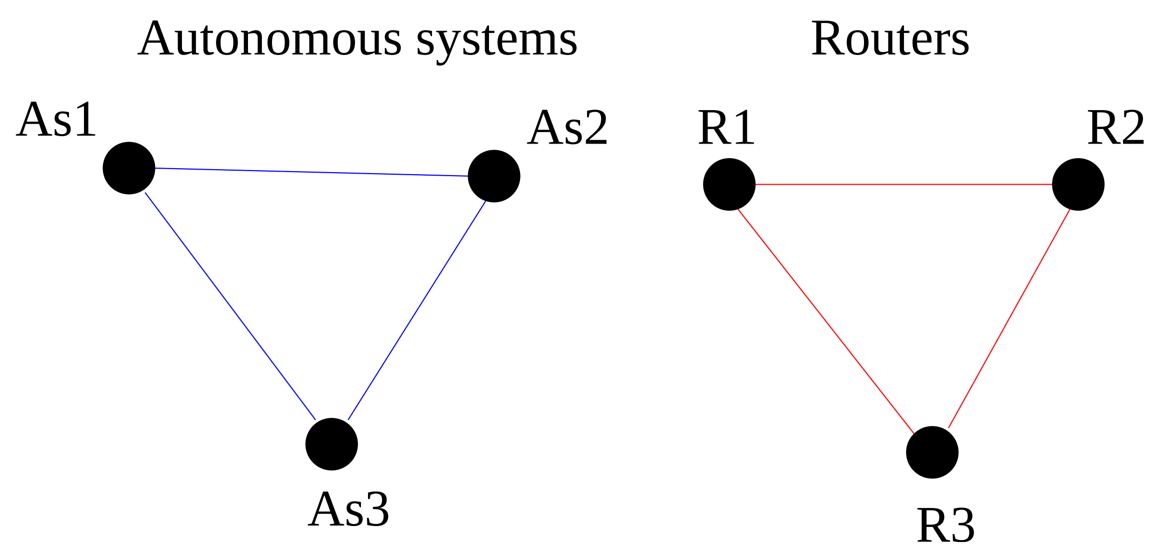
# *Empirical variance and short cycles in a power law data network model*

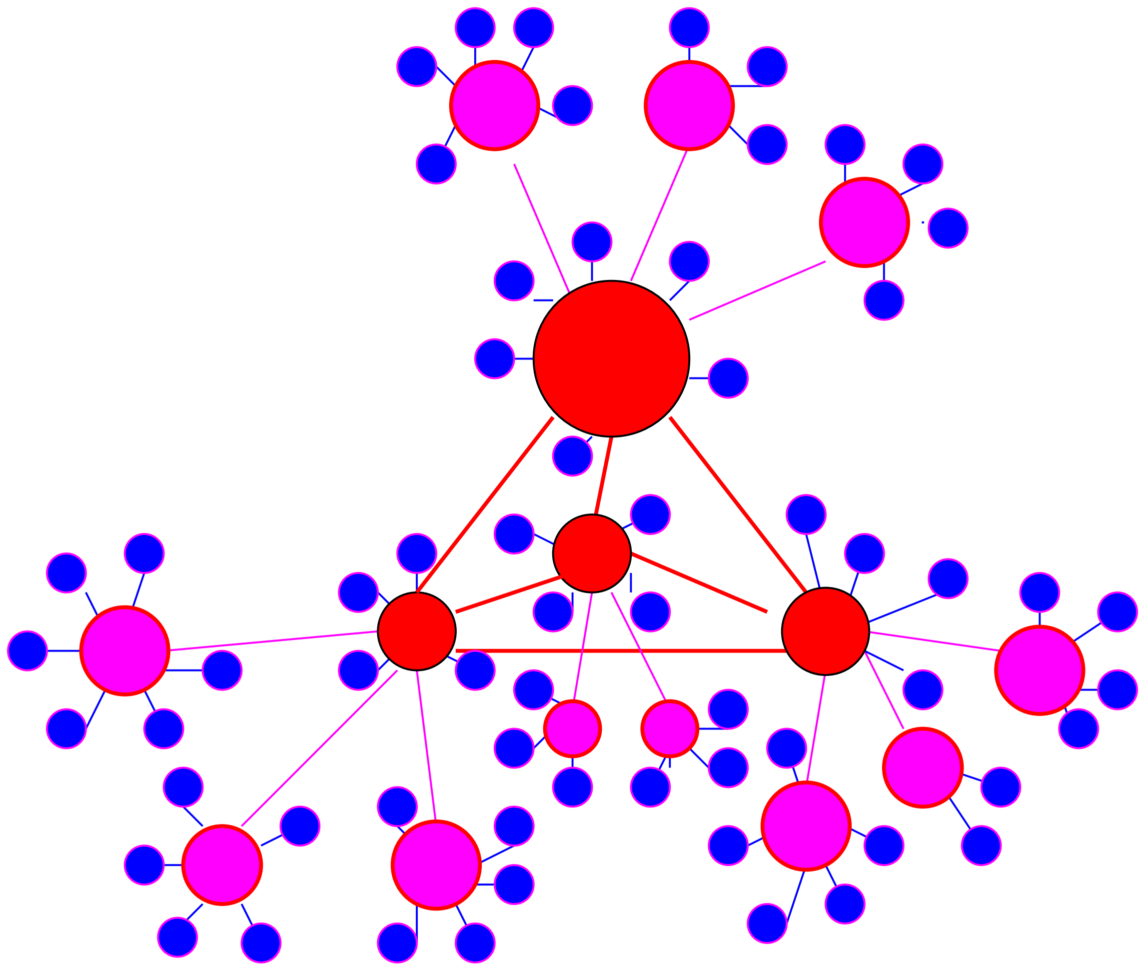
Hannu Reittu
VTT Information Technology
Hannu.Reittu@vtt.fi

