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

// in an expanding system, such as a growing organism,
freedom to change the pattern of performance is
one of the intrinsic properties of the organism itself //

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

challenge: characterize Internet traffic trends

motivation: lack of data since 1995
another motivation: way too much data

- 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/outreach/presentations/2002/ipam0203/
- 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

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 just an albatross (enough to inspire telco envy)
- no ’why’ tool: what’s causing my current problem?
- privacy/legal issues disincent research
- result --> meager shadow of careening ecosystem
- result --> discouraged (or worse) academics

if you’re not scared i’m not explaining this right

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 -- there are organisms and molecules and atoms all over the place to study
- more like astronomy (with no national virtual observatory)
  - or even decent telescopes
- or early quantum mechanics
  - in that you can't measure the particles when you need to

requires sophisticated tools And special access to data

the view from here

the data we do have

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

what we need

- globally relevant measurements
  - 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

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
- no revolutionary progress in the field for years
- and most of us are partial to revolution

not helping matters

- PACI tension betw. production support & pushing technology envelope
  - unfortunately PACI is not unique there
- artificial distinction between infrastructure and research
  - (Atkins complaint of original PACI program)
  - some of us have been whining about this for years

but ‘informational science’ is now an essential cyberinfrastructure goal

- sounds good to me
- need to assume that if we fail on this one, nsf gets nicked
- we’re need to secure such a field about the Internet itself
  - perfect example of field with so much information that meta-information becomes vital
  - (mark’s great thursday quote about ‘wait, why aren’t you just sharing the files? is there any other way [to make progress?]’ -- from napster ‘generation why’ kid)

what can be done

find way to fund researchers to share data

- scarce time and resources are required to share public data with other researchers
  - answering queries
  - providing data
  - answering inevitable questions about the data
- make a data catalog of available data sources -- a single clearinghouse for information on available data sets

need ‘well-curated’ Internet measurement data repository

- Atkins report goal (i recently learned)
- measurements need pedigrees describing them, how to navigate
- audit trails, portable analysis scripting language to support reproducibility
- well-managed meta-data
- understand sampling implications and technology better
- anonymization tools & reduction agents

Atkins report vision of such a repository

increasingly important to science and engineering research

- long-term and sustained support of such repositories
- more than simply running large storage facilities
- supported by research into cyberinfrastructure
- better ways to organize and manage large repositories
  - metadata (machine readable and searchable)
  - dynamics, reclassification supported
- software tools to analyze
- standards to allow data to be self-documenting/discoverable automatically insure interoperability necessary to use data across disciplines
- high speed access
  - network, storage, I/O subsystem issues

btw, much here already been/being solved by google, amazon, orkut

- tech transfer might should go both ways

CAIDA trends project: mission

establish meta-repository for network measurement data

- facilitate access to raw data
- enable testing of analitic methodologies
- publicize, promote, and implement the results
- long-term storage of data
- bring together researchers and developers

create universal annotation system

- applicable to various heterogeneous data sets
- enables cross-correlation and comparative analysis
- convenient to navigate
- indispensable for large distributed data bases

framework to help ’cultivate culture of, and passion for, sound measurement, as science and discipline’ [-- vern’s talk]

CAIDA trends project: approach

The Internet Measurement Data Catalog

- single source for information about data including:
  - who created it
  - how it was collected
  - when it was collected
  - where it is stored
  - access policies
  - format, packaging, and compression
  - annotations to allow known features and problems with the data to be shared with other users
- eventual expansion to include
  - mapping data to tools that read/write it
  - grouping related data
  - tagging data and tools used to do published research

trends project: requirements

most important: receptive to community input

- maximally representative data sets
  - traces
  - active probing
  - routing information
  - geographic data
  - bandwidth measurements
  - ? what else ?
- strategic approach to sampling of traffic
  - reality: we can’t capture it all
  - monitor high bandwidth commodity backbone links ("core")
  - define schedules and durations
  - implement high-precision clock synchronization
  - collect long bidirectional traces
  - make collection process application specific as necessary
  - ? what else ?

trends project: tasks

- deploy strategic Internet measurement instrumentation
- improve measurement tools
  - advanced hardware for monitoring OC48 links
  - advanced software for pre-processing the data various levels of aggregation
  - modules for storage and manipulation of data
  - expand security related monitoring
    - ability to capture DoS attacks in progress
- develop and support a large data storage infrastructure at SDSC
- coordinate movement of traffic measurement data
- create multi-faceted sets of data (datakits)
- universal annotation system (next slide)

trends project: universal annotation system

requirements

- accomodate heterogeneous raw data sets
- handle data sets distributed among many sites
- facilitate community access to data repositories
  - data sharing and comparative analysis
- flexible and extensible
  - define meaningful data cross-mappings
- community-based approach to develop common formats
- encourage wide use of common formats
- leave control and security issues to data owners
- ? what else ?

