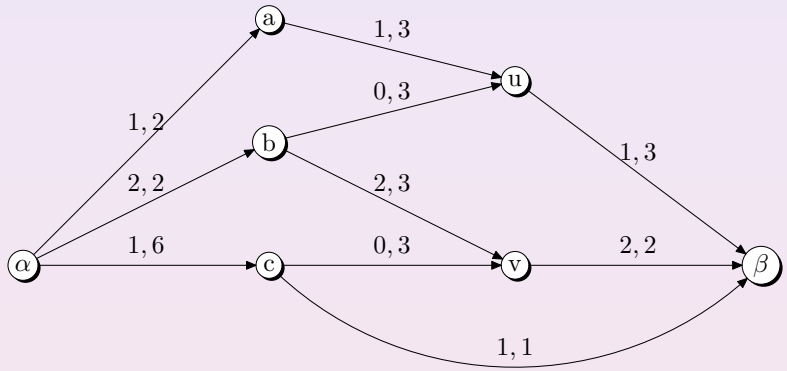
# Modelling data networks – research summary and modelling tools




Richard G. Clegg (richard@richardclegg.org)— December 2011
Available online at `http://www.richardclegg.org/lectures` accompanying printed notes provide full bibliography.
(Prepared using LaTeX and beamer.)

# Research topics and modelling demonstration

- This final lecture describes some important topics in network research.
- These are a personal view of some interesting research topics which remain to answer.
- The mathematical techniques from previous lectures today are all important in many of these research areas.
- Finally a demonstration is given of a well-known modelling tool ns-2.

# Areas of modelling interest(1)

Now let us focus on several specific areas of interest to modellers.

## Topology models

How are the nodes in the internet connected to each other? How are things connected? Is the internet "big" or "small"?

## User/flow arrival modelling

How does traffic arrive on the internet? We can see arrivals as a stochastic process. How long do connections last?

## Application level protocols

What traffic do applications place on the internet? For example: peer-to-peer networks use an overlay (graph theory again?) A web page might make connections to many different places.

# Areas of modelling interest(2)

## Traffic statistics

What does the traffic along a link look like in statistical terms? Again see internet traffic as a stochastic process (queuing theory). How does TCP congestion control alter this?

## Transport/network protocols

How do TCP/IP protocols affect the traffic? See internet traffic as a feedback process (control theory). How do these protocols interact with the rest of the network?

Other things to model:

- Reliability modelling — what happens when links or nodes fail?
- Overlay networks — P2P increasingly important.

# Internet topology

- Two levels of topology are usually considered "router level" and "autonomous system" (AS) level.
- Router level topology is still the least well-known — often ISPs take trouble to protect this information for security reasons.
- Topology metrics — these quantities are all rigorously defined and can be found in the literature:
  1. Graph diameter (longest possible "shortest path" between nodes).
  2. Node degree distribution (what proportion of nodes have $k$ neighbours).
  3. Assortivity/disassortivity (do well-connected nodes connect with each other?) – sometimes called "rich club".
  4. Clustering (triangle count) – are the neighbours of a node also neighbours of each other.
  5. Clique size – largest group where everyone is everyone's neighbour (a clique in graph theory).

# AS level topology

## Power law networks

The node-degree distribution in AS networks is particularly well-studied. Let $P(k)$ be the proportion of nodes with degree $k$ (having $k$ neigbours). To a good approximation

$$P(k) \sim k^{-\alpha},$$

where $\alpha$ is a constant.

- Power law topology of the AS graph shown by [Faloutsos x3].
- This graph has some interesting properties — some extremely highly connected nodes, what happens if they fail?
- Same type of graph as:
  1. Links on websites, wikipedia and many other similar online systems.
  2. Academic citations in papers.
  3. Human sexual contacts.

# Mathematics to generate AS topology

## Albert–Barabasi [Barabasi 99] "Preferential attachment" model

Constructive Start with a small "core" network. When a new node arrives, attach it to an old node with the following probability

$$\mathbb{P}\left[\text{Attaching to node } i\right] = \frac{d(i)}{\sum_{j \in \text{all nodes}} d(j)},$$

where $d(i)$ is the degree of node $i$.

- This model "grows" a network with a powerlaw.
- Many similar models have been created which are more general.
- Current best model may be [Zhou 2004] Positive Feedback Preference which adds a small "faster than exactly proportional" term.

# User/flow arrival modelling

- As a first approximation the arrival of users can be modelled as a Poisson process.
- You might want to consider periodic effects:
  1. Daily – with people's sleep cycles.
  2. Weekly – weekends different.
  3. Yearly – year-on-year growth in traffic.
- Perhaps simpler just to simulate some peak hour and some estimate of growth?

# Application level protocols

- If you are modelling a specific application there will be details associated with this.
- Common applications (www, ftp, p2p) will have existing research — read what is done before setting out on your own.
- If no studies are done what could you compare your application to?
- Could your application be viewed as:
  1. A series of ftp-like transfers of data.
  2. UDP bursts at a given rate for given periods of time
  3. A p2p application which might use existing p2p research methods.
- An important thing to simulate is the length of transfers and for many applications this is heavy-tailed.

