Internet Measurements and Public Policy: Mind the Gap

Hadi Asghari
Delft University of Technology, Faculty of Technology Policy and Management
Michel J. G. van Eeten
Delft University of Technology, Faculty of Technology Policy and Management
Milton L. Mueller
Syracuse University, School of Information Studies
{ h.asghari, m.j.g.vaneeten, mueller } @ { tudelft.nl, syr.edu }

Abstract

Large and impressive data collection efforts often fail to make their data useful for answering policy questions. In this paper, we argue that this is due to a systematic gap between the ways measurement engineers think about their data, and how other disciplines typically make use of data. We recap our own efforts to use the data generated by a number of such projects to address questions of Internet and telecommunication policy, and based on our experience, propose five points for engineers to consider when building measurement systems to reduce the gap. Ignoring the gap means that fewer researchers use the data and significantly lowers a project's impact on policy debates and outcomes.

1. Introduction

Policy researchers in areas such as cybersecurity, privacy protection or network neutrality, can benefit substantially from large-scale empirical data, as findings based on global and longitudinal evidence are more reliable and insightful than those based on secondary sources and anecdotes.

Luckily, there is a substantial number of projects that generate large dataset that could potentially inform policy development in these areas. However, in our experience, a gap or mismatch exists between what measurement engineers tend to record, and what the social scientists, economists and policy researchers can typically consume. Consider packet dumps: for the measurement engineers they provide the ultimate accountability and flexibility to answer new questions in the future. For the policy researchers who typically answer questions around larger aggregates such as months or ASNs, the individual dumps are a big hurdle; they mandate downloading gigabytes to extract the few interesting pieces of information. One could judge this as merely a nuisance or lost machine time; but in practice, it might impose a serious barrier to further use of the data if a parsing tool has to be first written by the social-scientists to extract and interpret those interesting bytes.

Social scientists also deal with problems of measurement error and statistical validity of data samples in different ways than technical researchers.

This mismatch is not simply a matter of inconvenience for policy researchers; it directly undermines the potential policy impact of the measurement project. Computer scientists and engineers that build large-scale measurement tools often hope that their systems will impact Internet policy by increasing transparency in a particular realm. While this impact occasionally happens, more often than not valuable data never reaches the policy debate.

In this paper we shed light on this problem by briefly describing our team's efforts to estimate deep packet inspection deployment in ISPs worldwide using a measurement test called Glasnost [1]. Glasnost allows ordinary Internet users to detect whether their ISP is differentiating between flows of specific applications. We briefly discuss the challenges we faced while parsing, analyzing and interpreting the logs as an illustration of a public dataset that can be very informative for the policy discourse. The challenges where not unique to this measurement set, however. We discuss several other large measurement efforts later in the paper.

The contribution of this paper is to provide guidelines on how Internet measurements *could* be stored, structured and supported to make it easier to use for a wider range of researchers. The significance of this exercise is a two way discussion; one that obviously benefits the policy researchers; but also increases the chances that many more measurement projects can create the policy impacts that their designers had in mind.

2. Accessible Measurements

Understanding the gap between measurement experts and policy researchers

Internet measurement datasets can be an invaluable tool for policy research. They provide valuable empirical evidence to support the policy discussions in many areas, such as botnets, network neutrality, or SSL certificate use. That being said, it's important to note that developing the instruments and maintaining the infrastructure that runs and stores the measurements constitute only *half* the work required to use them for policy research.

The other half includes a mundane task of transforming the measurement logs into datasets that can be handled by common statistical packages; and finally, experimenting with models, adding independent variables, uncovering patterns, and validating and interpreting the results. Policy researchers would *prefer* to focus only on this last part, as that is where their main contribution lies. In practice however making sense of the raw data and the *transformation step* turns out to be extremely time-consuming. It is also this step that forms the gap between the two disciplines.

The gap is structural, rooted in the different requirements of computer scientists and policy researchers in the way they work and use data. For example, computer scientists often need to remain as close as possible to the raw data. This allows for accountability and validation. If an ISP denies deploying DPI, they can be presented with the packet dumps. It also makes it easier to mine the data in previously unthought-of ways in the future. In many cases, test results or interim reports are not even saved; the idea being that one could regenerate them from source at any point. In some measurement projects, historical data is not kept at all - you are offered the live feed and can decide to log it from now on.

