Online Social Networks and Media

Network models

What is a network model?

- Informally, a network model is a process (radomized or deterministic) for generating a graph
- Models of static graphs
  - input: a set of parameters Π, and the size of the graph n
  - output: a graph $G(\Pi,n)$
- Models of evolving graphs
  - input: a set of parameters Π, and an initial graph G_0
  - output: a graph G_t for each time t

Families of random graphs

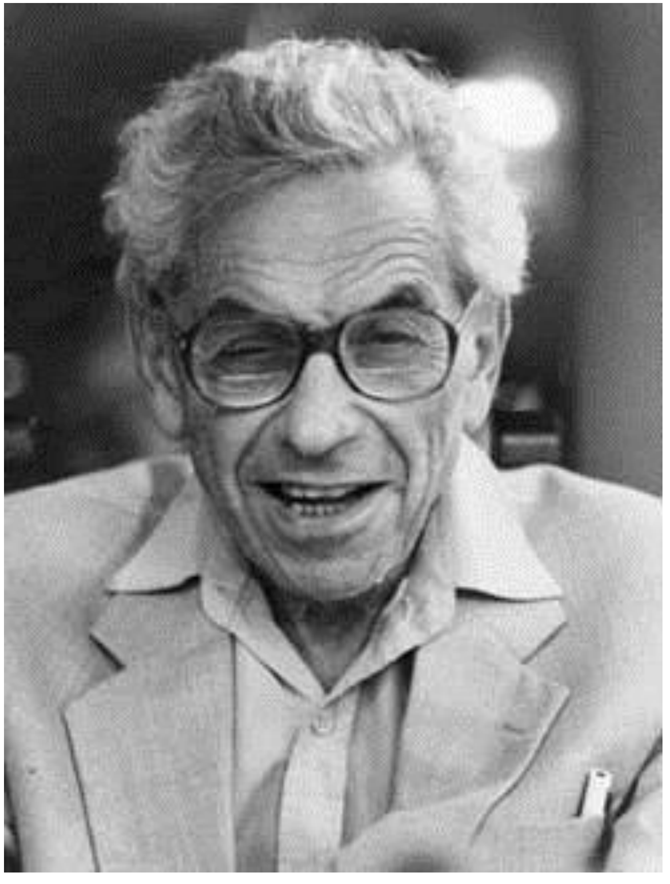
- A deterministic model D defines a single graph for each value of n (or t)

- A randomized model R defines a probability space $\langle G_n, P \rangle$ where G_n is the set of all graphs of size n, and P a probability distribution over the set G_n (similarly for t)
  - we call this a family of random graphs R, or a random graph R

Why do we care?

- Creating models for real-life graphs is important for several reasons
  - Create data for simulations of processes on networks
  - Identify the underlying mechanisms that govern the network generation
  - Predict the evolution of networks

Erdös-Renyi Random graphs



Paul Erdös (1913-1996)

Erdös-Renyi Random Graphs

- The $G_{n,p}$ model
  - input: the number of vertices n, and a parameter p, $0 \leq p \leq 1$
  - process: for each pair (i,j), generate the edge (i,j) independently with probability p

- Related, but not identical: The $G_{n,m}$ model
  - process: select m edges uniformly at random

Graph properties

- A property P holds almost surely (a.s.) (or for almost every graph), if

$$\lim_{n\to\infty} P[G \text{ has } P] = 1$$

- Evolution of the graph: which properties hold as the probability p increases?
  - different from the evolving graphs over time that we saw before

- Threshold phenomena: Many properties appear suddenly. That is, there exist a probability p_c such that for $p<p_c$ the property does not hold a.s. and for $p>p_c$ the property holds a.s.

The giant component

- Let $z=np$ be the average degree
- If $z < 1$, then almost surely, the largest component has size at most $O(\ln n)$
- if $z > 1$, then almost surely, the largest component has size $\Theta(n)$. The second largest component has size $O(\ln n)$
- if $z = \omega(\ln n)$, then the graph is almost surely connected.

The phase transition

- When $z=1$, there is a phase transition
  - The largest component is $O(n^{2/3})$
  - The sizes of the components follow a power-law distribution.

Random graphs degree distributions

- The degree distribution follows a binomial

$$p(k) = B(n;k;p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Assuming $z=np$ is fixed, as $n \to \infty$, $B(n,k,p)$ is approximated by a Poisson distribution

$$p(k) = P(k;z) = \frac{z^k}{k!} e^{-z}$$

- Highly concentrated around the mean, with a tail that drops exponentially

Other properties

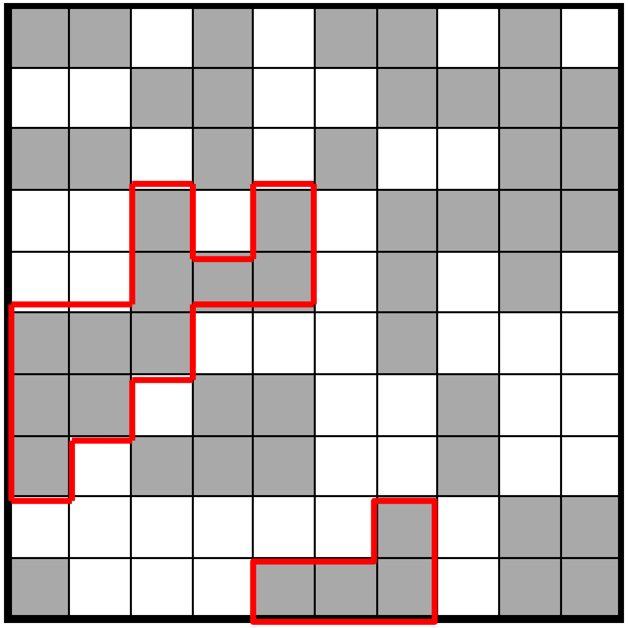
- Clustering coefficient
  - $C = z/n$
- Diameter (maximum path)
  - $L = \log n / \log z$

Phase transitions

- Phase transitions (a.k.a. Threshold Phenomena, Critical phenomena) are observed in a variety of natural or human processes, and they have been studied extensively by Physicists and Mathematicians
  - Also, in popular science: “The tipping point”
- Examples
  - Water becoming ice
  - Percolation
  - Giant components in graphs
- In all of these examples, the transition from one state to another (e.g., from water to ice) happens almost instantaneously when a parameter crosses a threshold
- At the threshold value we have critical phenomena, and the appearance of Power Laws
  - There is no characteristic scale.

