to some actor in the system). This more opaque Internet poses a huge challenge for the measurement and continued evolution of the system.

Ethics

With the crucial role of the Internet in our everyday lives, the ethical considerations of our work as Internet empiricalists again are becoming increasingly important. While well-managed but (potentially) disruptive experiments and measurements were acceptable in the past, the implications of disrupting peoples' communication have become far greater, and therefore now must receive heightened scrutiny.

As a simple example, sending a single probe to an arbitrary remote host is highly unlikely to be disruptive. On the other hand, the odds are good that transmitting probes to an arbitrary host at 1 Gbps for an hour will be viewed as a highly disruptive attack. Although the two ends of the spectrum are clear, where to draw the line between "non-disruptive" and "disruptive" is at best difficult.

Additionally, we use the Internet to exchange ever-more private information. Therefore, even passive measurement that does not perturb the system now must undergo increased scrutiny to ensure that any personal information captured is handled in an appropriate manner.

Another important aspect of measurement that requires ethical foresight is in terms of dealing with side effects. The Internet has dramatically democratized information exchange, and in many cases, freed information from government and traditional media control. However, this has triggered broad, state-sponsored surveillance efforts all over the world. In a non-trivial number of places, even seemingly benign communication across the Internet is viewed as incriminating. At the same time, some of our measurements can make traffic appear to be coming from a particular computer. Therefore, we must exercise care in not conducting measurements that will implicate individuals in activity that is viewed as problematic, but in which they have no part. Increasingly, researchers involved in Internet measurement must consider the non-technical side effects of their work.

Whether it rises from complexity or intentional obfuscation, the Internet's opacity makes the process of soundly measuring the system immensely difficult.

In This Issue

This special issue attracted a large number of submissions. After several rounds of reviews and personal interactions with the authors, we selected five articles from 26 submissions. The selected articles provide a glimpse into diverse topics in this rich field of investigation.

Alok Tongaonkar's "A Look at the Mobile App Identification Landscape" provides a survey of methods that allow an ISP to understand which mobile applications generate certain traffic. ISPs need this information to monitor resource consumption by various applications, and to identify and block malicious activities. Yet assigning traffic to an application is challenging, because much of the traffic — regardless of the responsible application — runs over HTTPS (with encrypted payloads and common ports), and different applications might interact with overlapping sets of servers in the course of their operation.

"Measuring, Characterizing, and Avoiding Spam Traffic Costs" by Osvaldo Fonseca and his colleagues considers an interesting issue of which networks profit from, and which networks bear the cost of, delivering spam traffic through the Internet. The study measures the extent to which smaller networks bear the bulk of the cost of spam traffic delivery and sketches an algorithm that uses these measurements to identify profitable partnerships among networks for blocking spam.

Next, Glauber Gonçalves and his colleagues' article "The Impact of Content Sharing on Cloud Storage Bandwidth Consumption" focuses on traffic exchanged between an organization and a cloud storage service such as Dropbox. By analyzing traces collected at several vantage points, this study quantifies the amount of potentially avoidable traffic due to repeated updates downloaded from the cloud, either by the device that already has these updates or by multiple devices sharing the content. The article consequently investigates the use of a shared cache to eliminate some of this traffic.

"Empirical Study of Router IPv6 Interface Address Distributions" by Justin Rohrer and his colleagues addresses the issue of IPv6 router topology mapping. While topology measurements through traceroutes are routinely performed across the IPv4 address space, the size of IPv6 address space presents hard challenges to conducting such measurements. The present study performs exhaustive probes of every /48 prefix within every advertised /32 address block and uses the resulting dataset to analyze subnetting and address usage practices by IPv6 network providers.

The final article in our collection — "Cuckoo Cache: A Technique to Improve Flow Monitoring Throughput" by Salvatore Pontarelli and Pedro Reviriego — is not a measurement study in itself, but rather addresses technology that enables large-scale measurements. Specifically, it proposes an enhancement to Cuckoo hashing, an efficient approach to implementing hash tables. An efficient hash table is key to a wide range of high-volume network measurements. In particular, this article demonstrates the benefits of their enhancement on the example of traffic flow monitoring on a link, where each packet leads to an update of a per-flow state, such as the amount of data carried by the flow.

We thank everyone for their submissions. We also thank the large number of colleagues who reviewed the submissions for this special issue. This issue would not have been possible without the reviewers' time and expert opinions. We hope that *IC*'s readership will find these articles informative and enjoyable.

Michael Rabinovich is a professor in the Electrical Engineering and Computer Science Department at Case Western Reserve University. His research interests revolve around the Internet, especially concerning issues related to performance, measurement, and security. Rabinovich has a PhD in computer science from the University of Washington. He serves on the editorial boards of *IEEE Internet Computing* and *ACM Transactions on the Web*. Contact him at michael.rabinovich@cwru.edu.

Mark Allman is a senior scientist with the International Computer Science Institute (ICSI) and adjunct faculty in the Electrical Engineering and Computer Science Department at Case Western Reserve University. His current research focuses on network architecture, security, transport protocols, congestion control, and network measurement. Allman has an MS in computer science from Ohio University. He is a member of the ACM. Contact him at mallman@icir.org.