As we shall see in the next section, the needs of policy researchers are different from these. Just to give some examples: researchers employing econometric techniques are interested in having a well-defined and consistent measurement approach as the starting point of their work; they prefer to work with observations spanning several years, and typically use organizations and companies as their unit of analysis (rather than individual machines). Robust, longitudinal, and aggregated, in contrast to flexible, real-time, and granular data.

Is it possible to have store and structure data in a fashion that addresses both sets of needs?

The "ideal" measurement set

In this section, we elaborate the needs of policy research, based on our experience in working with large measurement datasets. We shall in later sections assess how several other datasets compare with this criteria. Implementing all these suggestions might be hard and perhaps impractical in certain cases. They are meant as guidelines for measurement projects that want to increase their potential policy impact by enabling the analyses of other researchers.

1. ***Measurement sets ought to keep archives and will benefit from being up to date.*** This should really be seen as an entry level requirement, as policy research benefits most from looking at patterns over time.
2. ***Providing spatially and temporally aggregated versions of the data is helpful.*** Typical units of analysis in policy research include the organization and country level (versus individual IP addresses), and over periods of weeks, months, quarters and years (versus days). Making such aggregated versions of the data available for download will be very helpful, despite the drawbacks of duplication. Not only will it reduce download and processing times, but also resolve privacy issues with disclosing IP addresses, thus opening up the data for more researchers.[1]
3. ***Measurements ought to have clear verdicts and interpretations***. If the meaning of a particular measurement is ambiguous, life will be very hard for other researchers. It is very hard to interpret the results of a test created by others, as it forces a researcher with a different background to understand all details of a system they have not implemented. Anomalies and corner cases, most notably false positives/negatives, make this process even harder.

   Please note that we are not advocating oversimplification and binary verdicts; measurements will obviously many times be messy, just like the

[1] With regards to spatial-aggregation, geo-location and IP-to-ASN data from the time of the measurement should be used. If this data is not already stored along with the measurements, historical lookup databases can be consulted.

real world. In practice, this recommendation would mean having (i) good documentation regarding the typical and unusual cases; (ii) keeping the interim verdicts and reports (the complete-trace); (iii) providing parsers. As a last resort, supporting researchers attempting to use the data, via a mailing list or other means, can be highly beneficial.

4. ***Consistency of the measurement instrument over time is important.*** This recommendation is as important as it is difficult to execute. The difficulty comes from the fact that measurement researchers often experiment with various parameters in their systems to see which creates the most accurate results. Unfortunately, this practice can be very harmful for econometric analysis, as one cannot simply pool the results derived from the different measuring instruments together.

   Keeping parallel versions of the tests running while experiments are conducted might be one solution; changes should be well documented in any case. Monitoring of the testing infrastructure and its storage can also be crucial to avoid gaps in recorded data.[2]

5. ***Data collection should be organized in a manner that promotes sample validity.*** The number of measurements, and its balanced distribution over ASNs, countries or use cases, is extremely important for statistical validity. Guaranteeing this is again understandably very hard; enlisting the aid of researchers from other disciplines, e.g. interaction designers, might be very fruitful in incentivizing different user groups to participate in measurements.

3. Case I: Deep Packet Inspection & Glasnost

Researching DPI

We shall start by briefly describing the nature of our interest in using one of the datasets, Glasnost. Our research project involved the deployment and governance of deep packet inspection (DPI) technology. DPI can disrupt the Internet's end-to-end principle in both beneficial and controversial ways, e.g., thwarting spam and malware, rationing bandwidth, blocking access to censored content, and building user profiles for advertisers. With its dual potential, DPI use is contested politically [2]. In broad terms, we wanted to find out what impact this new technological capability has on how states and companies govern the Internet.

This leads to a set of sub-questions, such as which Internet providers are using DPI, to what extent, and how they respond to various regulations and laws aimed at privacy, network neutrality and censorship.