# What is a Heavy-Tailed distribution?

## Heavy-Tailed distribution

A variable $X$ has a heavy-tailed distribution if

$$\mathbb{P}[X > x] \sim x^{-\beta},$$

where $\beta \in (0, 2)$ and $\sim$ again means asymptotically proportional to as $x \to \infty$.

- Obviously an example of a power law.
- A distribution where *extreme values* are still quite common.
- Examples: Heights of trees, frequency of words, populations of towns.
- Best known example, Pareto distribution $\mathbb{P}[X > x] = (x/x_m)^{-\beta}$ where $x_m > 0$ is the smallest value $X$ can have.

# Heavy tails and the internet

- The following internet distributions have heavy tails:
  1. Files on any particular computer.
  2. Files transferred via ftp.
  3. Bytes transferred by single TCP connections.
  4. Files downloaded by the WWW.
- This is more than just a statistical curiousity.
- Consider what this distribution would do to queuing performance (no longer Poisson).
- Non mathematicians are starting to take an interest in heavy tails (reference to “the long tail”.

# Long-Range Dependence (LRD) and the Internet

- In 1993 LRD was found in a time series of bytes/unit time measured on an Ethernet LAN [Leland et al '93].
- This finding has been repeated a number of times by a large number of authors (however recent evidence suggests this may not happen in the core).
- A higher Hurst parameter often increases delays in a network. Packet loss also suffers.
- If buffer provisioning is done using the assumption of Poisson traffic then the network will probably be underspecifed.
- The Hurst parameter is "a dominant characteristic for a number of packet traffic engineering problems".

# Long-Range Dependence (LRD)

Let $\{X_1, X_2, X_3, \dots\}$ be a weakly stationary time series.

## The Autocorrelation Function (ACF)

$$\rho(k) = \frac{\mathrm{E}\left[(X_t - \mu)(X_{t+k} - \mu)\right]}{\sigma^2},$$

where $\mu$ is the mean and $\sigma^2$ is the variance.

The ACF measures the correlation between $X_t$ and $X_{t+k}$ and is normalised so $\rho(k) \in [-1, 1]$. Note symmetry $\rho(k) = \rho(-k)$.
A process exhibits LRD if $\sum_{k=0}^{\infty} \rho(k)$ diverges (is not finite).

## Definition of Hurst Parameter

The following functional form for the ACF is often assumed

$$\rho(k) \sim |k|^{-2(1-H)},$$

where $\sim$ means asymptotically proportional to and $H \in (1/2, 1)$ is the Hurst Parameter.

# More about LRD

- Think of LRD as meaning that data from the distant past continue to effect the present.
- LRD was first spotted by a hydrologist (Hurst) looking at the flooding of the Nile river.
- For this reason Mandelbrot called it “the Joseph effect”.
- Stock prices (once normalised) also show LRD.
- LRD can also be seen in the temperature of the earth (once the trend is removed).
- Models include Markov chains, Fractional Brownian Motion (variant on Brownian motion), Chaotic maps and many others.

# Transport and network level protocols

- It might be important if we are considering a packet level model to model specific details of the TCP/IP protocols.
- Usually this will involve simulating the window size (additive increase multiplicative decrease) of the TCP protocol.
- Remember that a detailed simulation to this level will extremely limit the number of nodes which can be simulated.
- A mathematical model will be demonstrated in the next section.
- In addition, the ns-2 model will be shown which is a packet level simulation of TCP/IP.

# Other things to model

- Of course depending on the nature of your modelling, there may well be other aspects of the network to be modelled.
- Some examples might be:
  1. Reliability of nodes and links.
  2. An overlay network.
  3. Possible hostile attacks to the network.
- In all cases, an important starting point is to find out what research already exists in the area.
- Are any real-life data sets available which could inform your modelling? Could you gather such data?

# The ns-2 simulation

- ns-2 is a freely available event-driven simulator which simulates packet-level traffic.
- It is available from `http://www.isi.edu/nsnam/ns/`
- The simulator is written in C++ but uses tcl for simulations.
- The scripts used for the rest of this lecture are available at `http://www.richardclegg.org/lectures`

# Summary of today's lectures

## Stochastic processes

A building block for many mathematical models of the internet. Characterising some part of the network in terms of a stochastic process allows you to understand it better.

## Markov chains

A flexible mathematical tool allowing you to model how a system evolves in time and a necessary prerequisite for queuing theory.

## Simulation modelling

Sometimes simulation is the only way to proceed but study of the existing system is necessary to make the simulation match reality.

# Final thoughts

- Select an appropriate level of modelling — if you need to model the whole internet you cannot do packet level modelling. If you need to model intricate protocol details for packets you cannot model the whole internet.
- Check against real data where possible that your modelling assumptions are justified.
- Is your experiment repeatable? Do you get similar results if you try slightly different starting scenarios?
- Remember sensitivity analysis: What happens if the bandwidth is a little less? What if the demand is a little more?
- Can statistical analysis of your results help?
- Remember that what you model today is out of date in a year and hopelessly obsolete in ten years.