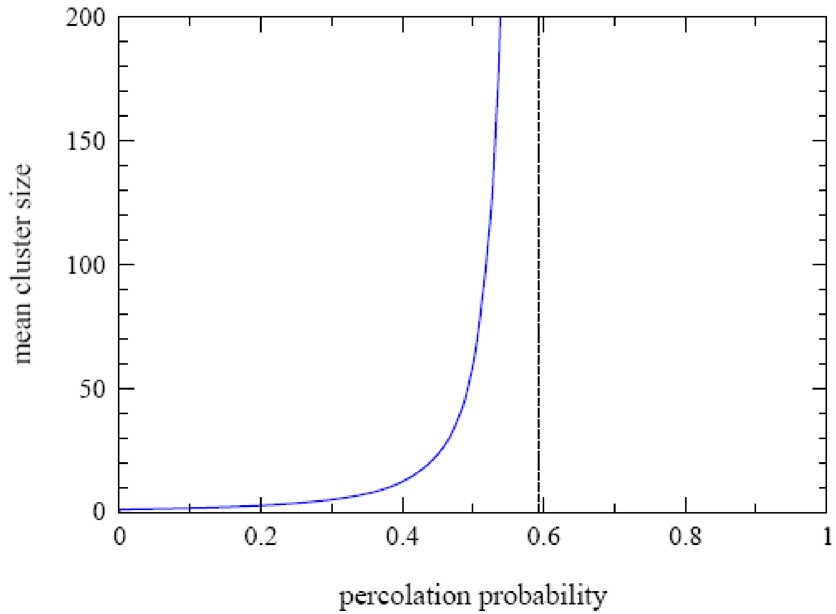
Percolation on a square lattice

- Each cell is occupied with probability p



- What is the mean cluster size?

Critical phenomena and power laws

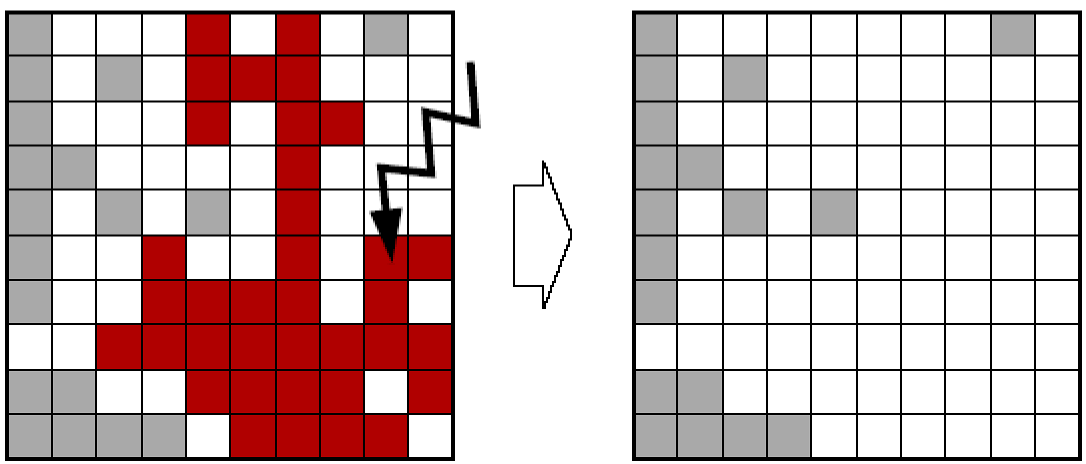


$p_c = 0.5927462...$

- For $p < p_c$ mean size is independent of the lattice size
- For $p > p_c$ mean size diverges (proportional to the lattice size - percolation)
- For $p = p_c$ we obtain a power law distribution on the cluster sizes

Self Organized Criticality

- Consider a dynamical system where trees appear in randomly at a constant rate, and fires strike cells randomly



- The system eventually stabilizes at the critical point, resulting in power-law distribution of cluster (and fire) sizes

The idea behind self-organized criticality (more or less)

- There are two contradicting processes
  - e.g., planting process and fire process
- For some choice of parameters the system stabilizes to a state that no process is a clear winner
  - results in power-law distributions
- The parameters may be tunable so as to improve the chances of the process to survive
  - e.g., customer’s buying propensity, and product quality.

- Could we apply this idea to graphs?