How could we answer these questions? One way to get this information is to ask the operators, but as a research method that is quite problematic. It would be costly and time consuming to survey all Internet service providers, many of them will refrain from responding, and for the others we have to doubt the truthfulness of their answers. So an alternative strategy is for network users to run tests that reveal what is actually happening to their traffic. This is precisely what the *Glasnost* test does. Glasnost, developed by researchers at the Max Plank Institute, enables users to detect blocking or throttling of BitTorrent and other protocols by their access provider, and whether this is done using DPI or the TCP port. (The detailed workings are described in [3]). The M-Lab platform gave Glasnost a global reach, with the test being run thousands of times by users from 2009 to 2012. The test-logs are stored and made publicly available. All of this seems ideal for our research purposes.

Evaluating Glasnost data

In this section of the paper we will describe the steps involved in processing the glasnost logs into the dataset suitable for our research. Our initial expectation was that this should be relatively straightforward, but this turned out to be far from the case. We compare Glasnost data against the guidelines laid out earlier. This is not meant to criticize the Glasnost project, but to better understand the issues that many projects seeking policy impact face.

Archives & ongoing logging

The Glasnost test has been on M-Lab platform [4] since early 2009, and it is still maintained and live.

[2] A commitment to maintain the tests over time might prove costly or infeasible for many projects. Deploying the tests on shared and open platforms such as M-Lab could be one way to reduce maintenance costs for such cases.

Level of aggregation

The Glasnost data is stored at the level of individual tests on Google Storage. For each test, a server log and a packet dump are stored, of which only the log is useful in our work. Furthermore, out of each log, only the header and a few summary lines at the end are needed. This means that much more data needs to be downloaded than is actually needed, even in the absence of spatially and temporally aggregate data. For example, for just February 2012, 115 GB of compressed data has to be downloaded. This take several hours to download on a campus gigabit connection due to the way Google storage functions. After extraction and removal of the packet-dumps, we are left with 33 GB of uncompressed data. Extracting the useful lines brings us down to 40 MB for per-test metrics, which is less than 0.1% of the downloaded data. Geo location and ASN information is also not stored with the data, making it necessary to use historical records for accuracy.

Turning logs to verdicts

We wished to have a verdict for each test run: does it indicate the presence of application based throttling or blocking (hence, DPI), or not? Although these results are shown to the user when they conduct the test, they are not stored in the logs. This made it necessary to parse and analyze the logs to re-calculate the result.

This turns out to be very involved. The test logs store details information about each measurement flow. The Max Plank researchers provided a parser that works on logs after May 2010; for the first batch of log files, we had to write our own parser by reverse engineering the Glasnost server code. The parsers took us only part of the way: it provides separate verdicts for upload, download, throttling and blocking, while we required *combining* them for a final verdict. Due to anomalies, corner cases and scarce documentation, this was not straightforward.

We will briefly explain these steps, but not bother the reader with all the complexities. In short, Glasnost works by recording and comparing the speeds of several network flows. Data is transferred between the client and server using different protocols (the application being tested versus a random bit-stream) and on different TCP ports (the application assigned port versus a neutral port), and detected interruptions are also recorded. If for instance the speed of the application flow is much slower than the control flow, it can be concluded that the ISP is performing application-based throttling. Since the Internet routes traffic on a best-effort basis, speed fluctuations are normal. The test developers came up with thresholds that could indicate speed differences reliably, and also determine if a connection is too *noisy* to make any inferences. The cases get more complicated when considering noisy flows, and *failed flows* (when zero bytes are transferred). Further complications include a large number of aborted tests, and tests with anomalous results, for instance where the BitTorrent flow is found to be significantly faster than the control flow which does not make sense. These yield a large number of possible combinations that required us to decide on the verdict for each combination.

This got us involved in the details of the workings of the measurement tool, took over two months of full time work to accomplish, and was very far removed from the policy research we planned to conduct.

Consistency over time

The Glasnost data has several discontinuities in the logs. First, the log formats changed three times. This change did not cause any major statistical problems, only extra parsing work. The second discontinuity was statistically relevant, however, and was caused by changes to the default parameters used by Glasnost, including the flow durations, repetitions, and directions. These were changed mid-2009 to find the optimal settings that yielded the fewest false results without making the test too long. These changes create very visible jumps in the results' percentages. (As Figure 1 shows, the number of false positives and negatives drops considerably after October 2009). The third discontinuity, also significant, was that for several periods in 2010 and 2011, the longest of which spans several weeks, no test logs exist. This was due to an unfortunate *rsync* problem between M-Lab and Google Storage, resulting in data loss.

