15-744: Computer Networking

L-19 Measurement

Motivation

- Answers many questions
  - How does the Internet really operate?
  - Is it working efficiently?
  - How will trends affect its operation?
  - How should future protocols be designed?
- Aren't simulation and analysis enough?
  - We really don't know what to simulate or analyze
    - Need to understand how Internet is being used!
  - Too difficult to analyze or simulate parts we do understand

Internet Measurement

- Process of collecting data that measure certain phenomena about the network
  - Should be a science
  - Today: closer to an art form
- Key goal: Reproducibility
- "Bread and butter" of networking research
  - Deceptively complex
  - Probably one of the most difficult things to do correctly

Measurement Methodologies

- Active tests – probe the network and see how it responds
  - Must be careful to ensure that your probes only measure desired information (and without bias)
  - Labovitz routing behavior – add and withdraw routes and see how BGP behaves
  - Paxson packet dynamics – perform transfers and record behavior
  - Bolot delay & loss – record behavior of UDP probes
- Passive tests – measure existing behavior
  - Must be careful not to perturb network
  - Labovitz BGP anomalies – record all BGP exchanges
  - Paxson routing behavior – perform traceroute between hosts
  - Leland self-similarity – record Ethernet traffic

Types of Data

Active

- traceroute
- ping
- UDP probes
- TCP probes
- Application-level “probes”
  - Web downloads
  - DNS queries

Passive

- Packet traces
  - Complete
  - Headers only
  - Specific protocols
- Flow records
- Specific data
  - Syslogs …
  - HTTP server traces
  - DHCP logs
  - Wireless association logs
  - DNSBL lookups
  - …
- Routing data
  - BGP updates / tables, ISIS, etc.

5

Overview

- Active measurement
- Passive measurement
- Strategies
- Some interesting observations

6

Active Measurement

- Common tools:
  - ping
  - traceroute
  - scriptroute
  - Pathchar/pathneck/… BW probing tools

7

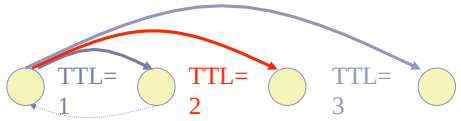
Sample Question: Topology

- What is the topology of the network?
  - At the IP router layer
  - Without “inside” knowledge or official network maps
  - Without SNMP or other privileged access
- Why do we care?
  - Often need topologies for simulation and evaluation
  - Intrinsic interest in how the Internet behaves
    - “But we built it! We should understand it”
    - Emergent behavior; organic growth

8

How Traceroute Works

- Send packets with increasing TTL values



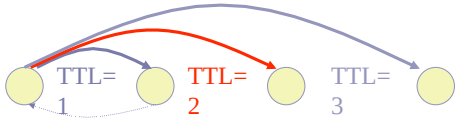
ICMP
"time
exceeded

- Nodes along IP layer path decrement TTL
- When TTL=0, nodes return "time exceeded" message

9

Problems with Traceroute

- Can't unambiguously identify one-way outages
  - Failure to reach host : failure of reverse path?
- ICMP messages may be filtered or rate-limited
- IP address of "time exceeded" packet may be the outgoing interface of the return packet



10

Famous Traceroute Pitfall

- Question: What ASes does traffic traverse?
- Strawman approach
  - Run traceroute to destination
  - Collect IP addresses
  - Use "whois" to map IP addresses to AS numbers
- Thought Questions
  - What IP address is used to send "time exceeded" messages from routers?
  - How are interfaces numbered?
  - How accurate is whois data?

