Complex Networks and their Analysis with Random Walks

Daniel R. Figueiredo
Federal University of Rio de Janeiro, Brazil

Konstantin Avrachenkov
INRIA Sophia Antipolis, France

ITC 27 – Tutorial – Ghent, Belgium

COPPE UFRJ

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Objectives and Organization

- First contact
  - o "networks everywhere"
- Empirical findings of networks
  - o important commonalities
- Mathematical models for networks

1h

Daniel

1h

break

- Random walk premier
  - o simple yet profound
- Applications of random walks
  - o sampling, ranking, clustering, etc

.5h

Kostia

1.5h

How to study networks?

- Networks obtained empirically provide one, complicated instance
  - eg., Facebook, Web, Neurons
- Need to work with abstractions

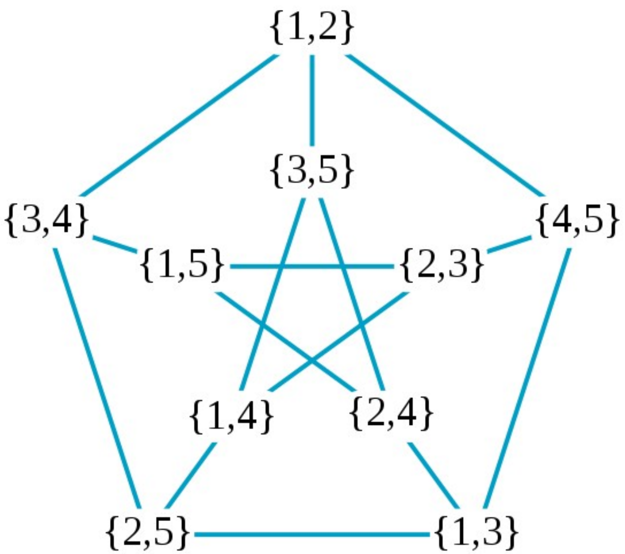
Mathematical models to the rescue!

- Simplify reality for fundamental understanding of various properties
- Models for network structure (connectedness)

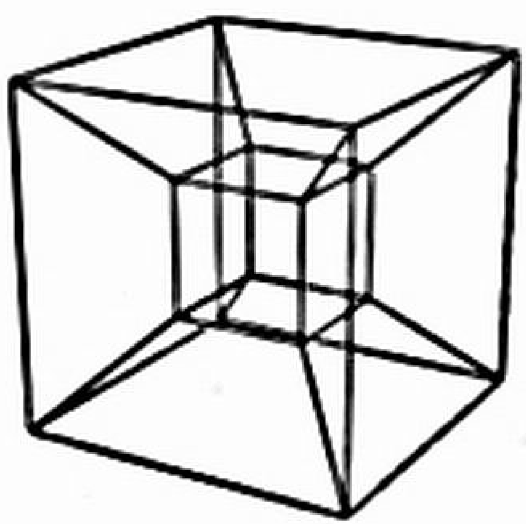
Deterministic Models

- Network structure is deterministic
  - o rules uniquely determine network formation
- Structural properties are deterministic

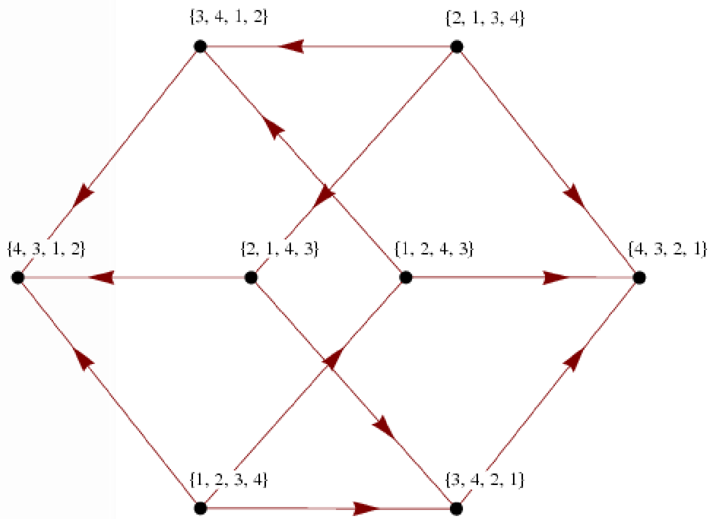
Examples of models?