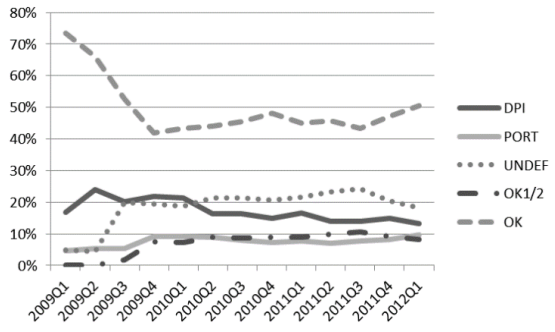

Figure 1 - Test verdict percentages over time

Issues with sample size

One cannot blame a crowd-sourced project for inadequate sample size, but can think of remedies. Two specific problems of sample size exist in the Glasnost data. First, although the test has supported testing different protocols since May 2010, BitTorrent still makes up the vast majority of the tests, because it was set as the default choice on the test interface. The other issue is regarding ASNs with 10 or less tests over a period, where one false positive or negative can make a large difference in the results of that ASN. A future remedy for both situations would be that the M-Lab website recommends visitors to run Glasnost when the total number of tests from that visitor's ASN is low, and for it to set the default protocol similarly.

On the plus side, Google made good efforts to publicize the M-Lab initiative generally. The number of tests conducted directly responds to these publicity efforts. When those efforts dwindled, the number of users conducting tests goes down.

Support infrastructure

Overall, the M-Lab team makes a strong effort to ease use of the Glasnost data. The support included an active mailing list, access to base parsing scripts. In terms of documentation, the original Glasnost paper provided a good starting point, but as mentioned more was desirable. The test authors also clarified some issues at one or two points. Over time, the quality of support improved, indicating a learning process for all parties. This was in our opinion what enabled our project to eventually succeed.

At this point and after much work, we have our "cleaned" dataset – one that is ready for econometric analysis. In the ideal situation, this would have been our starting point.

4. Other Cases

We compare Glasnost to four other measurement projects in Table 1. These include a spam trap, the SANS DShield database, the EFF SSL-observatory and finally the Conficker sinkhole, all of which have provided valuable input to policy research.

These datasets benefit from being (mostly) longitudinal and enjoy a relatively good sample size. The aggregation level is unfitting in two of the five cases; the verdicts lack clarity in two of the five cases. The discontinuity problem exists in two of the five. The main take away message here is that the guidelines can be used to evaluate all these large datasets; the criteria are meaningful and of value in all the cases.

Table 1 – Evaluating five measurement sets

	Time Period	Aggregation & format	Logs to verdicts	Consistency	Sample size
Glasnost [1]	2009-now	**(-)** Individual test logs	(-) Parsing involves many steps	(-) Multiple discontinuities	(+/-) Mixed
Spamtrap	2005-now	(+/-) Logs, daily. IP based.	(+) Relatively clear, excellent support	(+) Yes	(+) Good
DShield [5]	2006-now	(+/-) Logs, daily. IP & port.	(+/-) False positives still unclear	(+) Yes	(+) Good
SSL-Observatory [6]	Only 2010	(+) SQL dump; Certificates	(+) Good documentation and support	N/A	(+/-) Full, but once
Conficker sinkhole [7]	2009-now	**(-)** Logs, hourly; per connect	(+) Relatively Clear	(-) Log format change	(+) Full

5. Related Work

The Internet measurement community already appreciates a number of strategies for what they call "sound measurements". Our guidelines are comparable in many points. For example, Paxson states the importance of keeping the complete audit-trial, tying this to the need for reinterpreting the data at a future time when the original rich research context is forgotten [8]; now, compare this to our criteria of clear verdicts and interpretations. In this literature stream, our work simply reiterates and highlights some of the already known sound strategies that are important for interdisciplinary research.

On a different note lies the benefits and hardships of interdisciplinary research. Thuraisingham talks specifically about reasons to pursue research projects between computer science and the social sciences [9] and highlights a number of challenges along the way, e.g. that computer scientists need to be ready to develop new tools and avoid one-size-fits-all solutions. The National Academies book "Facilitating Interdisciplinary Research" provides a long list of key conditions for interdisciplinary work [10]. Most of them are essentially linked to conversations, connections and combinations

among teams of different disciplines, which in our opinion is similar to the gap presented in this paper.

