INTERNET TRAFFIC MEASUREMENT (PART I)

Gaia Maselli

maselli@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

Overview

- Basic concepts
- Characterization of traffic properties that are important to measure
- Challenges in traffic measuring
- Tools available
- The most important results regarding properties of Internet traffic

Motivation

- Traffic measurement and modeling is important to a wide range of activities
- *Performance analysis* requires accurate traffic measurements in order to construct models useful for answering question related to throughput, packet loss, and packet delay induced by network elements (links and routers)
  - Performance analysis problems are generally concerned with timescale from microseconds to ten of minutes
- *Network engineering* is concerned with network configuration, capacity planning, demand forecasting, traffic engineering
  - The questions that arise in traffic engineering are usually concerned with timescales from minutes to years

Basic concepts

Stochastic processes

- One often encounters a situation in which measurements are presented in some order, for example over time
- To use the tools of probability in this setting we need to define a *sequence of random variables*, called *stochastic process*
- A stochastic process is a collection of random variables indexed on a set (usually the index denotes time)
- A continuous-time stochastic process

$$\{X_t, t \geq 0\}$$

- A discrete-time stochastic process

$$\{X_n, n = 1, 2, \ldots\}$$

- In computer networks we will be concerned with discrete-time stochastic processes

Stochastic processes relevant to Internet measurements

- Often one is concerned with events occurring at specific points in time, which we can generically call arrivals
- An *arrival process* is a stochastic process in which successive random variables correspond to instants of arrivals: $\{A_n, n=0,1,\ldots\}$
- The arrival process has the property that is non decreasing, and so it is not stationary
- Thus it it often more convenient to work with interarrivals, which may or may not be stationary
- Interarrivals can be modeled as an interarrival process, $\{I_n, n=1,2,\ldots\}$ where $I_n=A_n-A_{n-1}$
- A summarization of an interarrival process consists of a characterization of the distribution

Time series of counts

- Another useful model for a sequence of events
- The *time series of counts* is the most common form in which network traffic is reported
- One establishes fixed-size time intervals and counts how many arrivals occur in each time interval
- For a fixed time interval T, $\{C_n, n=1,2,\ldots\}$ where
  $C_n = \#\{A_m \mid nT < A_m \leq (n+1)T\}$
- The particular value of T chosen is called the *timescale* of the time series of counts
- The time series of counts contain less information than the arrival process view because precise arrival times are lost
- It is not possible to reconstruct the sequence of arrivals from the counts

Characterization of traffic properties that are important to measure

Traffic properties

The traffic properties that can be important to measure

- The basics: packets and bytes
- Traffic high-level structure
- Flows
- Traffic purpose

The basics: Packets and bytes

- Traffic can be viewed at the IP level as a *set of packets* (collection of packets passing through routers and over links)
- At some location (one end of the link) it is possible to capture or observe the packets
- How can we summarize the traffic observed?

1. Considering those points in time when a packet arrives at our observation point:
   a. arrival process $\{A_n, n=1,2,\ldots\}$
   b. Inter arrival process $\{I_n, n=1,2,\ldots\}$ where $I_n=A_n-A_{n-1}$
2. Time series of counts $\{C_n, n=1,2,\ldots\}$ where