# Introduction

- modelling of huge data networks
- Internet graphs, P2P graphs



- power law degree distributions (Faloutsos, ACM/SIGCOMM'99)
- infinite variance $\Rightarrow$ very large nodes
- why?
- typical graph is hierarchical (Reittu-Norros, Globecom 2001, Performance Evaluation, 55 (2004) 3-24)



# Three simple models

**1.’degree model’** node degrees are independent random variables with $\mathbb{P}(D \geq k) = k^{-\tau+1}$, $2 < \tau < 3$, $k = 1, 2, \cdots$ + maximum entropy (see Reittu-Norros, ibid.)

**2. ’edge model’** edge probabilities are independent (Chung-Lu, Internet Mathematics V.1, No 1, 2003, 91- 113) and produce the expected degree sequence.

Expected distance: $L = \frac{-2}{\log(\tau-2)} \log\log(N)$, $N$-number of nodes.

$L/2 \approx$ levels in hierarchy.

’complementary’ models

- good: large scale topology
- bad: small scale clustering incorrect

## 3. ’Jellyfish’

A measurement based model for AS-graphs

See:L. Tauro and C. Palmer and G. Siganos and M. Faloutsos, "A Simple Conceptual Model for the Internet Topology", "Global Internet, San Antonio, Texas", November,2001

- $N(1999) \approx 3N(1997)$
- however, constant distance
- a clique of large nodes

supports a simplified view on topology

## degree model

$D_{i^*} = \max_{i=1,2,\cdots N}(D_i)$

$\mathbb{P}(D_{i^*} \geq k) = 1 - \left(1 - k^{-\tau+1}\right)^N$

$D_{i^*}$ located around $N^{\frac{1}{\tau-1}}$

$y$ large

then $\mathbb{P}\left(D_{i^*} \geq yN^{\frac{1}{\tau-1}}\right) \approx y^{-\tau+1}$ and $\mathbb{P}\left(D_{i^*} \geq \frac{N^{\frac{1}{\tau-1}}}{y}\right) \approx 1 - e^{-y^{-\tau+1}}$