11

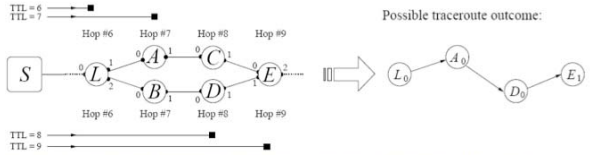
More Caveats: Topology Measurement

- Routers have multiple interfaces
- Measured topology is a function of vantage points
- Example: Node degree
  - Must "alias" all interfaces to a single node
  - Is topology a function of vantage point?
    - Each vantage point forms a tree

12

Less Famous Traceroute Pitfall

- Host sends out a sequence of packets
  - Each has a different destination port
  - Load balancers send probes along different paths
    - Equal cost multi-path
    - Per flow load balancing
- Why not just use same port numbers?



Soule *et al.*, "Avoiding Traceroute Anomalies with Paris Traceroute", *IMC 2006*

Designing for Measurement

- What mechanisms should routers incorporate to make traceroutes more useful?
  - Source IP address to "loopback" interface
  - AS number in time-exceeded message
  - ??
- More general question: How should the network support measurement (and management)?

Overview

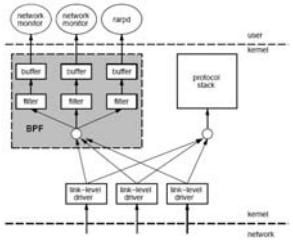
- Active measurement
- Passive measurement
- Strategies
- Some interesting observations

Two Main Approaches

- Packet-level Monitoring
  - Keep packet-level statistics
  - Examine (and potentially, log) variety of packet-level statistics. Essentially, anything in the packet.
  - Timing
- Flow-level Monitoring
  - Monitor packet-by-packet (though sometimes sampled)
  - Keep aggregate statistics on a flow

Packet Capture: tcpdump/bpf

- Put interface in promiscuous mode
- Use bpf to extract packets of interest
- Packets may be dropped by filter
  - Failure of tcpdump to keep up with filter
  - Failure of filter to keep up with dump speeds
- **Question:** How to recover lost information from packet drops?



Traffic Flow Statistics

- *Flow monitoring* (*e.g.*, Cisco Netflow)
  - Statistics about groups of related packets (*e.g.,* same IP/TCP headers and close in time)
  - Recording header information, counts, and time
- More detail than SNMP, less overhead than packet capture
  - Typically implemented directly on line card

What is a flow?

- Source IP address
- Destination IP address
- Source port
- Destination port
- Layer 3 protocol type
- TOS byte (DSCP)
- Input logical interface (ifIndex)

Flow Record Contents

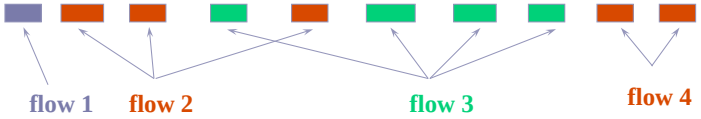
Basic information about the flow…

- Source and Destination, IP address and port
- Packet and byte counts
- Start and end times
- ToS, TCP flags

…plus, information related to routing

- Next-hop IP address
- Source and destination AS
- Source and destination prefix

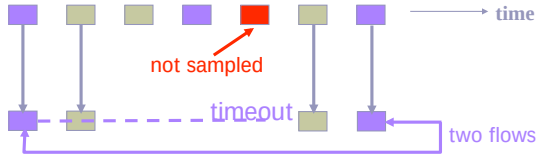
Aggregating Packets into Flows



- **Criteria 1:** Set of packets that "belong together"
  - Source/destination IP addresses and port numbers
  - Same protocol, ToS bits, …
  - Same input/output interfaces at a router (if known)
- **Criteria 2:** Packets that are "close" together in time
  - Maximum inter-packet spacing (e.g., 15 sec, 30 sec)
  - **Example:** flows 2 and 4 are different flows due to time