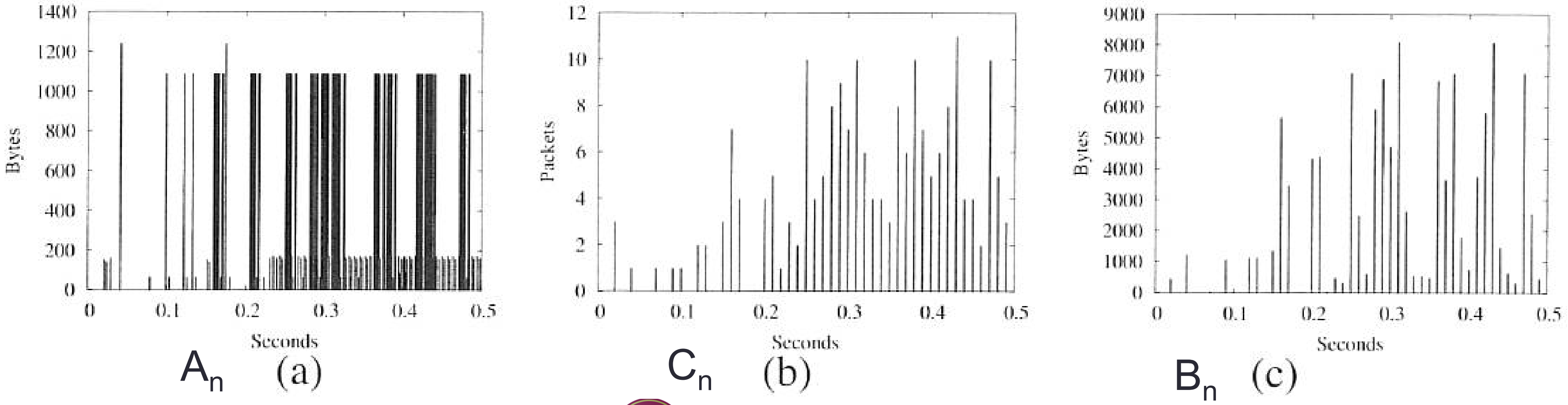
$$C_n = \#\{A_m \mid nT < A_m \leq (n+1)T\}$$

The basics: Packets and bytes (Cont)

3. Considering the bytes contained in packets (packets have varying size)
   a. Time series of counts (counting the number of bytes contained in the packets arriving in each interval)

      $\{B_n, n=1,2,\ldots\}$ where $B_n=\Sigma_{nT<Am\leq(n+1)T}\text{size}(A_m)$

      $\text{size}(A_m)$ is the size of the packet arriving at time A_m and includes payload + headers at all levels



A_n (a)

C_n (b)

B_n (c)

The basics: Packets and bytes (Cont)

- The time series of byte counts is the most commonly used measure of the workload represented by traffic, since it captures the amount of bandwidth consumed as packets pass through routers and over links
- The time series of packet counts is useful for understanding the workload generated by traffic on a per-packet basis (address lookups performed by routers)