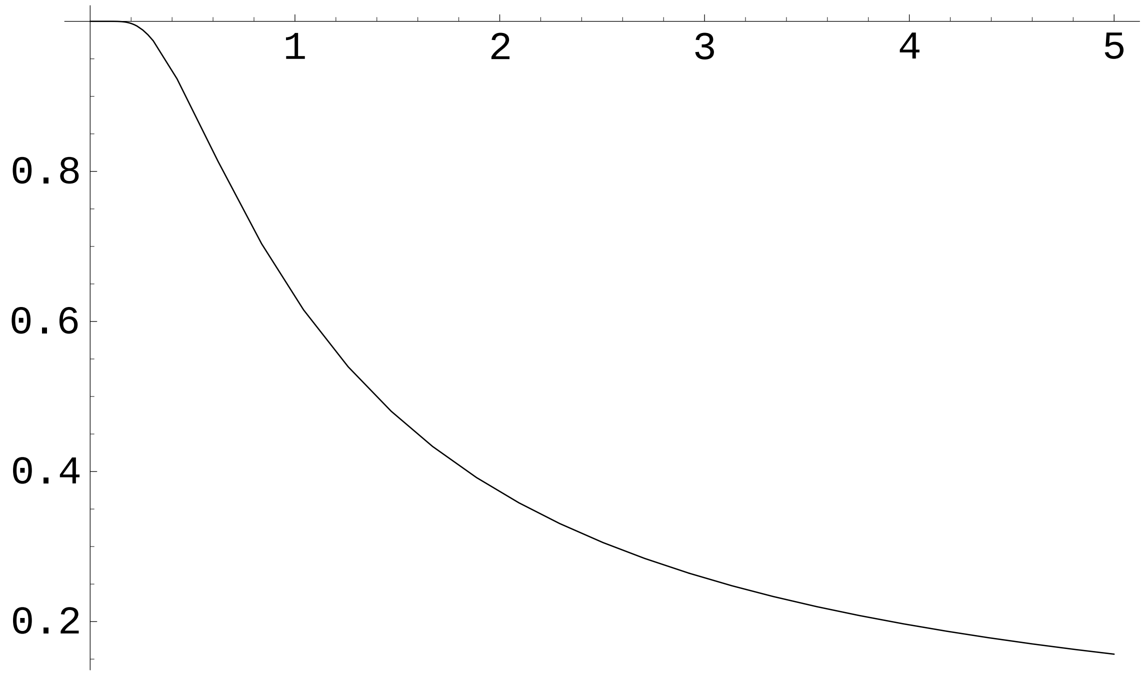



Figure 0.1: $\mathbb{P}(D_{i^*}) \geq xN^{\frac{1}{\tau-1}}), N = 10^6, \tau = 2.1$

Expected degree sequence:

$$d(r) = const \left(\frac{N}{r}\right)^{\frac{1}{\tau-1}}, r = 1, 2, \cdots, N$$

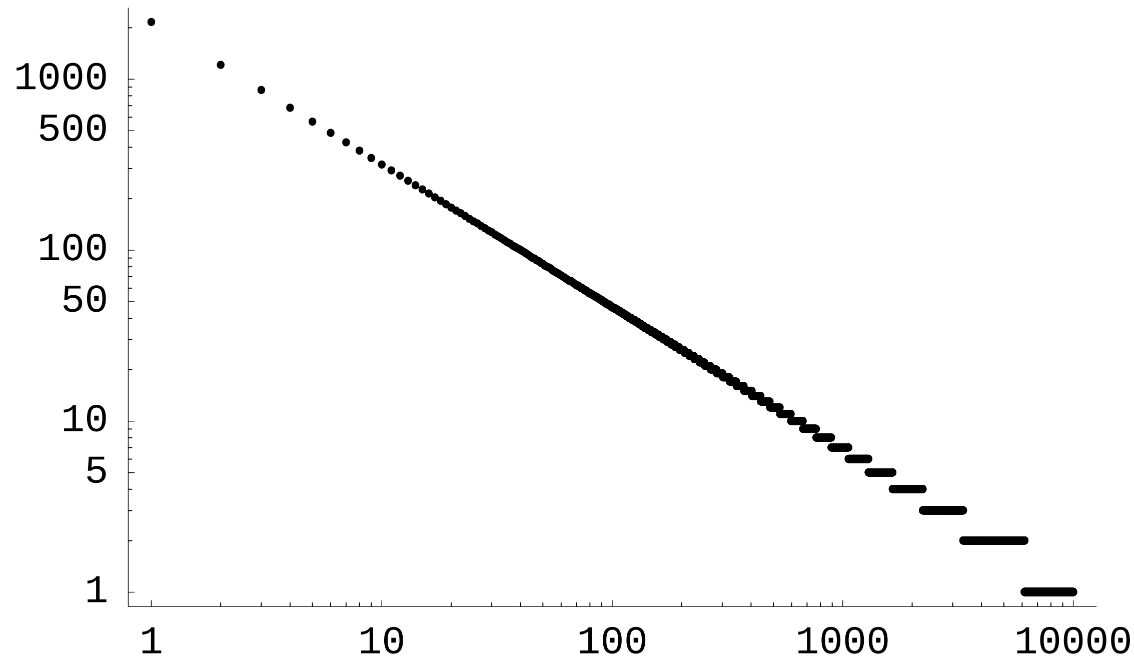



Figure 0.2: $[d(r)]$

As graph instance (CAIDA): $\tau \approx 2.2$, $N \approx 10000$, $\max D \approx 2000$.