present state of knowledge

- none for the Internet community
- draw from other sciences
  - biology, physics, astronomy

trends project: universal annotation system (2)

tasks

- create front-end user interface
  - Internet access to data
  - APIs
  - AUPs
  - compatibility with collection-based sofrtware
- create back end information management system
  - automatic methods of indexing
  - include: data, tools, analysis requests
  - distributed data collection and publication
- maintain and develop compelling tools
  - responsive to user needs
- solicit input from concerned research and standards groups
  - Grid Forum, IETF (IPFIX, IPPM, PSAMP), IRTF (IMRG)
  - NANOG, ISP community (security issues)

expected users of IMDC

- CAIDA currently receives dozens of queries for data every week
- CAIDA has available [|soon] hundreds of gigabytes of data, including:
- anonymized and unanonymized OC48 backbone traces
- network telescope data including:
  - host scan dynamics
  - the spread of Internet worms
  - Denial-of-Service backscatter
- making CAIDA data searchable via IMDC will encourage people to use

we've attempted a compromise between requiring so much context for contributed data that no one will contribute, and requiring so little background that searches don't provide meaningful information

expected uses of IMDC

cornerstone of [inter]national Internet observatory

- more than abilene observatory
  - http://abilene.internet2.edu/observatory/proposal-process.html
    - support collection and dissemination of abilene data
    - operational view of large-scale network
    - data on fundamental properties of network
  - which is .05% of what we need
    - (11 nodes. 2 racks. no commodity peering, address structure gone)

IMDC: application to current research problems

each research question requires:

- research plan
  - identification of data sources
  - scope of required data
    - specified time period
    - particular topology
    - certain physical link
  - analysis techniques
- implementation
  - preparation of data
    - selection
    - cleansing
  - analysis steps
    - data mining
    - scripts
  - publishing results
  - visualizations
  - web pages
  - articles

IMDC: research problems (cont.)

example: workload trends

- patterns of usage over time
- pace of new protocols' deployment
- growth of tunneling technologies
  - impact on fragmentation
- more users or more traffic per user?
  - per host, prefix, site, AS
- behavioral characteristics
  - for classification
  - for engineering purposes
- comparison of various flow models
- traffic load and geography
  - local
  - regional
  - international
- tracking distributed denial-of-service activity

expected uses of IMDC

exploding myths

- e.g., RIAA claimed in august "P2P traffic dropped"
  - http://www.pewinternet.org/reports/pdfs/PIP_File_Swapping_Memo_0104.pdf
  - march/may 2003 -> december 2003 brought 29% -> 14% "usage"
  - data sources: telephone surveys nov18->dec14 (huh?); software downloads
  - not data sources: Internet data (wth?)

real data

- have never seen a trace at time t with less p2p traffic than at time t-1
  - frankly i don't see that happening soon

being able to verify/refute this claim is actually a huge deal

- (and not just about changing how we must think of ownership of everything that comes out of our brains)
- will change Internet engineering as we know it today
- current stability and profitability/usability assumptions of asymmetric utilization
  - (btw also driving community to re-evaluate issues of privacy and anonymity;
  - won't ever see a p2p protocol again that doesn't support encryption)

trends project: meta-commentary

end game: legitimate tracking of trends

- caveat: trends are really not very 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 sterlings's nsf april 2004 grand challenge workshop keynote talk
  - http://www.cra.org/Activities/grand.challenges/sterling.html
  - exceptionally worth reading anyway
- "digital imprimateur" -- 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

this project's website (neutral about falling sky)

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

IMDC: interim progress (14 months in)

- short answer: not done yet
  - design process complete, including user interface
  - database configured and functional
  - prototype implementation in progress
- medium answer: impediments on our minds
  - ineffective data cataloging
  - disparate formats
  - inadequate documentation
  - inadequate or missing information or quality control
  - inadequate analysis tools
  - inadequate local storage for data analysis
- long answer: workshop in early june 2004
  - co-chair with IRTF's IMRG chair to maximize community input
  - introduce community to and solicit feedback on architecture and user interface
    - get architecture to fit data, not vice-versa
    - discuss typical user modes for researchers, engineers
  - discuss logistical issues
    - supporting processing tools
    - anonymization techniques
    - security of database
  - future workshop 'reverse engineering the Internet' theme (--neil spring's paper)
  - relationship to and support for distributed observatory