Random graphs and real life

- A beautiful and elegant theory studied exhaustively

- Random graphs had been used as idealized network models

- Unfortunately, they don’t capture reality…

Departing from the ER model

- We need models that better capture the characteristics of real graphs
  - degree sequences
  - clustering coefficient
  - short paths

Graphs with given degree sequences

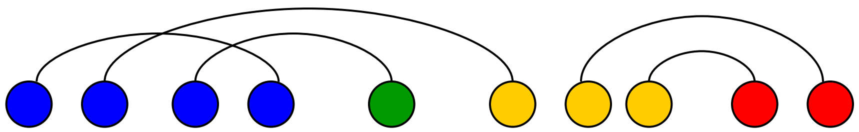
- The configuration model
  - input: the degree sequence $[d_1,d_2,\ldots,d_n]$
  - process:
    - Create d_i copies of node i
    - Take a random matching (pairing) of the copies
      - self-loops and multiple edges are allowed

- Uniform distribution over the graphs with the given degree sequence

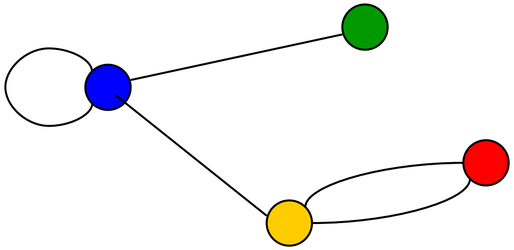
Example

- Suppose that the degree sequence is

- Create multiple copies of the nodes



- Pair the nodes uniformly at random
- Generate the resulting network



Other properties

- The giant component phase transition for this model happens when

$$\sum_{k=0}^{\infty} k(k-2)p_k = 0$$

p_k: fraction of nodes with degree k

- The clustering coefficient is given by

$$C = \frac{z}{n}\left(\frac{\langle d^2\rangle - \langle d\rangle}{\langle d\rangle^2}\right)^2$$

- The diameter is logarithmic

Power-law graphs

- The critical value for the exponent α is

$$\alpha = 3.4788...$$

- The clustering coefficient is

$$C \propto n^{-\beta} \qquad \beta = \frac{3\alpha - 7}{\alpha - 1}$$

- When α<7/3 the clustering coefficient increases with n

Graphs with given expected degree sequences

- Input: the degree sequence $[d_1, d_2, \dots ,d_n]$
- m = total number of edges

- Process: generate edge (i,j) with probability $d_i d_j/m$
  - preserves the expected degrees
  - easier to analyze

However…

- The problem is that these models are too contrived
- It would be more interesting if the network structure emerged as a side product of a stochastic process rather than fixing its properties in advance.

Preferential Attachment in Networks

- First considered by [Price 65] as a model for citation networks (directed)
  - each new paper is generated with m citations (mean)
  - new papers cite previous papers with probability proportional to their indegree (citations)
  - what about papers without any citations?
    - each paper is considered to have a “default” a citations
    - probability of citing a paper with degree k, proportional to $k+a$

- Power law with exponent $\alpha = 2+a/m$

Practical Issues

- The model is equivalent to the following:
  - With probability m/(m+a) link to a node with probability proportional to the degree.
  - With probability a/(m+a) link to a node selected uniformly at random.
- How do we select a node with probability proportional to the degree?
  - Select a node and pick on of the nodes it points to.
  - In practice:
    - Maintain a list with the endpoints of all the edges seen so far, and select a node from this list uniformly at random
    - Append the list each time new edges are created.

Barabasi-Albert model

- The BA model (undirected graph)
  - input: some initial subgraph G_0, and m the number of edges per new node
  - the process:
    - nodes arrive one at the time
    - each node connects to m other nodes selecting them with probability proportional to their degree
    - if $[d_1,\ldots,d_t]$ is the degree sequence at time t, the node $t+1$ links to node i with probability

$$\frac{d_i}{\sum_i d_i} = \frac{d_i}{2mt}$$

- Results in power-law with exponent $\alpha = 3$

The mathematicians point of view [Bollobas-Riordan]

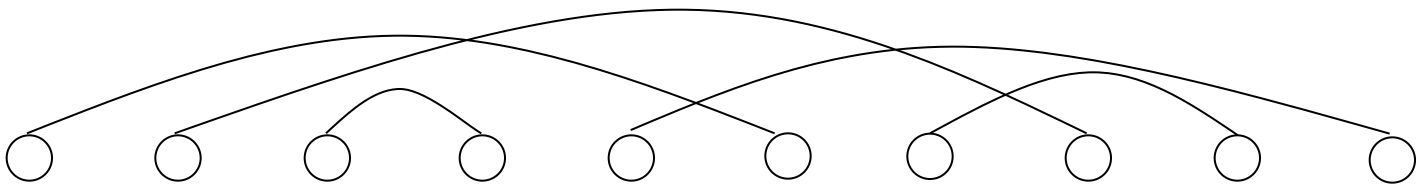
- Self loops and multiple edges are allowed
- For the single edge problem:
  - At time t, a new vertex v, connects to an existing vertex u with probability
    $$\frac{d_u}{2t-1}$$
  - it creates a self-loop with probability
    $$\frac{1}{2t-1}$$
- If m edges, then they are inserted sequentially, as if inserting m nodes
  - the problem reduces to studying the single edge problem.

The Linearized Chord Diagram (LCD) model

- Consider 2n nodes labeled {1,2,…,2n} placed on a line in order.