- nodes $1$ and $2$ with degrees $d_1$ and $d_2$ and $d_1 d_2 > N$ form an edge a.a.s.
- $\Rightarrow$ cluster of fully connected large nodes with degrees $\geq \sqrt{N}$
- measurements: something like this exists!(L. Tauro and C. Palmer and G. Siganos and M. Faloutsos, ibid.)

# Edge model

$$G = (E, V), G \subset K(N)$$

take the expected degree sequence as granted:

$$d(r) = const \left(\frac{N}{r}\right)^{\frac{1}{\tau-1}}, r = 1, 2, \cdots, N$$

edges are drawn independently with probabilities:

$$\mathbb{P}(\{i, j\} \in E) = d(i)d(j)/Z, Z = \sum_{k=1}^{N} d(k),$$

A problem: $Z = \sum_{k=1}^{N} d(k) = constN$, and $d(i)d(j)/Z > 1$ for a subset of nodes.

A solution (Chung-Lu, ibid.)

$$d(r) = const \left(\frac{N}{r+m}\right)^{\frac{1}{\tau-1}}, r = 1, 2, \cdots, N$$

take $m$ such that $d(i)d(j)/Z < 1$ for $\forall i, j$

A cut-off:

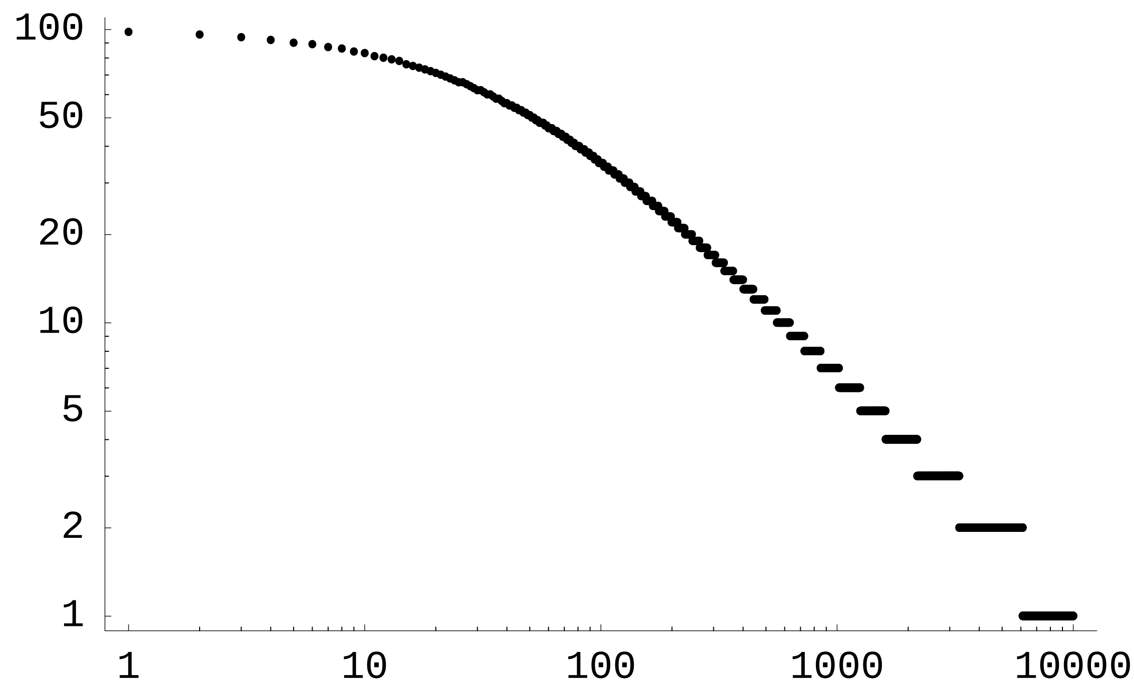



Figure 0.3: Chung-Lu cut-off

Means that the clique of largest nodes is excluded totally!

For As-graph this is clearly not correct

Large As are the most important!