Kneser Graph



Hypercube

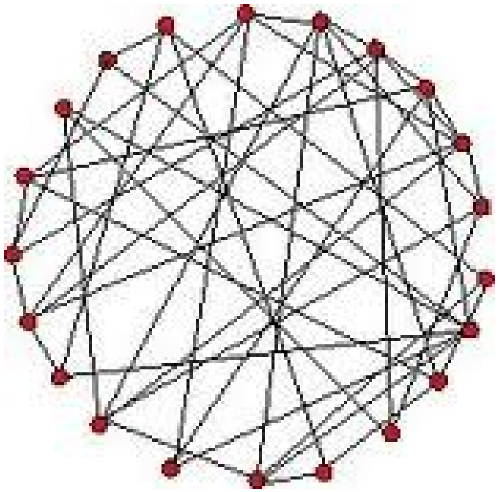


Caley Graphs

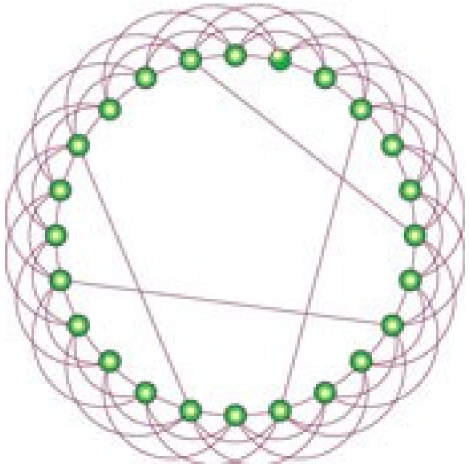
Probabilistic Models

- Network structure is random
  - probabilistic rules determine network formation
- Structural properties are random

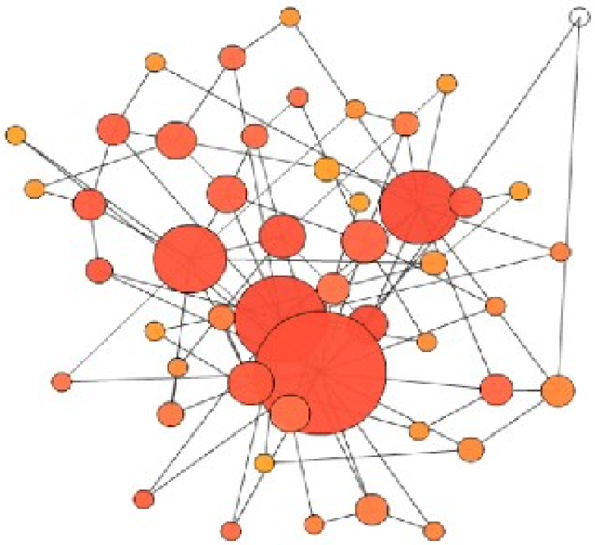
Examples of models?



Erdös-Réyni



Watts-Strogatz



Barabási-Albert

G(n,p) Model

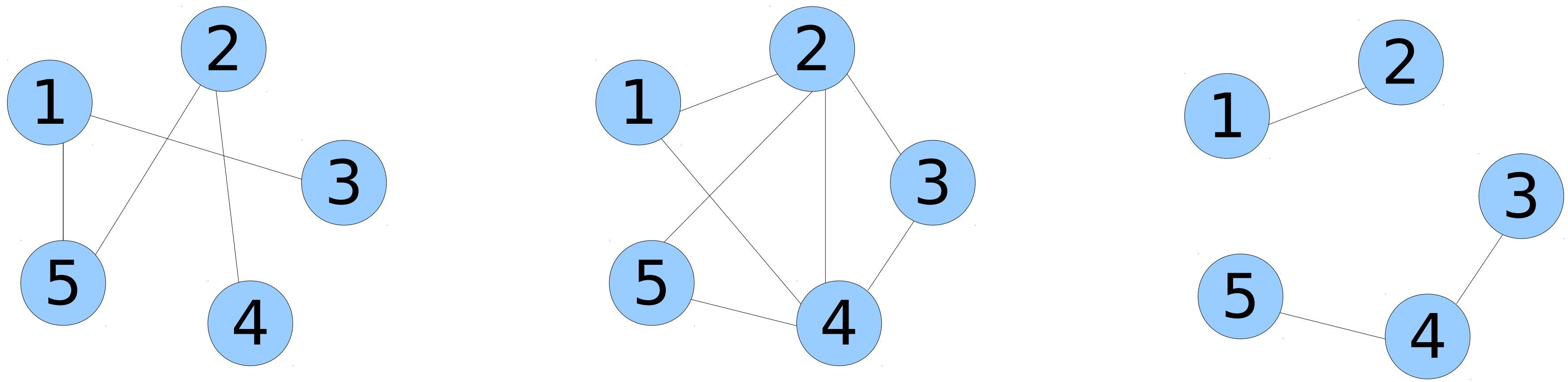
- Classic and most widely studied model for random graphs
  - o first studied by Erdós and Rényi in 50s
  - o aka. Binomial model, Erdós-Rényi model
- The model
  - o network has n labeled nodes
  - o each possible edge is present with probability p, independently