6. Summary and Discussion

This paper developed a five part framework to guide computer scientists' design of large-scale data collection efforts so that they can be useful for social science and policy work. These guidelines are as follows:

1. Measurement sets ought to keep archives and will benefit from being up to date.
2. Providing spatially and temporally aggregated versions of the data is helpful.
3. Measurements ought to have clear verdicts and interpretations.
4. Consistency of the measurement instrument over time is important.
5. Data collection should be organized in a manner that promotes sample validity.

As a case study, we highlighted the challenges faced by our own team in using a number of datasets, including Glasnost. The dataset was invaluable, and provided us with empirical insights into deep packet inspection use that would otherwise remain unanswered. However, due to the way Glasnost logs are stored and structured, we were forced to spend considerable time upfront to process the logs into a suitable format. For many social scientist, this hurdle will be insurmountable not only because of time constraints, but because they lack access to the technical competencies that are needed to move forward. This was a major distraction from policy research and in our opinion represents a structural dissonance between the different goals of computer scientists and policy researchers in using Internet measurements. Although data transformation is inevitable, the complexity and amount of time required to do so directly impacts the number of research teams willing to use a particular dataset. The five-part framework is an attempt to close this gap.

As the final paragraphs, we would like to address the number one comment we have received regarding the framework: "you make suggestions without consideration of the costs involved". We have two responses to this comment. First, one could look at incurred costs as simply the costs of doing business. This of course depends on the aims of the measurement project owners: if they wish their work to have an impact on telecommunication policy and related fields, and to aid policy makers to make informed decisions based on data, which we suspect to be often the case, then bearing these costs will be simply a necessity. Its basic economics: lowering the costs involved in using your data means more researchers will use it, thus increasing the odds of it being impactful.

The second response is that implementing the issues will not be as costly as one thinks; the objective is to lower the barriers to entry, not eradicate them. Simply thinking about the issues in the design and deployment phase of measurement projects will already accomplish much. *Hallway usability testing* [11], championed as a common sense approach to software engineering, can also work here. A hallway usability test is where you grab the next person that passes by in the hallway and force them to try to use the code you just wrote. We could expand the idea as persuading the next econometrician passing through the department to use the measurement sets, and resolving the top issues faced. This should neither be formal nor costly; it might even turn out to be fun.

References

[1] MPI. (Accessed April 1, 2013). *Glasnost: Test if your ISP is shaping your traffic*. Available: http://broadband.mpi-sws.org/transparency/bttest-mlab.php

[2] M. Mueller, A. Kuehn, and S. M. Santoso, "DPI and copyright protection: A comparison of EU, US and China," 2011.

[3] M. Dischinger, M. Marcon, S. Guha, K. P. Gummadi, R. Mahajan, and S. Saroiu, "Glasnost: Enabling end users to detect traffic differentiation," in *Proceedings of the 7th USENIX conference on Networked systems design and implementation*, 2010, pp. 27-27.

[4] M-Lab. (Accessed April 1, 2013). *About Measurement Lab*. Available: http://www.measurementlab.net/about

[5] DShield. (Accessed April 1, 2013). *About the Internet Storm Center*. Available: http://www.dshield.org/about.html

[6] EFF. (Accessed April 1, 2013). *The EFF SSL Observatory*. Available: https://www.eff.org/observatory

[7] CWG. (Accessed April 1, 2013). *Conficker Working Group*. Available: http://www.confickerworkinggroup.org/wiki/

[8] V. Paxson, "Strategies for sound internet measurement," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 263-271.

[9] B. Thuraisingham. (2009, accessed April 1, 2013). Why is Interdisciplinary Research Hard. Available: http://www.utdallas.edu/~bxt043000/Motivational-Articles/Why_is_Interdisciplinary_Research_Hard.pdf

[10] *Facilitating Interdisciplinary Research*: The National Academies Press, 2004.

[11] J. Spolsky. (2000, accessed April 1, 2013). The Joel Test: 12 Steps to Better Code. *Joel on Software*. Available: http://www.joelonsoftware.com/articles/fog0000000043.html