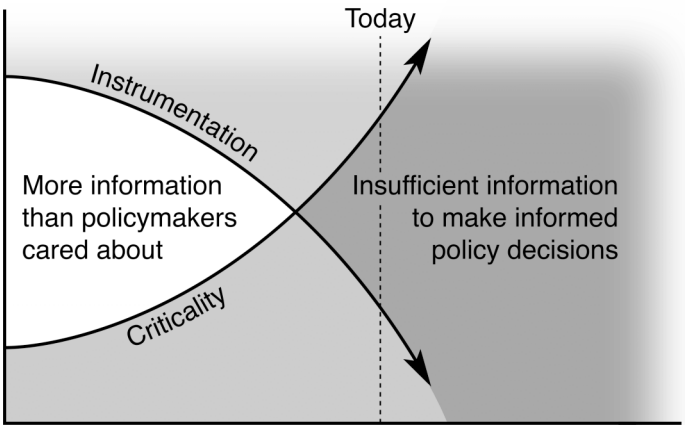
Instrumenting the Internet to support the development of informed policy

Position paper for NSF/OECD meeting
Bill Woodcock
January 31, 2007

In the Internet's nascence, its quantitative features were easily grasped. Traffic flowed through a single well-instrumented backbone, the whole scope of the network was small, and it occupied a well-understood subset of the world's institutions. Those institutions were largely research-oriented and were populated, to no small degree, by people whose attention was devoted to understanding the network. Concerns of competition, customers, and finance did not exist in the private-sector sense that we would appreciate them today. The portion of the network which existed outside of that well-illuminated sphere was small, relatively easily estimated, and not terribly significant.

The privatization of the core of the network, as envisioned in the National Information Infrastructure and carried out in the beginning of the 1990s, turned this on its head. In the years which followed, the Internet grew through private-sector investment in commercial Internet service providers and carriers, most of whom had little interest in instrumenting their networks beyond what was needed for immediate network operations and capacity planning, and no interest whatsoever in sharing or publicizing that information. Things grew even worse at the end of the decade, when the telecommunications investment bubble arrived, and huge amounts of new capacity were added to the network in a seat-of-the-pants way, rather than based upon measurement, needs-analysis and capacity-planning, and these new additions were completely uninstrumented. Indeed, instrumentation, which might reveal less-than-hyperbolic growth, was actively shunned, and psuedo-scientific pronouncements about the size, growth, and importance of the Internet strayed further and further from any substantive basis. In the wake of the crash which followed, surviving carriers found themselves in a virtual ghost-town of overbuilt facilities, and capacity planning was the second-to-last thing on their minds, just ahead of instrumenting their networks to gather data for capacity planning.

In the mean time, the Internet has grown to occupy a position of criticality in the developed world, and has thus moved to the forefront of governmental and economic policy-making consideration. We have passed the crossing-point of the downward trend of understanding and the upward-trend of needs-to-know, and this has left us in the most unfortunate position of having a large and unmet need for information which simply isn't being collected.



Because the data in question is transient, every day which goes by is a day's worth of data which the world will never have. Merely instrumenting the Internet now would be sufficient to solve the long-term problem, since data which has at least been captured can always be analyzed by those who come later.

That instrumentation need be neither complex nor expensive. It's already in place in some Internet exchange points, and all we need to do is encourage the remainder to adopt the best-practices of their industry. This would provide the information we need to assure industry and Internet users at large that their interests are being looked after, that the magnitude of threats can be determined, that infrastructure providers will know when to build more capacity, and that policy-makers will have the input they need to allocate public resources.

Questions which I propose to address:

1) **How can growth in and means of Internet resource production, and correspondence of sources and destinations of traffic, be measured on the commercial Internet?**

   We propose to regularize, institutionalize, and expand the two methods which have proven most successful to-date: documentation of IXP participation and Netflow sampling. Taken together, these two methods provide insight into the topology of the Internet, and the size of its parts. By measuring the total volume of traffic through an IXP, knowing who all of its participants are, and sampling the traffic of a statistically-significant set of those participants, one can extrapolate a floor for publicly-exchanged traffic. By correlating IXP participation and Netflow matrixes with two of the additional datasets which we collect, observed topology and inferred relationships, one can further begin to extrapolate the volume of Internet traffic being generated outside the purview of IXPs.

2) **What are the benefits of greater instrumentation, understanding, and transparency to industry, consumers, policy makers, researchers, and for the future of the Internet?**

   Presently, the very most basic questions about the Internet are unanswerable. How large is

it? How fast is it growing? Where is it growing? How much of the Internet's bandwidth is produced at Internet exchange points, and how much is produced within individual carriers? These questions are the fundamental inputs to sound policy-making and economic planning. With answers to these questions, public-sector resources and policy could reinforce accurately-targeted private-sector investment to maximize the growth and value of the most critical utility of the twenty-first century. Well-informed development will increase the reliability and speed, and decrease the cost of communications services. This in turn will lubricate the general economic machine, and allow markets to more transparently and efficiently serve the world's peoples. More specifically, network service providers need accurate information to plan infrastructure expansion; regulators need accurate information to ensure that the effects of competition are optimized; consumers need accurate information to help them select services and have confidence in the market's ability to provide a reliable utility; and the academic community needs broad insight into the workings of the network in order to make the intellectual leaps necessary to advance the state of the art. None of these needs are presently served, because none of this information is being collected.

3) How can the privacy concerns of individuals and industry be balanced with the information requirements of the public sector and academia?

When new information becomes available, there is always the possibility of overshooting the mark: releasing proprietary information to those who would abuse it, or who would take advantage of the window of time in which the new information is not fully distributed and digested, to arbitrage their greater knowledge to the detriment of the public, or of those who have made the new information available. In the case of Internet traffic statistics, the principle dangers are competitive and regulatory. If overly-detailed information about specific individual carriers finds its way into the hands of their competitors, this information may, if it is unflattering, be used to discredit them in the marketplace, thereby disincentivising them to make further information available. Likewise, if overly-specific information makes its way into the hands of those who would stifle growth by applying punitive use-based charges and taxes to users or providers, the Internet would be harmed as well. It is thus critical to collect only the minimum necessary information about traffic flows, and to aggregate the information of multiple carriers together to anonymize it, before publication. This is the current best-practice at Internet exchange points, and has proved palatable and safe over the past fifteen years.