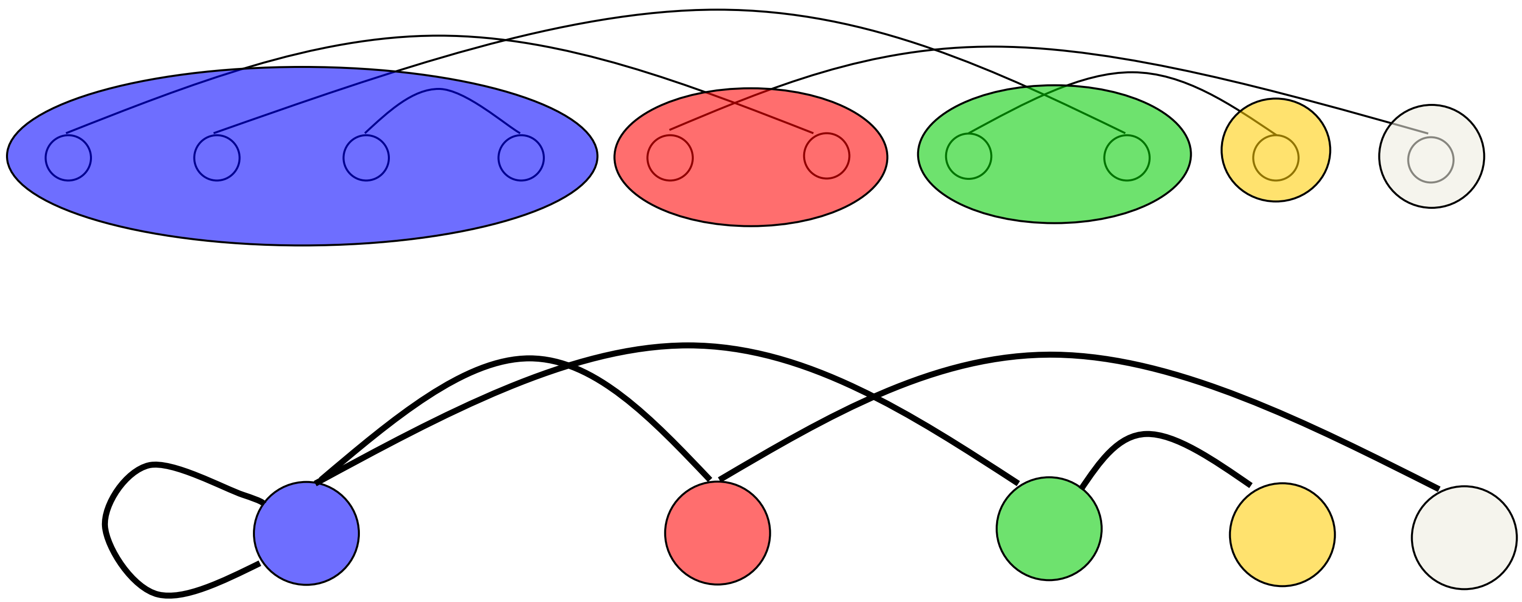
Linearized Chord Diagram

- Generate a random matching of the nodes.



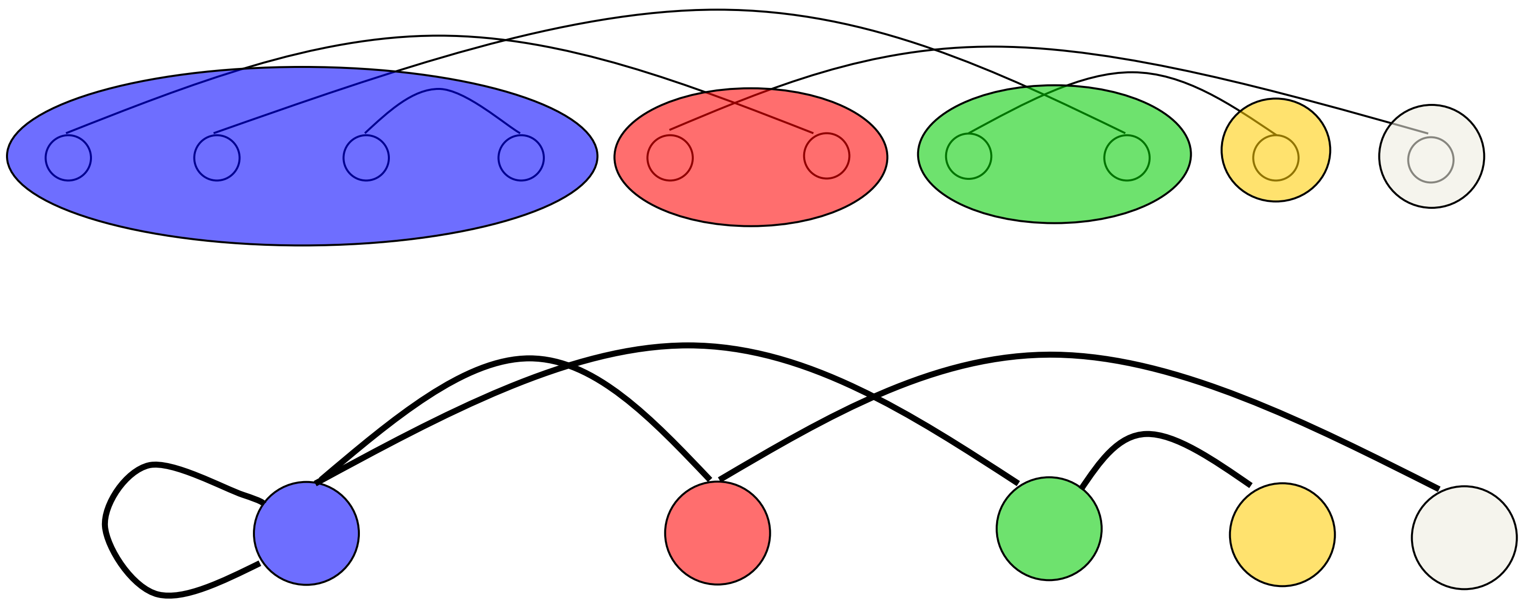
Linearized Chord Diagram

- Starting from left to right identify all endpoints until the first right endpoint. This is node 1. Then identify all endpoints until the second right endpoint to obtain node 2, and so on.



