

measurements and signal processing of Internet data

// the wonderful thing about science is that eventually
nature tells you when you are fooling yourself. real
objects can be measured again and measured by somebody
else -- false signals will eventually be weeded out. //

-- robert kirshner _the extravagant universe_

kc claffy, ucsd/sdsc/caida
november 2004
kc@caida.org
www.caida.org

challenge: characterize Internet traffic trends

more precisely: correlating heterogeneous measurement data to achieve system-level analysis of Internet traffic trends

motivation: lack of data since 1995

another motivation: way too much data

(s/data/signal/g)

challenge: correlating Internet data

problem: even getting signal to process can suck up careers

- admissions about dealing with Internet data
 - vern's 2001 talk www.icir.org/vern/talks/vp-nrdm01.ps.gz
 - david moore's 2002 talk www.caida.org/publications/presentations/2002/ipam0203/
 - vern's 2004 imc paper <http://www.icir.org/vern/papers/meas-strategies-imc04.pdf>
- longitudinal data are highly ad hoc
- measurement tools lie to us
 - packet filters, clocks, "simple" tools...
 - no culture of calibration
- measurements carry no indication of quality
 - lack of auxiliary information
- measurements are not representative
 - there is no such thing as **typical**
- analysis results are not reproducible
- large-scale measurements are required
 - that overwhelm our home-brew data management
- we do not know how to measure real traffic (topology, routing)
 - or in some cases we know we can't
 - not that this stops us from measuring something else

just so i don't understate the case

- for the most part we really have no idea what's on the network
 - can't measure topology effectively in either direction. at any layer.
 - can't track propagation of a bgp update across the Internet
 - can't get router to give you its whole RIB, just FIB (best routes)
 - can't get precise one-way delay from two places on the Internet
 - can't get an hour of packets from the core
 - can't get accurate flow counts from the core
 - can't get anything from the core with real addresses in it
 - can't get topology of core
 - can't get accurate bandwidth or capacity info
 - not even along a path much less per link
 - SNMP an albatross (it has inspired being envy of telcos)
 - no 'why' tool: what's causing my current problem?
 - privacy/legal issues disincen research
 - result --> meager shadow of careening ecosystem
 - result --> discouraged (or worse) academics
- signal processing, ha, most of us just signal seeking

obstacles to Internet/network research

where is the data?

- Internet grew organically, incorporating useful technologies as less useful ones obsolesced
- scientifically rigorous monitoring & instrumentation not included in post-NSFNET Internet
- data often proprietary; research use outside owning administrative domain is rare
- researchers can't find out about what little data **is** available
- Internet research fundamentally different from physics/biology/chemistry -- although we have their problems as well
- more like astronomy w/ no national virtual observatory or even decent telescope
- or early quantum mechanics
 - in that you can't measure the particles when you need to

requires sophisticated tools And special access to data

obstacles to Internet/network research

problems caused by lack of data

- results with predictive power elusive since every link/node has its own idiosyncracies/policies
- makes it hard to assess the quality of any result
- fundamental research cannot be accomplished
- tools designed to combat major problems cannot be tested
 - DoS attack mitigation
 - virus/worm spread
- can't validate theory, model, or simulation against real network
 - not to mention code bugs, methodology flaws

result: weak Internet science

- it's not just soft, it's slippery
- and stunted
- i've just stopped reviewing papers that don't even try

100 year view of network science

The modern field of elementary particle physics depended crucially on the establishment of a huge volume of data gathered mainly in the period 1945-65. Only then was it possible for the synthesis of the Standard Model to take place (1967-74).

-- Peter Galison, Harvard Professor of History, Physics

risk analysis

reminder:
the risk of being hurt by lack
of signal processing techniques
is less than the risk of being hurt
by lack of relevant signal to process

explosion in interest in measurement notwithstanding,
we need to be honest with ourselves (and reporters,
and the government) regarding what we really know
about the Internet

- how much spam? viruses? p2p traffic?
- how many routers?
- how much bandwidth?
- how much encryption?
- how many hops?
- what does global peering topology look like?
- what does router/IPv6/multicast/dns/etc topology look like?
- we have absolutely no idea

more sophisticated approach to data analysis inevitable

why?: the data we do have

- disparate
- incoherent
- limited in scope
- scattered
- unindexed

what we do not have

- rational architectures for data collection
- instrumentation suitable for above OC48 links (that number tends to grow..)
- archiving and disseminating capabilities
- data mining and visualization tools for use in (nearly) real time?
- historic data for baseline
- cross-domain analysis of multiple independent data sets
- local phenomena vs. global behavior

what we need

- creativity
- persistence
- epistemological perspective

epistemological questions

- what is the right way to describe data?
- what is the data's description complexity?
- how can we reduce descriptions?
- which parameters can affect data variation?
- can we prove independence from a parameter?
- how much dependence is between parameters?
- how do we scan experiment design space?

CAIDA's vested interest in this workshop

- trends project (repository for correlation of heterogeneous data)
 - plan for long-term and sustained support of such a repository
 - examples in other sciences, e.g., chemistry's Protein Data Bank
- improved timestamps mechanism (see darryl's talk/code)
- overhaul of skitter project (better use of resources?)
- getting better signal (measurement) in general

look for Internet Measurement Data Catalog alpha release in early 2005

- <https://www.caida.org/projects/trends/>

appendix: meta-commentary on Internet analysis

end game: legitimate analysis of trends

- caveat: trends are really not good
- the more we see, the less we like
- see kc's talk '[top problems of the Internet & how researchers can help](#)'
- grep for garbage in bruce sterling's nsf april 2004 grand challenge workshop keynote talk
 - <http://www.cra.org/Activities/grand.challenges/sterling.html>
 - exceptionally worth reading anyway
- "digital imprimatur" -- john walker
 - <http://www.fourmilab.ch/documents/digital-imprimatur/>
 - "how big brother and big media can put the Internet genie back in the bottle"
 - rich 'optimistic pessimism'
- geoff huston's nznog talk
 - video <http://s2.r2.co.nz/20040129/>
 - slides <http://www.nznog.org/ghuston-trashing.pdf>
 - not so much with the optimism

reminder: IMDC's website (neutral about falling sky)

- <http://www.caida.org/project/trends/>