Moving From Opportunistic to Systematic Measurement

Mark Allman
International Computer Science Institute

1 THE WRONG QUESTION

The WOMBIR call is flawed from the start by posing the following as the first critical question: "what **data** is needed?".

The answer obviously depends on why one is collecting data. Are we interested in routing? DNS? Performance? Privacy? Censorship? Obviously in a global sense we are interested in all of these—and many more! So, rather than trying to arrive at some general notion of "what data?" we need to focus on myriad kinds of data and more crucially on how to collect data.

Further, we have been asking questions like "what data?", "what infrastructure?", "how to share?", etc. for decades now. Perhaps it should be telling that we are still asking these sorts of questions because we have not yet to come up with acceptable answers.

However, the goal of WOMBIR—as I understand it—is crucial. That is, we wish to understand the Internet through an empirical lens and, therefore, it behooves us to think about how we can improve that lens.

2 STATUS QUO: AD-HOC MEASUREMENT

Our current approach to building an empirical understanding of the Internet is fundamentally ad-hoc and *opportunistic*. For instance:

- We setup a border packet monitor at our own institution.
- We actively probe the Internet or some subset thereof.
- We probe from vantage points provided by "friends and family". I.e., whatever end-points we can scrounge together when we need measurements.
- We form some ad-hoc partnership with industry to get access to some data.
- We can monitor some infrastructure due to how the Internet developed (e.g., a DNS root server, a network telescope).
- We leverage systems setup by operators as part of running the network (e.g., RouteViews).
- Etc.

This ad-hoc approach means that even though we all have many questions about the Internet that measurement and data could shed light on, we tend to tackle problems for which we have data either on-hand or can gather relatively easily. Obviously, this approach has proven quite useful in building the community's understanding about the Internet's operation and behavior. However, this situation can also be maddening when we have questions for which no ready data is available to answer. Or, we have to end an investigation without digging as deeply as we would like into the underlying causes of some phenomenon because we do not have data to shed light on these puzzles.

Further, the Internet is constantly getting more complex. While the academic research community has proven itself to be quite clever at illuminating Internet behavior, our ad-hoc approaches are not well matched for the ever-increasing complexity. Industry has a role to play here due to their often privileged vantage point. While this can often shed light on the Internet's operation, the information is not necessarily openly shared. Of course, in some cases what is shared from industry in the open literature has been hugely useful. However, we all know of cases where a given company will not release information they view as having the potential to portray their business in a negative light. Therefore, what the broader community gets from industry is quite ad-hoc.

A final issue we have is that developing measurement infrastructure is often something researchers are interested in doing, however operating the infrastructure is not. We have seen many measurement systems and services die on the vine because researchers' main interest was in building the system and getting some specific measurements. However, once bit-rot and machine-rot sets in, researchers have moved on to some other idea/task/project.

3 FUTURES: A MORE SYSTEMATIC APPROACH

We are not suggesting we will ever remove the opportunistic and ad-hoc nature of some of our measurement activities. However, our position is that we should explore ways to make our measurement efforts more stable and systematic—i.e., less ad-hoc. This approach isn't somehow novel, radical or unwelcome to the research community. The problem with this approach is that *it costs money*! Until there is large investment in systematic measurement we will be stuck with a fragmented and ad-hoc understanding of how the Internet works.

There are no doubt countless ways to invest in better Internet measurement. The following is a list of notes meant to spur thought and discussion:

- Likely the easiest investments are in open tools that give fundamental new capabilities to everyone (e.g., *zmap*). Supporting these kinds of advances can happen through the traditional funding avenues.
- Fund long(ish)-term joint appointments between industry and academia. We have seen ad-hoc versions of this work well—e.g., faculty sabbaticals or faculty on partial leave at their own startups. However, these are fairly short term or specialized arrangements. Funding these sorts of arrangements could benefit a broader range of faculty and industry. Additionally, we can think about placing conditions on this sort of funding to ensure there is open release of information and this doesn't become a subsidy for industry.
- Fund specific journals and publish studies re-appraising previous results. Ensure we have curriculum resources to help faculty integrate re-appraisal into their classes (a la projects at Stanford and other universities).
- Subsidize Internet access for people who will host—passive and active—measurement infrastructure in their homes. (Obviously with a hefty amount of ethical due diligence involved.)

- Subsidize mobile Internet access for people willing to run measurement apps on their phones. (Obviously with a hefty amount of ethical due diligence involved.)
- Setup a system for sharing data from subsidized infrastructure within the community.
- Regardless of the specific structure, plan to invest for the long-term. I.e., stop thinking in terms of three year grants and instead think in terms of structures that will serve for a decade or more.
- Understand that much of this will take practitioners in addition to researchers. I.e., there will be aspects of these structures that hold no appeal to researchers but are crucial to maintaining solid measurement capabilities.