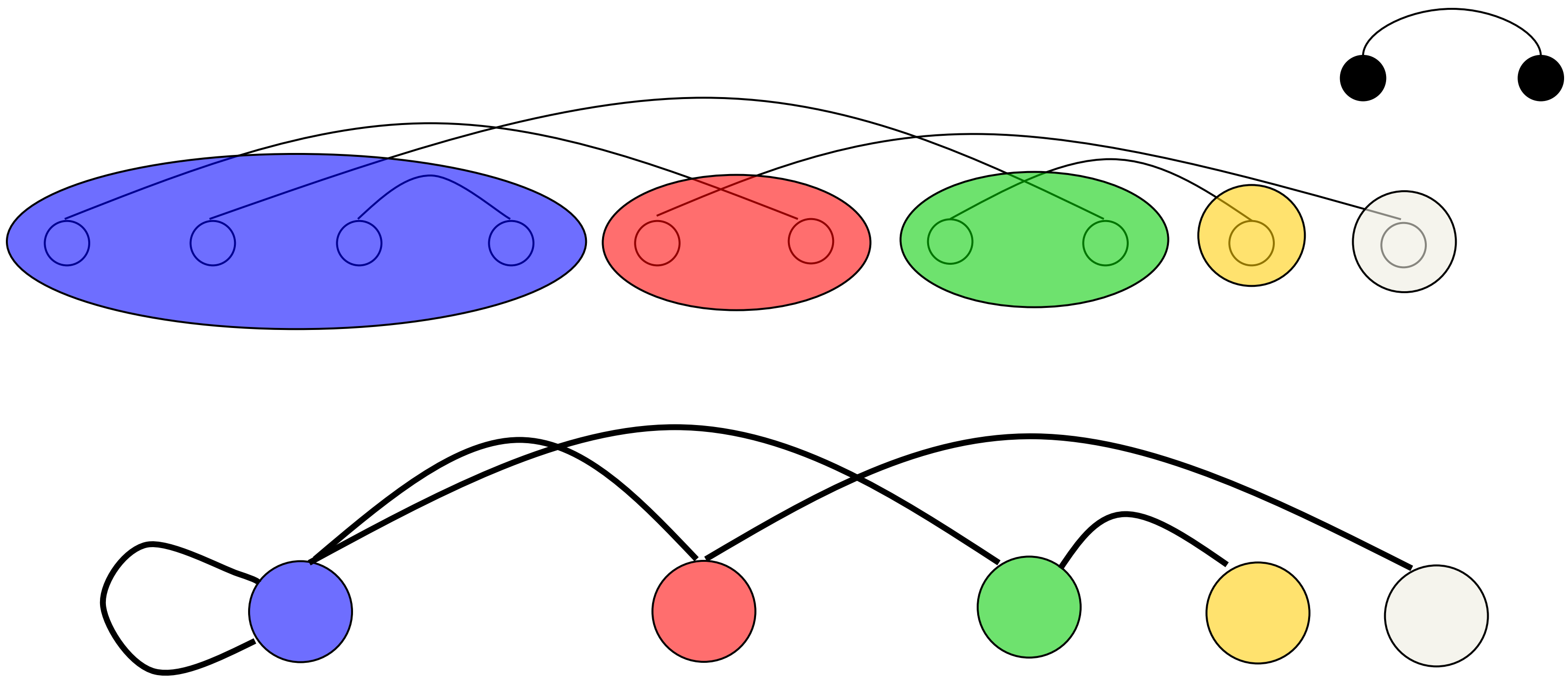
Linearized Chord Diagram

- Uniform distribution over matchings gives uniform distribution over all graphs in the preferential attachment model



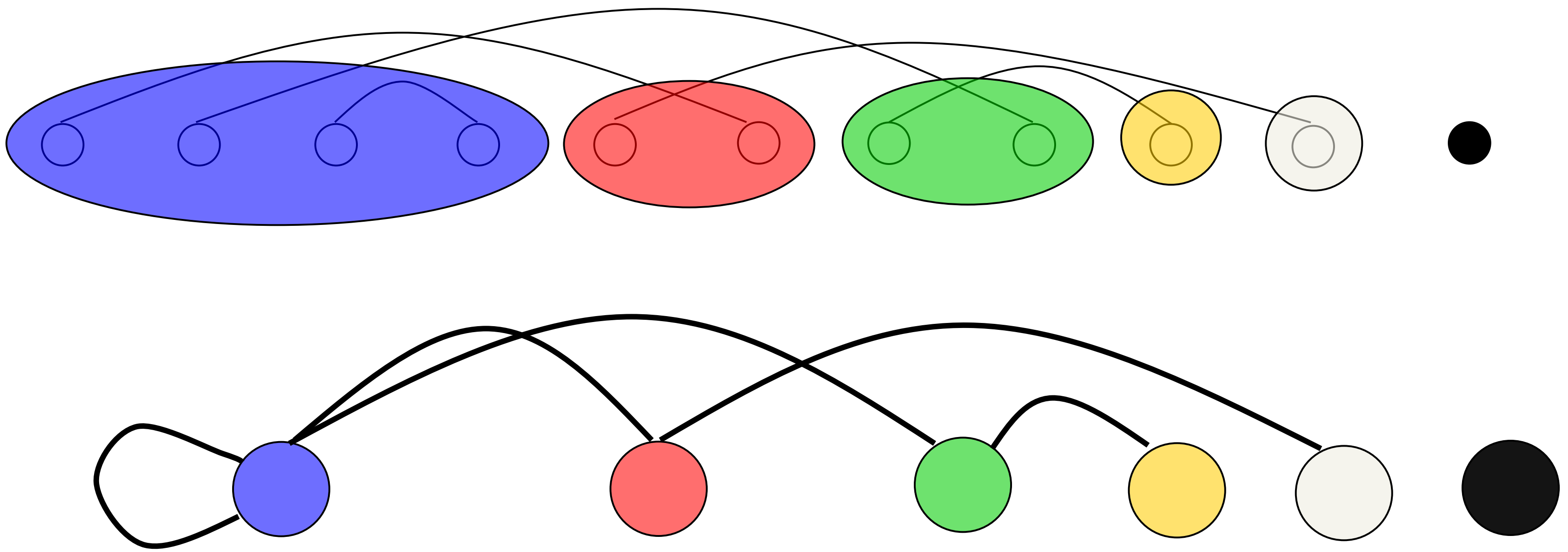
Linearized Chord Diagram

- Create a random matching with 2(n+1) nodes by adding to a matching with 2n nodes a new cord with the right endpoint being in the rightmost position and the left being placed uniformly