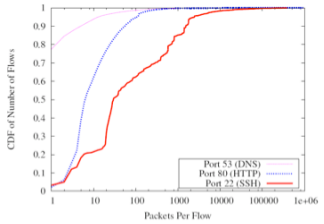
Packet Sampling

- Packet sampling before flow creation (Sampled Netflow)
  - 1-out-of-m sampling of individual packets (e.g., m=100)
  - Create of flow records over the sampled packets
- Reducing overhead
  - Avoid per-packet overhead on (m-1)/m packets
  - Avoid creating records for a large number of small flows
- Increasing overhead (in some cases)
  - May split some long transfers into multiple flow records
  - … due to larger time gaps between successive packets



Problems with Packet Sampling

- Determining size of original flows is tricky
  - For a flow originally of size *n*, the size of the *sampled* flow follows a binomial distribution
  - Extrapolation can result in big errors
  - Much research in reducing such errors (upcoming lectures)
- Flow records can be lost
- Small flows may be eradicated entirely



Overview

- Active measurement
- Passive measurement
- Strategies
- Some interesting observations

Strategy: Examine the Zeroth-Order

- Paxson calls this “looking at spikes and outliers”
- More general: Look at the data, not just aggregate statistics
  - Tempting/dangerous to blindly compute aggregates
  - Time series plots are telling (gaps, spikes, etc.)
  - Basics
    - Are the raw trace files empty?
      - Need not be 0-byte files (e.g., BGP update logs have state messages but no updates)
    - Metadata/context: Did weird things happen during collection (machine crash, disk full, etc.)

25

Strategy: Cross-Validation

- Paxson breaks cross validation into two aspects
  - Self-consistency checks (and sanity checks)
  - Independent observations
    - Looking at same phenomenon in multiple ways
- What are some other examples of each of these?

26

Longitudinal measurement hard

- Accurate distributed measurement is tricky!
- Lots of things change:
  - Host names, IPs, software
- Lots of things break
  - hosts (temporary, permanently)
  - clocks
  - links
  - collection scripts
- Paxson's “master script” can help a bit

29

Anonymization

- Similar questions arise here as with accuracy
- Researchers always want full packet captures with payloads
  - …but many questions can be answered without complete information
- Privacy / de-anonymization issues

30

PlanetLab for Network Measurement

- Nodes are largely at academic sites
  - Other alternatives: RON testbed (disadvantage: difficult to run long running measurements)
- Repeatability of network experiments is tricky
  - Proportional sharing
    - Minimum guarantees provided by limiting the number of outstanding shares
  - Work-conserving CPU scheduler means experiment could get more resources if there is less contention

31

Overview

- Active measurement
- Passive measurement
- Strategies
- Some interesting observations

32

Traces Characteristics

- Some available at http://ita.ee.lbl.gov
  - E.g. tcpdump files and HTTP logs
  - Public ones tend to be old (2+ years)
  - Privacy concerns tend to reduce useful content
- Paxson's test data
  - Network Probe Daemon (NPD) – performs transfers & traceroutes, records packet traces
  - Approximately 20-40 sites participated in various NPD based studies
  - The number of "paths" tested by NPD framework scaled with (number of hosts)^2
    - 20-40 hosts = 400-1600 paths!

33

Observations – Routing Pathologies

- Observations from traceroute between NPDs
- Routing loops
  - Types – forwarding loops, control information loop (count-to-infinity) and traceroute loop (can be either forwarding loop or route change)
  - Routing protocols should prevent loops from persisting
  - Fall into short-term (< 3hrs) and long-term (> 12 hrs) duration
  - Some loops spanned multiple BGP hops! → seem to be a result of static routes
- Erroneous routing – Rare but saw a US-UK route that went through Isreal → can't really trust where packets may go!