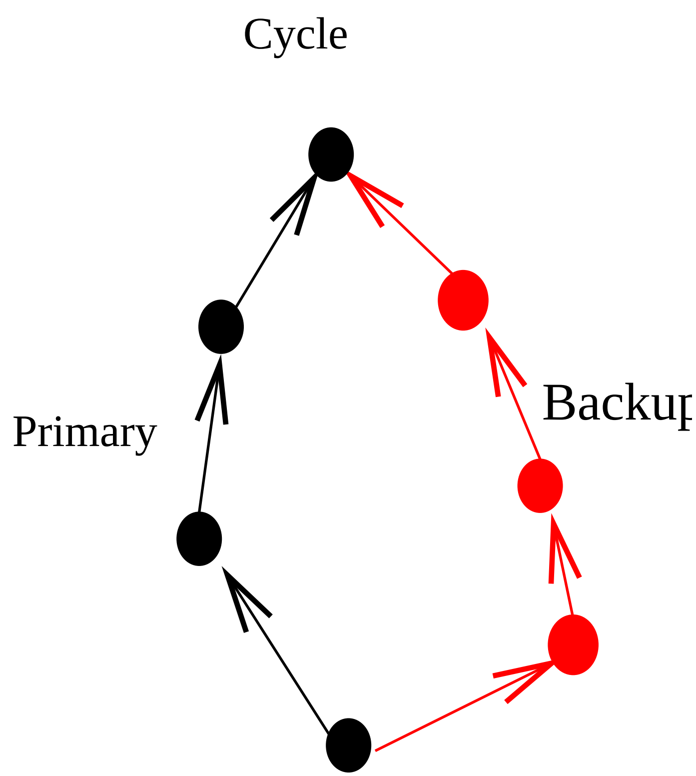
Take simply:

$$\mathbb{P}(\{i, j\} \in E) = \min\{1, d(i)d(j)/Z\}$$

- similar to degree model
- clustering features are easier to analyse

# Cycles

Cycles are resource for protection:



A general random graph result:
M. Kim and M. Medard
( thesis by M.Kim 'Robustness in Large-Scale Random Networks')

**cycles are related to** $\mathbb{E}\left\{D^2\right\}$

in degree model: $\mathbb{E}\left\{D^2\right\} = \infty$

'empirical variance'

$$\sigma^2_{emp} = \frac{1}{N}\sum_i^N D_i^2$$

we suggest (a.a.s.)

$$\sigma^2_{emp} = N^{\frac{3-\tau}{\tau-1}}(1+o(1))$$

and moreover

$$\sigma^2_{emp} = \frac{1}{N}\sum_{\{i:D_i^2>\sqrt{N}\}} D_i^2(1+o(1))$$

suggestion: ’support’ of empirical variance tells where the clustering is located

cycles are at the top of the hierarchy

in particular, short cycles in clique $C = \{i : D_i > \sqrt{N}\}$

Typical value of $\sigma^2_{emp}$

$\sigma^2_T = \sum_{r=1}^{N} \frac{d(r)^2}{N}$,

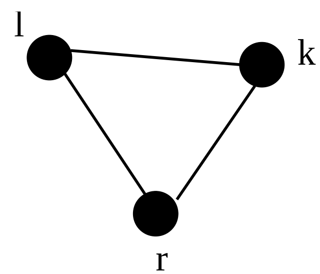
$d(r) = const \left(\frac{N}{r}\right)^{\frac{1}{\tau-1}}$

**edge model**

finite variance

probability of a triangle, $\mathbb{P}(\{r, k, l\} \in \triangle)$

$$\frac{d(r)d(k)}{N}\frac{d(k)d(l)}{N}\frac{d(l)d(r)}{N} = \frac{d(r)^2}{N}\frac{d(k)^2}{N}\frac{d(l)^2}{N}$$




same principle for any cycles

typical value of empirical variance $\sigma_T^2$ appears...,say

$\mathbb{E}\{| \triangle |\} \sim (\sigma_T^2)^3$

short cycles are concentrated in a subgraph with high rank nodes (C)

**real As graph:**

some clustering among small rank nodes as well

however, clique of high rank nodes (C) is reality

$\mid C \mid \approx 10$

number of links in C

$(\mid C \mid -1) \mid C \mid /2 \approx 50$

less than 1 percent of total

however, almost all alternate paths use them!