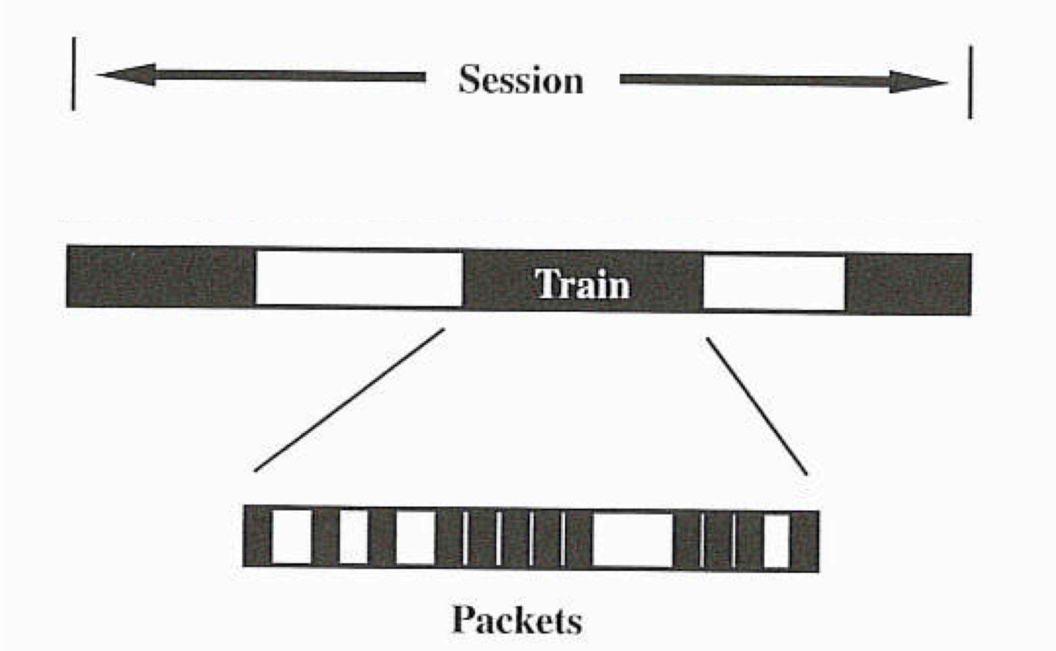
Traffic high-level structure

- Structure is imposed by transport protocols and by the ultimate generators of traffic (applications)
- Understanding this structure helps in interpreting the observed properties
- By observing traffic structure we can learn about the higher-level protocols and applications
- We can describe structure in traffic as a result of a collection of ON/OFF processes
- An ON/OFF process is one that alternates between a “on” state in which it generates network activity, and an “off” state in which it is silent

Three levels of structure in traffic

There are three principal levels of ON/OFF activity in network traffic:

1. Packets - network state at the link level alternates between packet transmission and silence
2. Trains – subset of packets in a trace corresponding to a single source and destination
3. Sessions – set of packet trains corresponding to a single execution of an application



Flows

- In many cases it is not important to identify application-level data items, but to collect all the packets associated with the exchange of data between two endpoints into a single entity (billing, modeling, traffic summarization)
- Flow means a set of packets passing an observation point during a time interval, with all packets having a set of common properties
- The set of common properties may be based on a packet's header filed contents, characteristics of the packet itself (such as layer-2 labels), or how the packet is processed (its next hop IP address or output interface)

Flows (Cont)

- *IP flows*: a set of packets distinguished by their source and destination addresses, or any other function of their IP or transport header fields (port, protocol, address, combination of these)

Example: packet train structure exists within IP flows between individual endsystems as defined by their IP addresses

- *Network-defined flows*: defined with respect to a particular network's workload

Example: the set of packets entering a network at a given point and exiting at another point is called an origin-destination or ingress-egress flow

Traffic purpose

- Traffic purpose is important for traffic measurement
- *Control traffic* is sent over the same media as regular data packets
  - packets implementing routing protocols (most commonly BGP and OSPF)
  - Measurement packets (SNMP)
  - General control packets (ICMP)
- Measuring control traffic is of interest in order to reproduce it in simulation or because it reflects the performance of routing systems
- *Malicious traffic* due to misuses or abuses

Challenges in traffic measuring

Challenges

- Practical issues in obtaining traffic measurements
  - Inability to observe
  - Inability to manage
  - Inability to share
- Statistical difficulties
  - Long tails and high variability
  - Stationarity and stability

Practical issues: observability

- The Internet architecture has certain distinct characteristics that tend to interfere with easy measurement of network traffic
- Core simplicity
  - Per-flow state is not maintained by routers
  - Packet capture is not available and must be implemented at endsystems or by additional hardware added to the network
- Distributed internetworking
  - There is not a single backbone network that provides a convenient measurement point for the majority of Internet traffic (large autonomous systems interconnect at multiple points, commercial and non commercial backbone networks)
  - Any measurement point presents a local view of traffic and network properties that may not be representative of other measurement points

Practical issues: data volume

- The most useful form of traffic monitoring is full packet capture (tcpdump, wireshark)
- But, as link speeds increase, full packet capture becomes increasingly challenging
  - Packet capture on an OC-48 link with 50% utilization yields 155 megabytes per second of data to be stored
  - For an OC-192 link, 625 megabytes per second!
  - Such data quantities are challenging to process, to store, and to manage
- Full packet capture on high-speed links is appropriate for short timescale (a minute or less)
- With sampling or summarization (next lecture) it can be feasible to scale up to days

Practical issues: data sharing

- Data sensitivity is another important aspect
- Full packet capture records the activity of network users
- Such traces can be used to extract information such as Web sites visited, passwords, and contents of email messages exchanged
- Even with less details, traces can provide information about the configuration and functioning of the network being monitored (information competition-sensitive for network service providers)
- The goals of protecting the privacy of network users and protecting the competitive secrets of network providers are in tension with network measurement