34

Observations – Routing Pathologies

- Route change between traceroutes
  - Associated outages have bimodal duration distribution
    - Perhaps due to the difference in addition/removal of link in routing protocols
- Temporary outages
  - Traceroute probes (1-2%) experienced > 30sec outages
  - Outage likelihood strongly correlated with time of day/ load
- Most pathologies seem to be getting worse over time

35

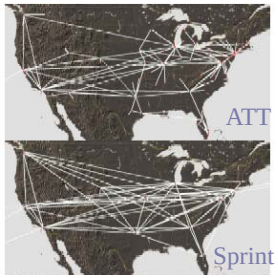
Observations – Routing Stability

- Prevalence – how likely are you to encounter a given route
  - In general, paths have a single primary route
  - For 50% of paths, single route was present 82% of the time
- Persistence – how long does a given route last
  - Hard to measure – what if route changes and changes back between samples?
  - Look at 3 different time scales
    - Seconds/minutes→ load-balancing flutter & tightly coupled routers
    - 10's of Minutes → infrequently observed
    - Hours → 2/3 of all routes, long lived routes typically lasted several days

36

ISP Topologies

- Rocketfuel [SIGCOMM02]
  - Maps ISP topologies of specific ISPs
    - BGP → prefixes served
    - Traceroute servers → trace to prefixes for path
    - DNS → identify properties of routers
    - Location, ownership, functionality



- However…
  - Some complaints of inaccuracy – why?
  - [IMC03] paper on path diversity

http://www.cs.washington.edu/research/networking/rocketfuel/

37

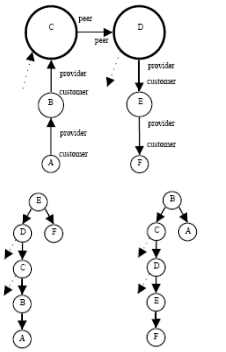
Network Topology

- Faloutsos^3 [SIGCOMM99] on Internet topology
  - Observed many "power laws" in the Internet structure
    - Router level connections, AS-level connections, neighborhood sizes
  - Power law observation refuted later, Lakhina [INFOCOM00]
- Inspired many degree-based topology generators
  - Compared properties of generated graphs with those of measured graphs to validate generator
  - What is wrong with these topologies? Li et al [SIGCOMM04]
    - Many graphs with similar distribution have different properties
    - Random graph generation models don't have network-intrinsic meaning
    - Should look at fundamental trade-offs to understand topology
      - Technology constraints and economic trade-offs
    - Graphs arising out of such generation better explain topology and its properties, but are unlikely to be generted by random processes!

38

Inter-Domain Relationships

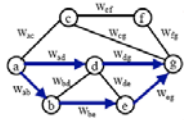
- Gao [TON01] → look at highest degree node
  - "Turning point" or plateau of valley-free path
- Subramanian [Infocom02] → merge views from multiple BGP tables, ranking each node
  - Peering edge (i, j): ranks are equal according to >50% vantage points
  - Customer-provider edge (i, j): rank(i) > rank (j) according to >50% vantage points



http://www.cs.berkeley.edu/~sagarwal/research/BGP-hierarchy/

Policies: Intra- and Inter-Domain

- Mahajan et. al. [IMW02]
  - Approximate link weights on ISPs
  - Actually, relative link weights
    - NOT real weights
  - Use observed paths, solve constraints
    - Only a snap-shot of the weights



1. $W_{ad} + W_{dg} = W_{ab} + W_{be} + W_{eg}$ [ADG=ABEG]
2. $W_{ad} + W_{dg} < W_{ac} + W_{cg}$ [ADG<ACG]
3. $W_{ad} + W_{dg} < W_{ac} + W_{cf} + W_{fg}$ [ADG<ACFG]
4. $W_{ad} + W_{dg} < W_{ab} + W_{bd} + W_{dg}$ [ADG<ABDG]
5. $W_{ad} + W_{dg} < W_{ad} + W_{de} + W_{eg}$ [ADG<ADEG]
6. $W_{ad} + W_{dg} < W_{ab} + W_{bd} + W_{de} + W_{eg}$ [ADG<ABDEG]