Very simple model!

- yet surprisingly rich structures emerge

G(*n*,*p*) Example

- Given its two parameter, *n* and *p*, what network is formed?
- Ex. n=5, p=0.25



- Network is random! A realization of the random process (choosing edges)

Characterizing the G(*n*,*p*)

What kind of networks does G(n,p) generates?

- Is it a connected graph? What is the degree distribution? What is the clustering? Etc

- Random structure depends on *n* and *p*
- Characterize structural properties for large *n* and scaling *p*
- Determine conditions for properties to be present with high probability

Simple Properties

- Sample space of G(*n*,*p*)?
  - S = all possible graphs with *n* nodes
- What is the sample space size?

$$|S| = 2^{\binom{n}{2}}$$

← Every possible edge can either be present or absent

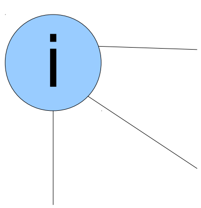
  - n=15, |S| > number of atoms on universe!
- Probability of generating a given graph, defined by E = {e_1, e_2, ..., e_k}?

$$P(generate\ set\ E) = p^{|E|}(1-p)^{\binom{n}{2}-|E|}$$

  - depends only on *k*, and not the set E

Degree in G(*n*,*p*)

- What is the degree of a given node?
  - degree is random!
- What is degree distribution of a given node?



$$P[D=k]? \quad k = 0, 1, \ldots, n-1$$

- Each edge incident on node i with probability *p*

$$P[D=k]=\binom{n-1}{k} p^k (1-p)^{n-1-k}$$

Binomial distribution

- Expected degree

$$E[D]=(n-1)p$$

Connected Components

- Is G(n,p) connected? Size of connected components?
- Let *p* be a function of *n*, thus *p*(*n*)
  - if p(n) = z/(n-1) for constant z, then E[D] = z
- z < 1 (subcritical)
  - all CC have size O(log n), many components
- z > 1 (supercritical)
  - largest CC has size Ω(n), all others O(log n)
- z = Ω(log n)
  - single CC, network is connected

Phase transitions on graph structure!

- results valid with high probability as *n* grows

G(n,p) and Real Networks

- Is G(n,p) a good model for real networks?

(most) Real Networks

- Short distances
- High clustering
- Heavy-tailed degree distribution

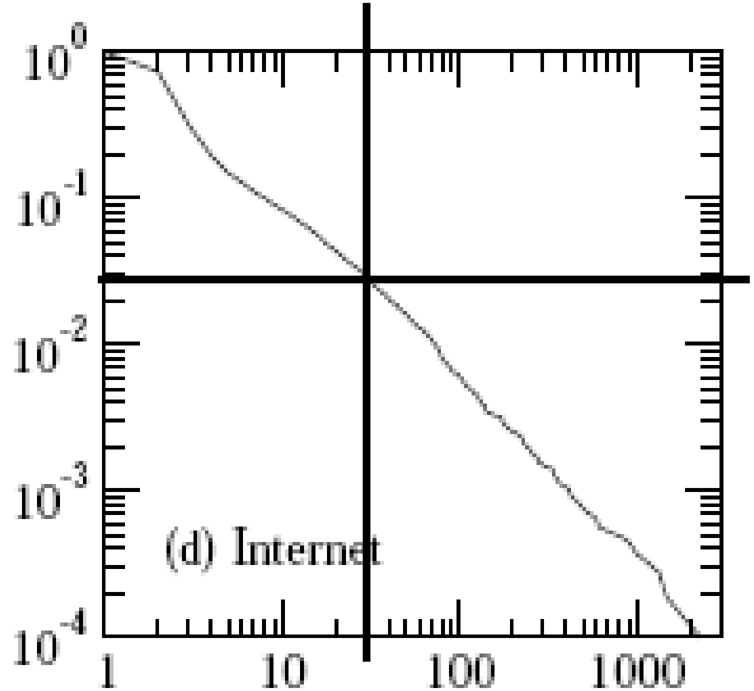
G(n,p)

