ACM SIGCOMM Student Mentoring Column

On Testbeds and Datasets

Dear students: This edition of the Student Mentoring Column focuses on various testbeds (for wired networking researching) and datasets. The questions below don't provide comprehensive coverage of either topic; as such, we may revisit them in future editions. I also hope to talk about wireless testbeds and datasets in a future column.

I got plenty of help in preparing this edition. In particlar, many thanks to *Aaron Gember-Jacobson (UW-Madison), Brighten Godfrey (UIUC), Ethan Katz-Bassett (USC), and Vyas Sekar (CMU).*

Testbeds:

Q: Is using Planetlab for my evaluation a good idea? Or, how do I pick a public testbed to use? And, what testbeds are available for public use?

A: The Internet is like Earth -- it has many different environments that are home to different animals. PlanetLab [1] is one environment with certain characteristics, like global reach, shared virtualized hosts, and often high bandwidth uplinks on university networks. Testing in PlanetLab is a good idea if those characteristics are similar to one of the environments for which you are proposing your system would be useful. Often, for a robust evaluation, you will want PlanetLab to be one of several environments that you test. For example, if you are working on a better TCP for inter-CDN transfers then PlanetLab could be a reasonable choice, but it would be a poor choice for a new data center TCP.

Another reason to test in multiple environments, one of which might be PlanetLab, is that you might not have access to the ideal environment to evaluate your particular work. You can partly overcome this hurdle by combining different environments that have different strengths and weaknesses. For example, PlanetLab gives you great control of the endhosts – you can run whatever you want – but may not be that diverse, since most hosts are in academic networks. Perhaps you can combine it with RIPE Atlas [2], which has much more diversity, but gives you much less control, in that you can only choose from a fixed set of measurements to issue. For some cases, performing controlled local experiments in your lab, in conjunction with at-scale experiments over public testbeds may be the right approach.

A more "meta" question that you may want to ask yourself before embarking on setting up an experiment on a testbed is whether your evaluation setup is representative enough for the question you are trying to address in the research and whether there are potential sources of bias. PlanetLab has served, and continues to serve, as an incredibly powerful platform for global measurements and running large-scale overlay services, but there are clearly aspects of the deployment (e.g., servers, bandwidth, network types, other users on the nodes, node reliability) that might introduce potential biases for the specific question you are answering. For instance, if you are doing work on intra-data center protocols or wireless network protocols, PlanetLab is not a good fit. Or if you are looking for running experiments of cloud providers, then PlanetLab nodes may be a poor substitute for actually running it using VMs in the cloud directly.

In addition to PlanetLab, there are many other testbeds that are great environments for conducting experiments. In most cases, your advisor simply needs to submit a request for access with a brief project description. Here is a subset of such testbeds. There are many others out there!

• CloudLab [3] and Chameleon [4] – CloudLab has three moderately-sized data centers with high-end servers, storage, and networking (SDN capable) are deployed at the University of Wisconsin-Madison, University of Utah, and Clemson University. The sites are connected by 100G links. Resources can be acquired in a similar fashion to Emulab. The Apt cluster, two Emulab sites, and a few other data center testbeds are federated with CloudLab.

Chameleon Cloud has two moderately-sized data centers with high-end servers, storage, and networking are deployed at the University of Chicago and University of Texas at Austin. Serves a similar purpose as CloudLab.

CloudLab and Chameleon are both incredibly useful for research into a variety of topics related to data centers, cloud computing, storage, and data analytics issues.

• GENI [5] – A federated testbed consisting primarily of racks of servers, storage, and switches deployed at universities across the U.S. Resources at different sites can be stitched together using the Internet2 backbone. GENI is an

excellent resource to experiment with disruptive ideas in a "global" setting.

• ESnet [6] – several high-bandwidth nationwide fiber or SDN networks.

• DETERlab [7] – infrastructure for cybersecurity research.

Q: It appears that most top papers these days seem to have some proprietary dataset. Is this good for the field? More pragmatically, how can I get access to such datasets for use in my own research?

A: I think this is a good topic for the community to think about, and I've been thinking about the role of data and proprietary data myself recently. I'm going to break this into a few questions.

(1) *Do most top papers have proprietary data?* Looking at all the award papers at SIGCOMM and NSDI in 2014 and 2015, a total of nine papers. Two of these papers (Conga at SIGCOMM 2014, and OVS at NSDI 2015) came from industry and benefited from this access, but none of the other seven papers required privileged access. This brief survey supports the observation that there are high profile papers based on proprietary datasets, but many good papers do not seem to need such data.