- Spring et. al. [SIGCOMM03]
  - Again, use lots of traceroutes
  - Quantify early exit between ISPs, Late exit, Load balancing

http://www.cs.washington.edu/research/networking/rocketfuel/

Routing Faults, Errors

- BGP misconfiguration – Mahajan et. al. [SIGCOMM02]
  - How prevalent?
    Upto 1% of BGP table size!
  - Impact on connectivity?
    Not much, but could increase router processing load
  - Causes?
    Router reboot, old configuration, redistribution, hijacks due to typos
- Routing failures – Feamster et. al [SIGMETRICS03]
  - How often do they occur?
    Often
  - Where do failures occurs?
    Everywhere, but most at edge networks and also within ASes
  - How long do they last?
    70% are < 5 minutes, 90% < 15 minutes
  - Do they correlate with BGP instability?
    Failures occur around instability → can use BGP to predict

http://nms.lcs.mit.edu/ron/

Observations – Re-ordering

- 12-36% of transfers had re-ordering
- 1-2% of packets were re-ordered
- Very much dependent on path
  - Some sites had large amount of re-ordering
  - Forward and reverse path may have different amounts
- Impact → ordering used to detect loss
  - TCP uses re-order of 3 packets as heuristic
  - Decrease in threshold would cause many "bad" rexmits
    - But would increase rexmit opportunities by 65-70%
  - A combination of delay and lower threshold would be satisfactory though → maybe Vegas would work well!

Observations – Packet Oddities

- Replication
  - Internet does not provide “at most once” delivery
  - Replication occurs rarely
  - Possible causes → link-layer rexmits, misconfigured bridges
- Corruption
  - Checksums on packets are typically weak
    - 16-bit in TCP/UDP → miss 1/64K errors
  - Approx. 1/5000 packets get corrupted
  - 1/3million packets are probably accepted with errors!

Observations – Bottleneck Bandwidth

- Typical technique, packet pair, has several weaknesses
  - Out-of-order delivery → pair likely used different paths
  - Clock resolution → 10msec clock and 512 byte packets limit estimate to 51.2 KBps
  - Changes in BW
  - Multi-channel links → packets are not queued behind each other
- Solution – many new sophisticated BW measurement tools
  - Unclear how well they really work ☹

Observations – Loss Rates

- Ack losses vs. data losses
  - TCP adapts data transmission to avoid loss
  - No similar effect for acks → Ack losses reflect Internet loss rates more accurately (however, not a major factor in measurements)
- 52% of transfers had no loss (quiescent periods)
- 2.7% loss rate in 12/94 and 5.2% in 11/95
  - Loss rate for “busy” periods = 5.6 & 8.7%
  - Has since gone down dramatically…
- Losses tend to be very bursty
  - Unconditional loss prob = 2 - 3%
  - Conditional loss prob = 20 - 50%
  - Duration of “outages” vary across many orders of magnitude (pareto distributed)

Observations – TCP Behavior

- Recorded every packet sent to Web server for 1996 Olympics
  - Can re-create outgoing data based on TCP behavior → must use some heuristics to identify timeouts, etc.
- How is TCP used clients and how does TCP recover from losses
  - Lots of small transfers done in parallel

Observations – TCP Behavior

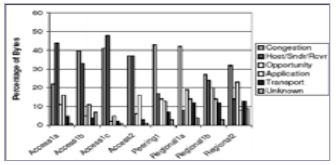
Trace Statistic	Value	%Age
Total connections	1,650,103	100
With packet reordering	97,036	6
With rcvr window bottleneck	233,906	14
Total packets	7,821,638	100
During slow start	6,662,050	85
Slow start packets lost	354,566	6
During congestion avoidance	1,159,588	15
Congestion avoidance loss	82,181	7
Total retransmissions	857,142	100
Fast retransmissions	375,306	44
Slow start following timeout	59,811	7
Coarse timeouts	422,025	49
Avoidable with SACK	18,713	4