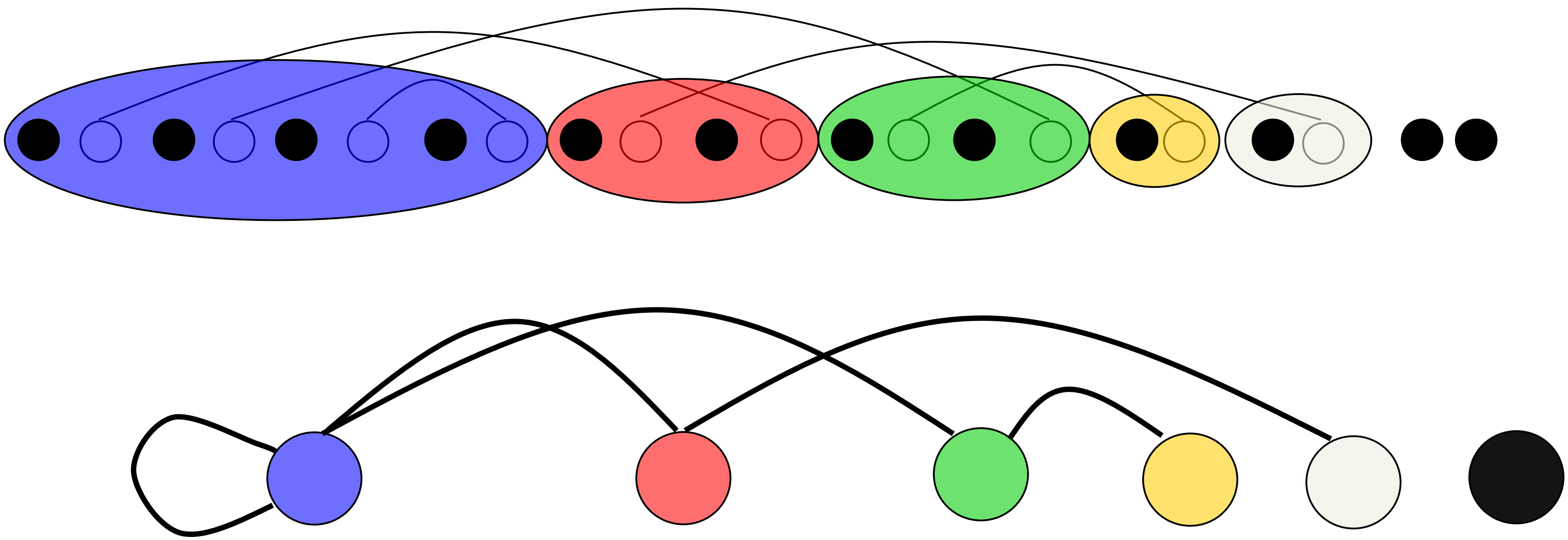
Linearized Chord Diagram

- A new right endpoint creates a new graph node



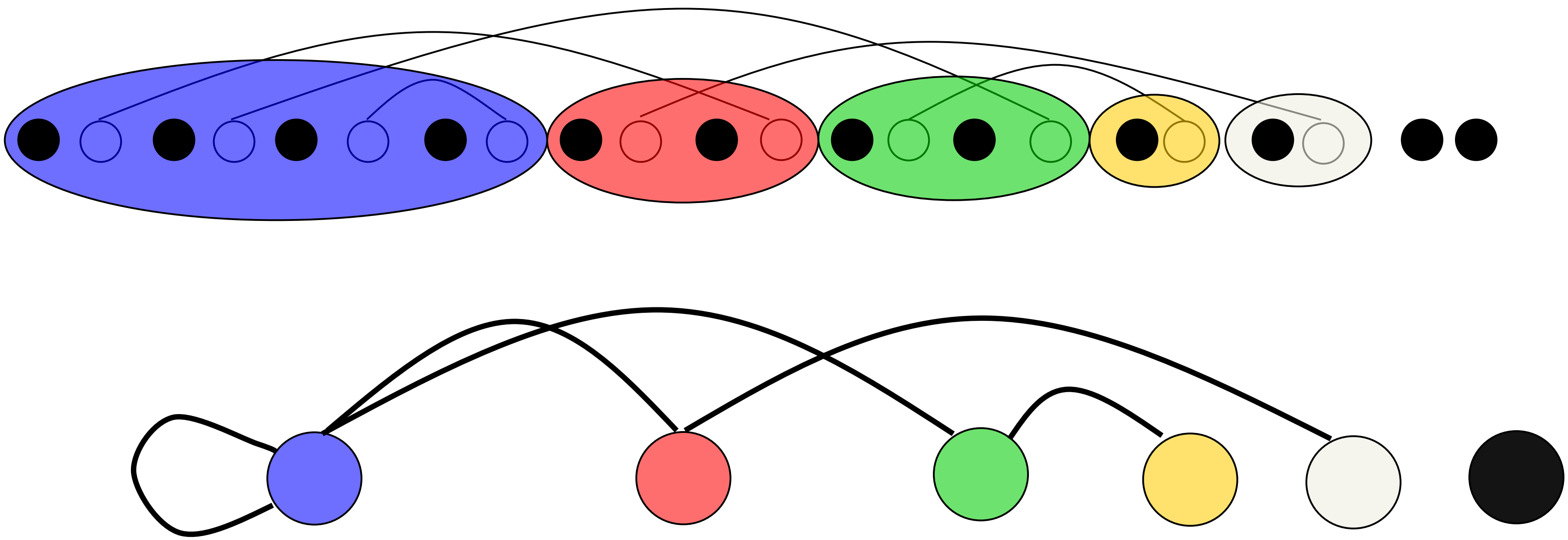
Linearized Chord Diagram

- The left endpoint may be placed within any of the existing "supernodes"