- Short distances
- Low clustering (*p*)
- Binomial degree distribution

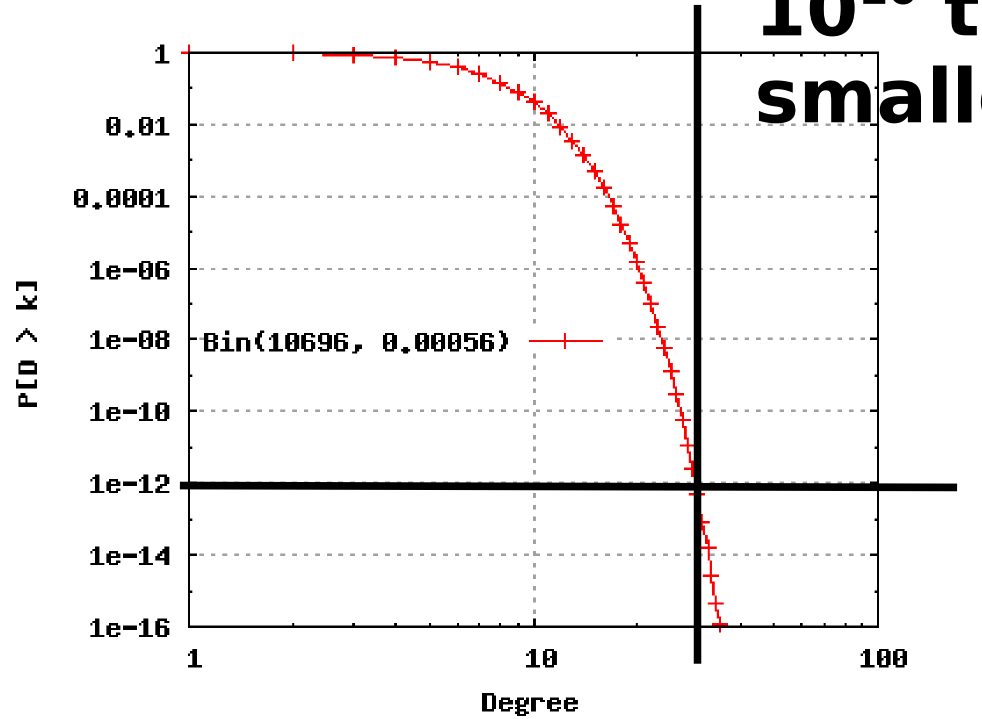
Fundamentally important, yet fundamentally different

Example

- AS Graph, 11K nodes, 32K edges
- Apply G(*n*,*p*) preserving *n* and avg deg (=6.16)
  - n=11K, p=0.00056
- Clustering: data=0.39, model=0.00056
  - almost **1000** times smaller
- Degree distribution

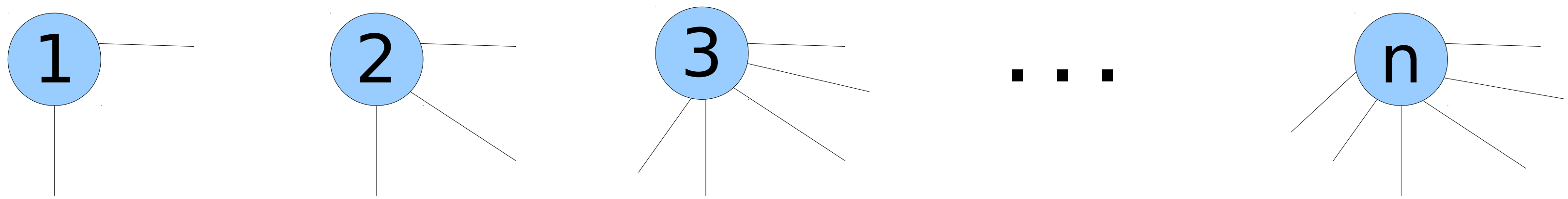


P[D>30] is 10^{10} times smaller!