(2) *Is it good for the field to have lots of papers that rely on privileged data?* There is clear value in having papers that rely on data that is not publicly available, but it is important that the field maintains a balance between such papers and papers that do not require such access. An over reliance on such data might bias the field towards short-term, industry-focused thinking, and it might limit access to the field for people who lack the connections. However, real data can help us understand networks and systems in the wild, steer us to real problems, and evaluate research, and sometimes access to the required data is only possible via privileged channels. One signal of the value the field places on papers based on such data is that SIGCOMM and NSDI have recently added Experience and Operational Systems tracks. IMC offers awards that share valuable datasets. CAIDA and Crawdad provide resources for sharing and documenting data. The value of privileged data can vary by subfield. For example, academic researchers may not have insight into details of workloads in huge industry data centers, and so Microsoft/Google/Facebook workloads can have tremendous value. Any top paper needs to establish that the problem it focuses on is important and that the paper makes a contribution in solving the problem. Using real data is one way to demonstrate importance and contribution, but it is not the only way. It makes sense to pick problems based on whether you will be able to establish these properties given the resources you have available.

(3) *How does one get access to datasets?* There are various ways to obtain access to data. A common approach is to intern at a company. Another approach: ask; reaching out to someone you think has the data you need. You will be surprised how willing people are to share data. Your chances increase if you provide value back (by your research contributions, or simple preliminary analysis of the collected data, say) and if you have an existing relationship with the people you are asking. It can help to establish a presence at industry events, e.g., NANOG, RIPE, and IETF meetings, or on mailing lists. Talk about your work (both giving presentations and in hallway conversations) and, perhaps more importantly, listen to what problems others have. When you want data for a project, you can try to reach out to these contacts. Finally, given the value of data and the barriers to acquiring it, think about providing data as a way you can have impact with your work.

In many case, an exciting new project may start when you decide to develop techniques to measure data you needed for other projects but which could not be measured using existing techniques. And, whenever possible, contribute data to the community. Sharing data can amplify the value of a project. For example, Rocketfuel and iPlane have provided data used by many projects. It requires extra work to make data available and, especially, to refresh it over time, but, if you have a way to measure useful data, it is worth it. Your work will be judged by the impact it has, not (just) your publications.

There are pros and cons to using proprietary datasets in research. The size of proprietary datasets is often one of their main advantages. Large companies have vast computing infrastructure that can provide a wealth of data for motivating a problem or evaluating a solution. Conversely, a key disadvantage of proprietary datasets is the inability for other researchers to gain access to the same data. Without access to the same data, other researchers cannot validate your solution or compare their (improved) solution to yours.

Ideally, there would be large, non-proprietary datasets available for researchers to facilitate strong motivation, evaluation, and comparison. In fact, there are several such

datasets for specific problem spaces: MapReduce, cluster filesystem, and social networking data is available from Yahoo Labs [8]; Internet routing data is available from the Route Views Project [9]; the topologies of many service providers' networks are available from the Internet Topology Zoo [10]; traffic traces and Internet routing data is available from the Center for Applied Internet Data Analysis (CAIDA [11]); etc.

If there isn't a large, public dataset that meets your needs, you can explore several other options: (1) Assemble a smaller dataset by collecting data from a variety of public sources. For example, if you are interested in studying network configurations, you can download configuration data from all of the routers in Internet2. (2) Talk to network and data center operators at your university. They may be willing to share configuration or monitoring data, especially if your research could improve the campus computing infrastructure. They also likely know people in the same roles at other universities and might be able to help you get access to data from other universities. Make sure you plan ahead and request data from these individuals a few months in advance, as they likely have many tasks vying for their attention, and they may need to get approval to give you the data. (3) Find a collaborator who has access to a large, proprietary dataset. An internship at the company which has the data you're interested in is probably the most common route to achieve this. However, many students have gained access to proprietary data without ever having an internship by talking with individuals from the company at conferences and asking their advisors to reach out to people they knew.

References:

[1] PlanetLab: https://www.planet-lab.org

[2] RIPE Atlas: https://atlas.ripe.net

[3] CloudLab: https://www.cloudlab.us

[4] Chameleon Cloud: https://www.chameleoncloud.org/

[5] GENI: https://www.geni.net

[6] ESNet: https://www.es.net

[7] DeterLab: https://www.deterlab.net

[8] Yahoo! Datasets: https://webscope.sandbox.yahoo.com/#datasets

[9] Routeviews: http://www.routeviews.org

[10] Internet Topology Zoo: http://www.topology-zoo.org

[11] CAIDA: http://www.caida.org

Aditya Akella

University of Wisconsin-Madison, USA