47

Flow Performance

- E2E performance – Zhang et al. [SIGCOMM02]
  - Packet-level traces collected at various ISP links and ISP summary flow stats
  - Flow *rate?* → *T-RAT*
    Not as skewed as flow size;
    But highly correlated with size
  - Reason for limited flow rate?
    Network congestion and receiver limitations

http://www.research.att.com/projects/T-RAT/



Inter-ISP links

	%bottlenecks	%all links
Tier 4 – 4, 3, 2, 1	14%	1%
Tier 3 – 3, 2, 1	17%	3%
Tier 2 – 2, 1	12%	4%
Tier 1 – 1	8%	6%

Intra-ISP links

Tier 4	3%	1%
Tier 3	9%	8%
Tier 2	12%	13%
Tier 1	25%	63%

- Classic Paxson97 paper
  - Traces of many TCP transfers
  - Observed packet reordering and corruption
  - Measured packet loss rates and showed loss events occur in bursts
- Wide-area performance – Akella et. al. [IMC03]

http://www-2.cs.cmu.edu/~aditya/bfind_release

48

Application Traffic Analysis

- P2P systems – Saroiu et. al. [MMCN02]
  - Bandwidths distribution? Mostly DSL; better upstream bw → client-like
  - Availability? Median ~ 60min
  - Peers often lie about bandwidth and avoid sharing

http://sprobe.cs.washington.edu/

- P2P traffic – Saroiu et. al. [OSDI02]
  - P2P traffic dominates in bw consumed
  - Kind of traffic carried: Kazaa → mostly video (bytes); web → text +images
  - File size distribution: P2P and HTTP (new)
    - P2P objects are 3X bigger
  - "Fetch-only-once" → popularity significantly different than Zipf

49

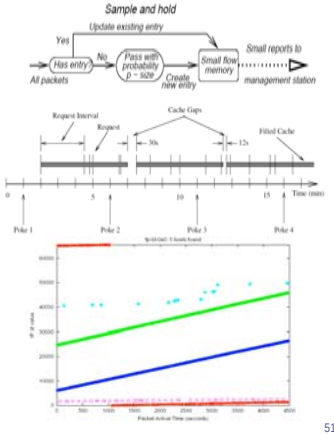
DNS Analysis

- Very interesting area, but little work ☹
- Danzig [SIGCOMM92] → analysis of DNS traffic
  - How config errors contribute to inflation
- Follow-up I: Jung et. al. [IMW01]
  - Failure-analysis and impact on latency
  - Cache hit rate (at MIT): 70%-80% → session-level = 0%!
  - Impact of TTLs: Low A-record TTLs are not bad for hit rates
  - Cache sharing: ~20 clients good enough for hit rate
- Follow-up II: Pang et. al. [IMC 04]
  - DNS infrastructure characteristics
  - Many authoritative and local name servers are both highly available
  - A few name servers get most of the load
  - Usage and availability correlated

50

Algorithms, Hacks

- Counting and Sampling
  - Estan et. al. [SIGCOMM02] [IMC03]
    - Sample and hold: counting "heavy hitters"
    - Multi-resolution bit-maps: counting all flows
- Cool hacks are always hot
  - Wills et. al. [IMC03]
    - Popularity of web-sites by probing LDNS caches
  - Bellovin [IMW02]
    - Count NAT-ed hosts by looking at IPid sequences



Security, Worms

- Code-Red case study from CAIDA [IMW02]
  - Origins of infected host (country, region)? US, Korea…
  - Rate of infection? Up to 2000 hosts infected per minute!
  - How quickly were patches developed, employed?
    - Patches developed only after attack
    - Patches employed only after Code-Red v2 arrived!
- Intrusion detection – Barford et. al. [SIGMETRICS03]
  - Look at >1600 firewalls logs for intrusions and extrapolate
  - Estimates about 25B intrusion attempts per day
  - Very few sources trigger a lot of attempts
    - Function in cliques
  - Intrusion attempts look normal from any single vantage point
    - Need global coordinated intrusion detection

http://wail.cs.wisc.edu/anomaly.html