Configuration Model

- **Idea**: specify degree of network nodes, connect them at random
- Parameters: degree sequence $d_1, d_2, \ldots, d_n$
  - degree sum has to be even



- Connect edge points at random
  - multiple edges unlikely if network is very large
- Generalization of G(n,p)
  - allows for arbitrary degree distribution
  - still very low clustering

Generative Models

- Grow network iteratively
  - add nodes and edges over time
- Capture some fundamental aspect of network formation
  - structure is consequence of iterative rules

Various models proposed in this class

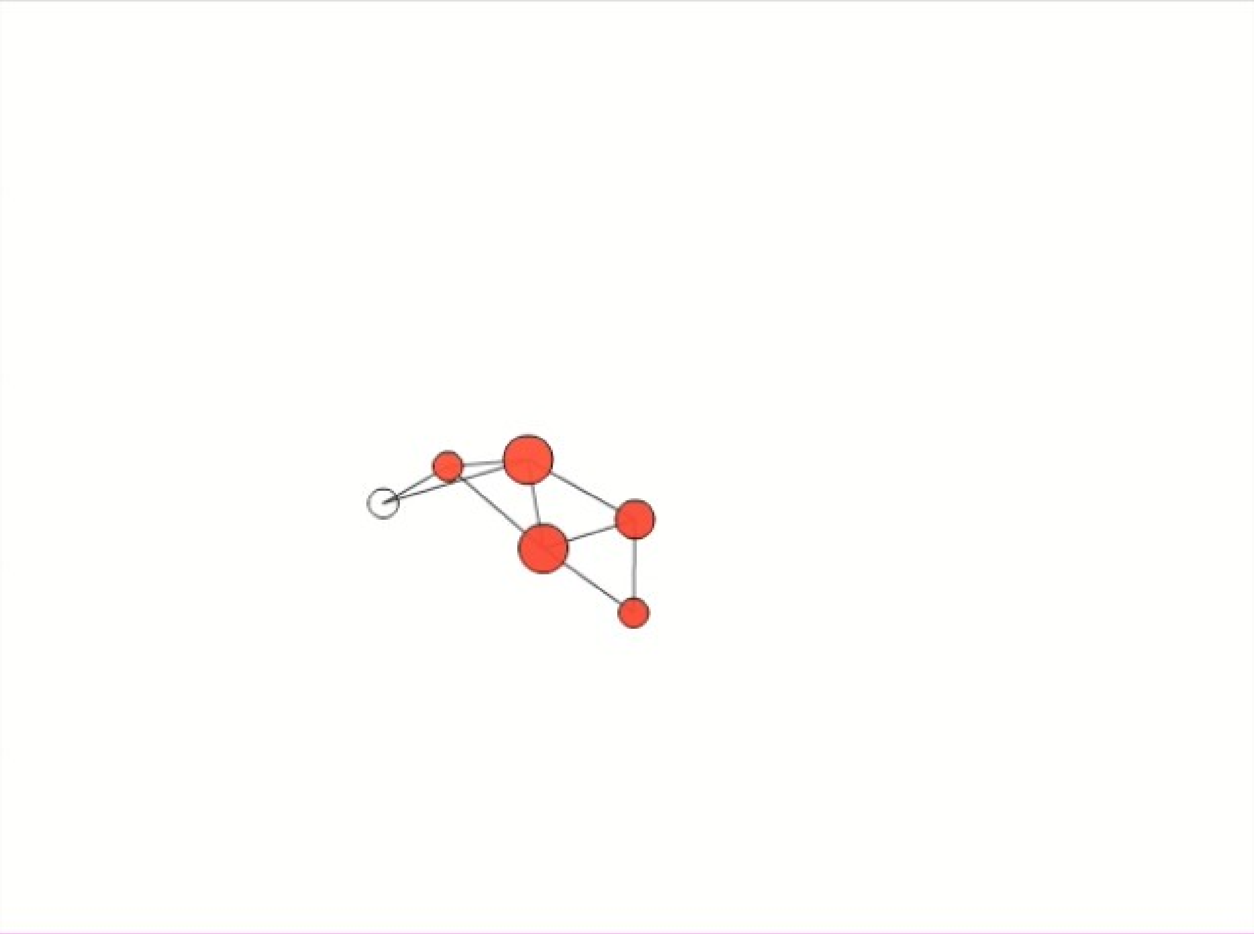
Preferential Attachment

- Old phenomenon for growing dynamics
  - cumulative advantage, rich-gets-richer, Matthew effect, etc
- **Idea**: accumulated resources promote the accumulation of further resources
- Various empirical observations
  - word usage, city growth, etc
- Applied to networks
  - paper citation networks: Solla Price, 50's
  - hyperlinks on web: Barabasi-Albert, 99