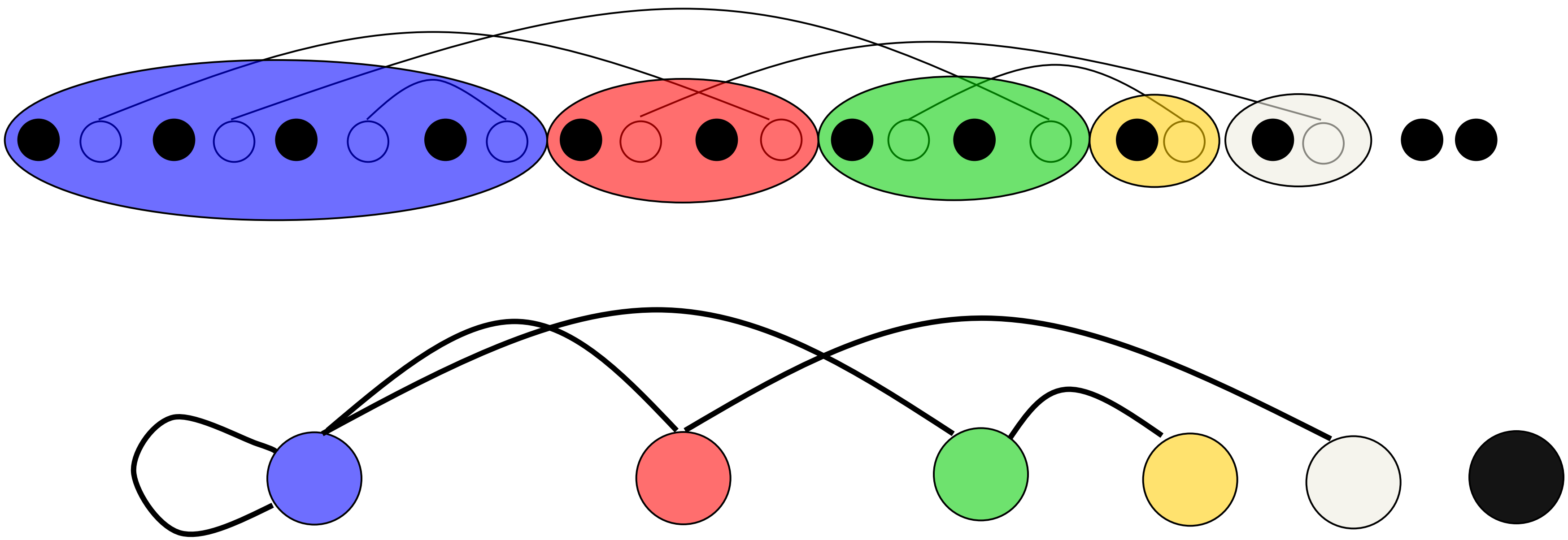
Linearized Chord Diagram

- The number of free positions within a supernode is equal to the number of pairing nodes it contains
- This is also equal to the degree



Linearized Chord Diagram

- For example, the probability that the black graph node links to the blue node is 4/11
  - $d_i = 4$, $t = 6$, $d_i/(2t-1) = 4/11$



Preferential attachment graphs

- Expected diameter
  - if $m = 1$, the diameter is $\Theta(\log n)$
  - if $m > 1$, the diameter is $\Theta(\log n / \log\log n)$
- Expected clustering coefficient

$$E\left[C^{(2)}\right] = \frac{m-1}{8} \frac{\log^2 n}{n}$$

Weaknesses of the BA model

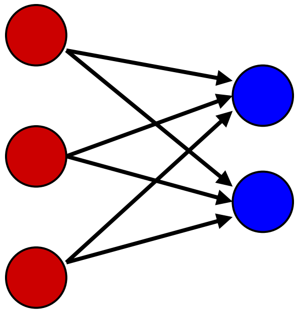
- Technical issues:
  - It is not directed (not good as a model for the Web) and when directed it gives acyclic graphs
  - It focuses mainly on the (in-) degree and does not take into account other parameters (out-degree distribution, components, clustering coefficient)
  - It correlates age with degree which is not always the case
- Academic issues
  - the model rediscovers the wheel
  - preferential attachment is not the answer to every power-law
  - what does “scale-free” mean exactly?
- Yet, it was a breakthrough in the network research, that popularized the area

Variations of the BA model

- Many variations have been considered some in order to address the problems with the vanilla BA model
  - edge rewiring, appearance and disappearance
  - fitness parameters
  - variable mean degree
  - non-linear preferential attachment
    - surprisingly, only linear preferential attachment yields power-law graphs

Empirical observations for the Web graph

- In a large scale experimental study by Kumar et al, they observed that the Web contains a large number of small bipartite cliques (cores)
  - the topical structure of the Web



a $K_{3,2}$ clique

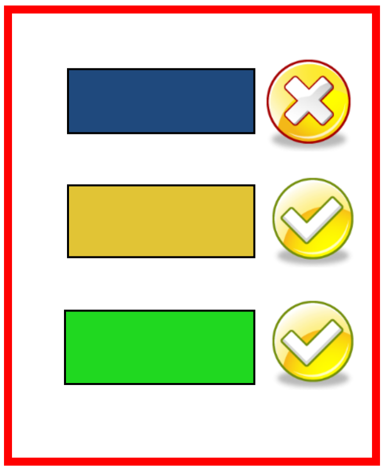
- Such subgraphs are highly unlikely in random graphs
- They are also unlikely in the BA model
- Can we create a model that will have high concentration of small cliques?