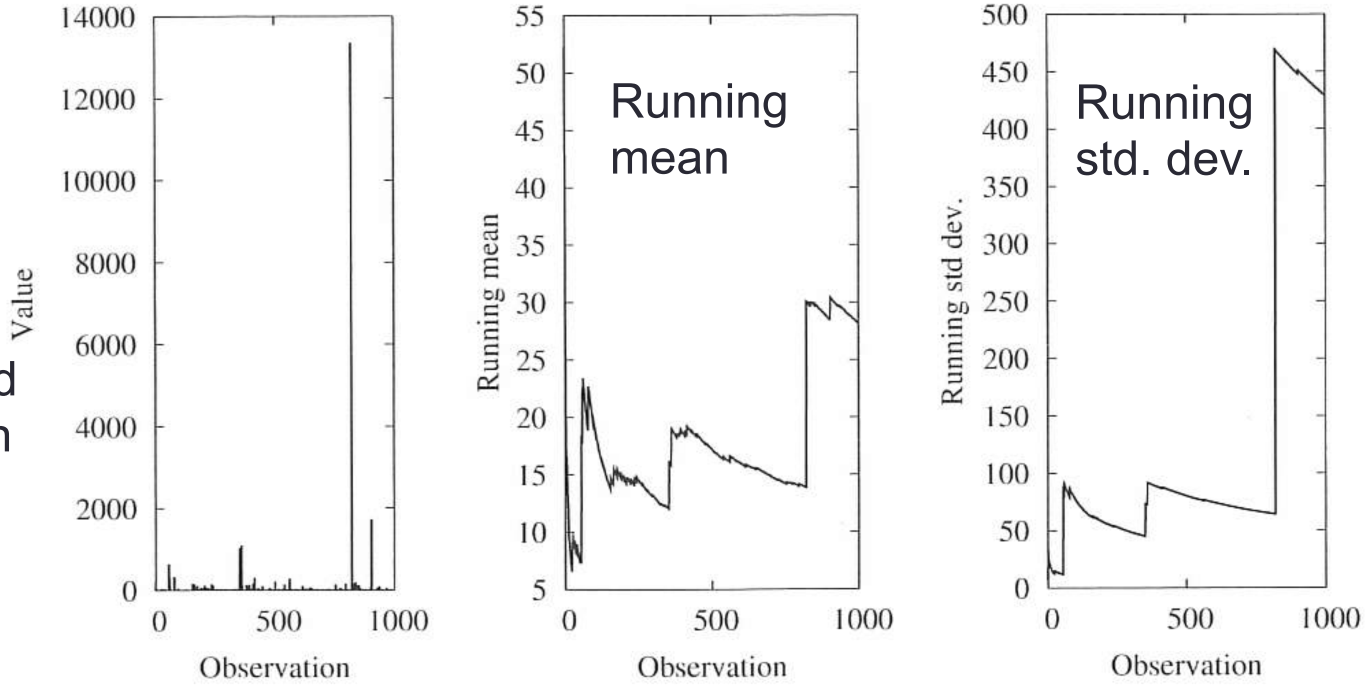
Statistical difficulties

Statistical difficulties: long tails

- A number of issues arise when statistically characterizing the results
- A particularly important part of a distribution is its upper **tail – the portion of the distribution that describes the probability of large values**
- Large values can dominate system performance, so a precise understanding of the probability of large values is often a prime concern
- For a random variable X having distribution F(x), we are concerned with the shape of 1-F(x) = P[X>x] for large x
- *Short-tailed* distributions have tails that decrease exponentially or faster (e.g., Exponential, Normal, Gamma)
- *Long-tailed* distributions have tails that decline more slowly than any exponential distribution (e.g., Pareto)

Statistical difficulties: long tails (Cont)

- A number of properties in Internet traffic show high variability, causing
- Instability of metrics: high variability data exhibits many small observations mixed with a few large observations which produce instable mean and variance