BA Model

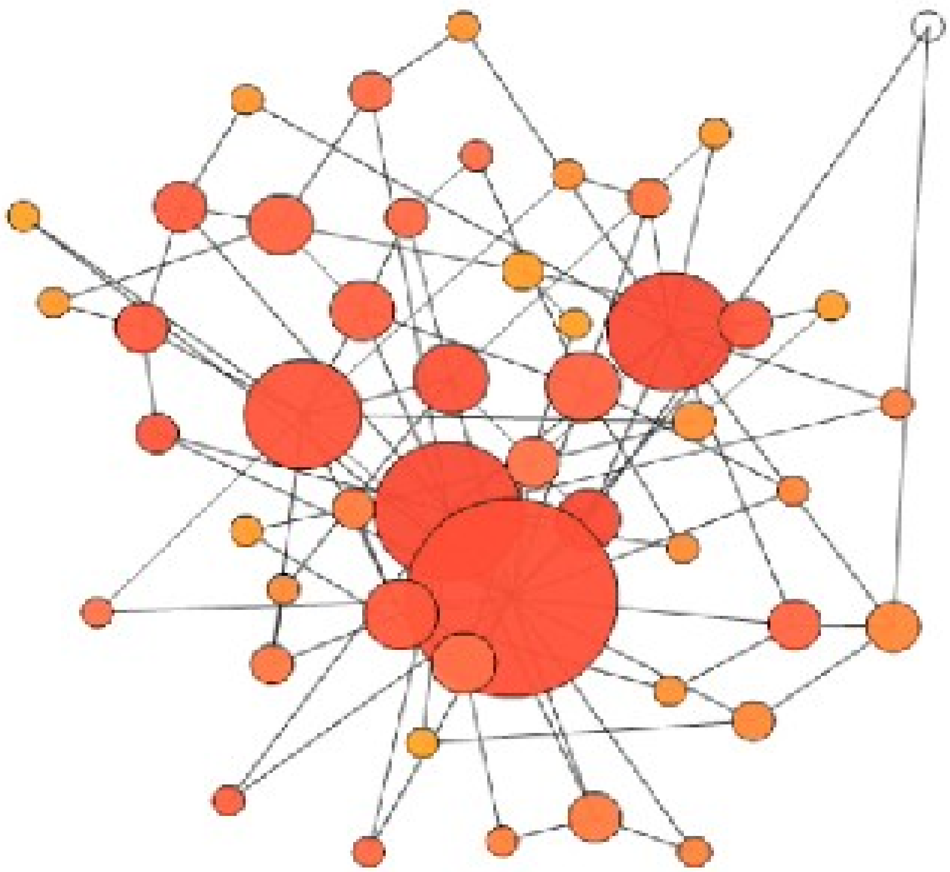
- Barabási-Albert model, proposed in 1999 (*Science* paper with over 23K citations)
  - preferential attachment based on node degree
- At each step
  - add one node with degree m
  - choose each neighbor with probability proportional to their degree
- Parameters
  - small initial network
  - m, number of edges added with each node

BA Model Example



- Initial network is a triangle, m = 2
  - size of node proportional to node degree

BA Model Example



- What is happening?
- many small degree, few high degree

Preferential attachment → **?** → **Heavy-tailed degree distribution**

BA Model Properties

- Analysis via continuous approximations with differential equations
- $d_u(t)$: average degree of node *u* at time *t*
  - t_u : time node *u* entered network

$$d_u(t) = m\left(\frac{t}{t_u}\right)^{1/2}$$

- Assuming t_u is uniformly chose [1, *t*]

$$P[d_u(t) = k] \approx \frac{2m^2}{k^3}$$

Power law degree distribution!

Limitations of BA

- Power law exponent is fixed, equal to 3
  - real networks have various decays
- Older nodes always have higher degrees
  - real networks new nodes can take over; Facebook
- Clustering coefficient is very low
- No new edges among existing nodes

Many, many more models!

- to address these and other limitations

References

Textbooks

- Mark Newman, *Networks: An Introduction*, 2010
- Albert-László Barabási, *Network Science,* 2015
- B. Bollobas, *Random Graphs, 2001*

Papers

- A.-L. Barabási, R. Albert, *Emergence of scaling in random networks*, Science 1999
- *Scale-Free Networks: A Decade and Beyond*, Science (special issue) 2009