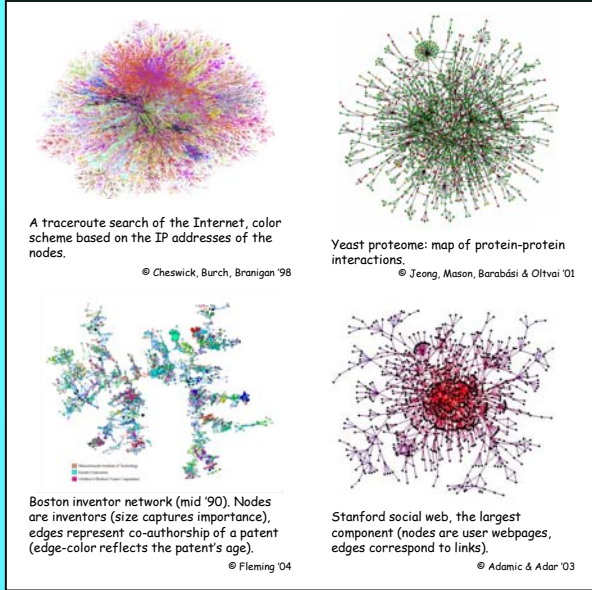


# Graph Model Selection using Maximum Likelihood

Ivona Bezáková, Adam Tauman Kalai, and Rahul Santhanam

## Real-world networks

A sample of real-world networks:



## Random graph models

A variety of popular random graph models:

### Preferential Attachment

[Mitzenmacher '01]

Parameters:  $n, p, q, \gamma$

1. Start with a single vertex.
2. In iteration  $i=2, \dots, n$  create edges between vertices  $i$  and  $\langle i$ . Repeat until c):  
With probability  
a)  $p$  create an outedge from  $i$ ,  
b)  $q$  create an inedge to  $i$ ,  
c)  $1-p-q$  start next iteration.

In 2. other end-point is chosen proportionally to **indegree**  $\rightarrow \gamma$ , or **outdegree**  $\rightarrow \gamma$ , respectively.

### Small World

[Watts - Strogatz '98, Kleinberg '00]

Parameters:  $s, \alpha, \beta$

1. Arrange vertices in  $s \times s$  grid.
2. Add an edge from  $u$  to  $v$  with probability  $\alpha \text{dist}(u, v)^{-\beta}$   
where  $\text{dist}(u, v)$  is the Manhattan distance from  $u$  to  $v$ .
3. Omit isolated vertices.

### Erdős-Rényi

Parameters:  $n, p$

For every pair of vertices include edge with probability  $p$ .

### Powerlaw Random Graph

[Bollobás '85; Aiello, Chung, Lu '00]

Parameters:  $n, \beta_{in}, c_{in}, \beta_{out}, c_{out}$

1. For every vertex: generate indegree, outdegree from powerlaw distribution with parameters  $\beta_{in}, c_{in}$  and  $\beta_{out}, c_{out}$ , respectively.
2. If the sum of indegrees differs from the sum of outdegrees, output an empty graph.
3. Otherwise, for every vertex create the corresponding number of in/outedges.
4. Randomly match the inedges with outedges.

### Powerlaw distribution

with parameters  $\beta$  (exponent),  $c$  (cutoff):  
Choose  $x \in \{1, \dots, c\}$  proportional to  $x^{-\beta}$ , i.e.  
 $\text{Prob}(x = a) = \alpha^{-\beta} / \sum_{i=1, \dots, c} i^{-\beta}$

These models (with the exception of the standard ER model) were designed to replicate properties observed in real-world networks, such as powerlaw distribution of the degrees, or the small world phenomenon (high clustering, small diameter).

## Which model is the best for a given real-world network?

### Contributions:

- proposed objective ranking method
- demonstrated feasibility of the approach
  - designed algorithms for the four models
  - applied ranking method to real datasets
- parameter estimation for random graph models

Our approach: ranking by Maximum Likelihood.

$$\text{Score (model)} = -\log \text{Prob (model generates } \mathcal{G} \text{)}$$

Previous graph comparison techniques generally ranked models by their ability to reproduce certain features observed in the dataset, introducing subjectivity (which features to select?) in the ranking mechanism.

### Involved technicalities:

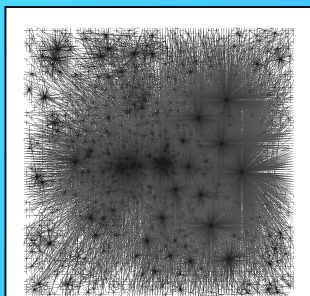
- To use our approach, a model should be able to generate any given instance graph with non-zero probability. We tweak some of the models for this purpose.
- Some models, like the PA model, impose ordering on nodes, irrelevant to the node labeling of the dataset. To account for this, we consider only symmetric models which generate every labeling equally likely.
- Algorithmic issues. Our algorithms range from simple to fairly elaborate parallel MCMC approach using the hit-and-run technique.

## Experiments

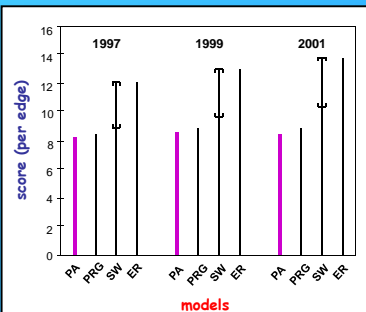
To demonstrate feasibility of the approach, we ranked the above models on three snapshots of the AS-level Internet topology graph, from 1997, 1999, and 2001.

### Dataset characteristics:

	vertices	edges
'97	3,117	6,024
'99	6,266	13,681
'01	11,080	25,485



The 2001 dataset. Here shown embedded in the grid as a byproduct of our small world computation.



A histogram showing the model's scores on the three datasets, scaled down by the number of edges in the dataset. For every dataset the preferential attachment model scores highest. For the small world model we computed an upper and a lower-bound on the score which are shown as an "uncertainty region".

	PA	PRG	SW	ER
'97	8.30 $p = 0.58$ $q = 0.08$ $\gamma = 0.5$	8.60 $\beta_{in} = 1.55, c_{in} = 610$ $\beta_{out} = 2.39, c_{out} = 69$	8.96 $s = 56$ $\alpha = 0.111$ $\beta = 1.9$	12.10 $p = 6.2e-4$
'99	8.55 $p = 0.61$ $q = 0.08$ $\gamma = 0.4$	8.83 $\beta_{in} = 1.57, c_{in} = 1410$ $\beta_{out} = 2.44, c_{out} = 172$	9.76 $s = 80$ $\alpha = 0.092$ $\beta = 1.8$	12.93 $p = 3.5e-4$
'01	8.58 $p = 0.63$ $q = 0.07$ $\gamma = 0.3$	8.85 $\beta_{in} = 1.57, c_{in} = 2421$ $\beta_{out} = 2.50, c_{out} = 214$	10.42 $s = 106$ $\alpha = 0.088$ $\beta = 1.8$	13.68 $p = 2.1e-4$

Blue numbers represent the models' scores (per edge) on the three datasets. For the small world model we state the computed lower bound on its score. We also show the optimal parameters which achieve the best score of the model.