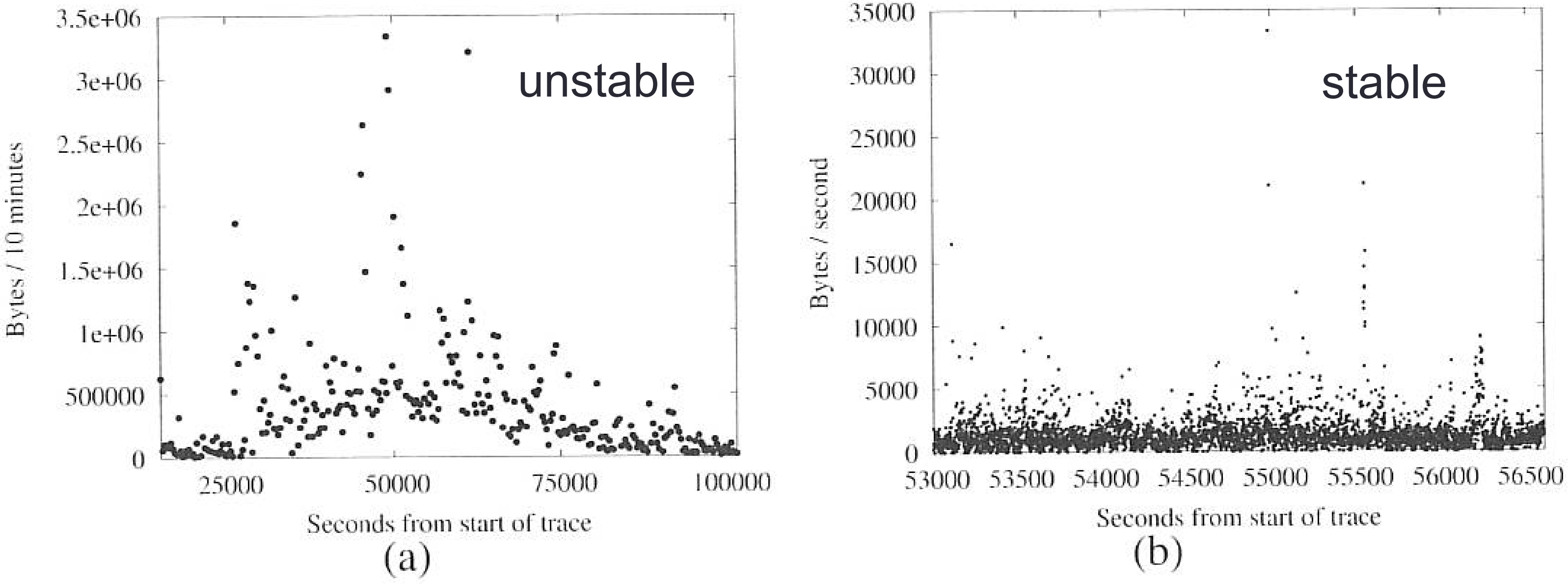
Long-tailed distribution

Statistical difficulties: stationarity and stability

- A central question in studying network traffic is the nature of traffic stability
- Stability refers to the consistency of traffic properties over time
- *Data stability*: data is stable if its empirical statistics do not seem to be changing over time
- To verify stability we need to define some objective metrics
- Approach:
  - Break the dataset into windows. For example 1000 observations can be divided up into 10 windows consisting of the first 100 observations, the second 100 observations, and so on
  - Compute the empirical statistics for each window (e.g., mean and variance)
  - If the resulting set of statistics do not seem to show any consistent or predictable variation (such as trend or a sudden shift) then dataset can be consider stable

Statistical difficulties: stationarity and stability (Cont)

- When measurement properties are stable, it may be appropriate to describe the measurements using *a stationary model*
- The question of stability (and the appropriateness of using a stationary model) is closely tied with *timescale*
- Many traffic properties may be thought of as unstable at some time scale and stable at other timescales
- Example: traffic volume (bytes or packets per unit of time) at two timescales: 24 hours and 1 hour



Statistical difficulties: stationarity and stability (Cont)

- Different goals for traffic measurement and modeling are associated with different assumptions about traffic stability
- Performance analysis question are generally concerned with short timescale (a hour or less) and usually assume that traffic is stable over these timescales
- The properties of congestion events differ depending on the timescale of measurement
  - Coarse timescale traffic measurements (e.g., 5 minute SNMP counts) will tend to obscure traffic burst which are important for performance analysis