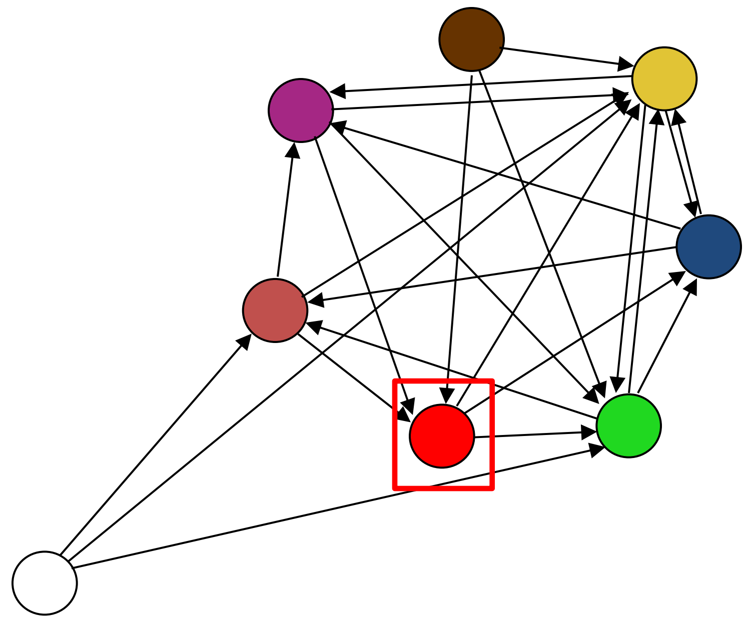
Copying model

- Input:
  - the out-degree d (constant) of each node
  - a parameter α
- The process:
  - Nodes arrive one at the time
  - A new node selects uniformly one of the existing nodes as a prototype
  - The new node creates d outgoing links. For the i^{th} link
    - with probability α it copies the i-th link of the prototype node
    - with probability 1- α it selects the target of the link uniformly at random

An example

- d = 3





Copying model properties

- Power law degree distribution with exponent $\beta = (2-\alpha)/(1-\alpha)$
- Number of bipartite cliques of size $i \times d$ is ne^{-i}

- The model has also found applications in biological networks
  - copying mechanism in gene mutations

Other graph models

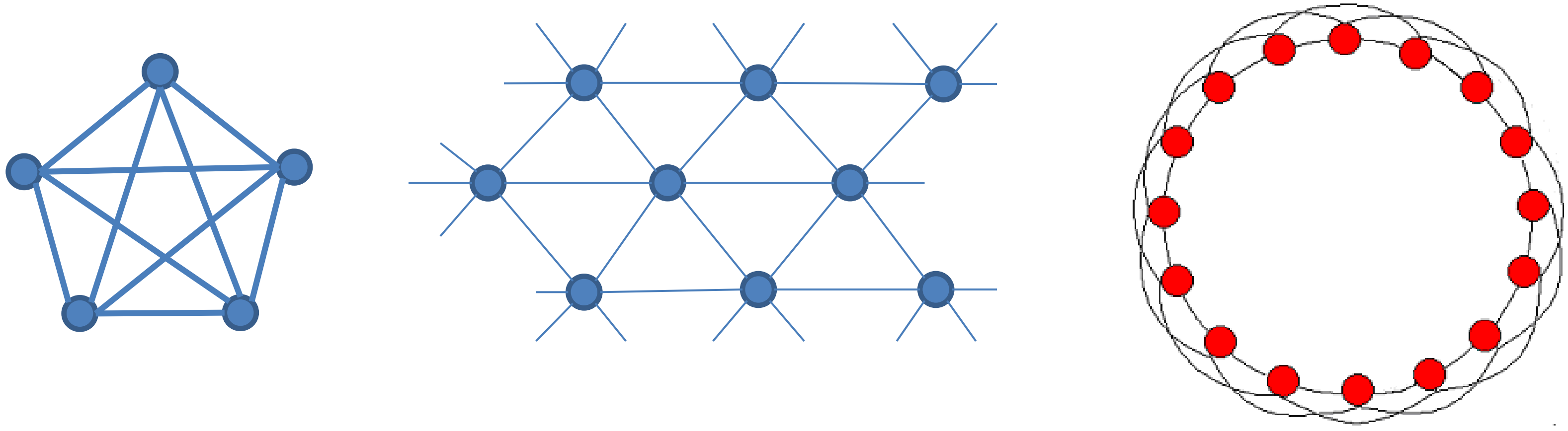
- Cooper Frieze model
  - multiple parameters that allow for adding vertices, edges, preferential attachment, uniform linking
- Directed graphs [Bollobas et al]
  - allow for preferential selection of both the source and the destination
  - allow for edges from both new and old vertices

Small world Phenomena

- So far we focused on obtaining graphs with power-law distributions on the degrees. What about other properties?
  - Clustering coefficient: real-life networks tend to have high clustering coefficient
  - Short paths: real-life networks are "small worlds"
    - this property is easy to generate
  - Can we combine these two properties?

Clustering Coefficient

- How can you create a graph with high clustering coefficient?



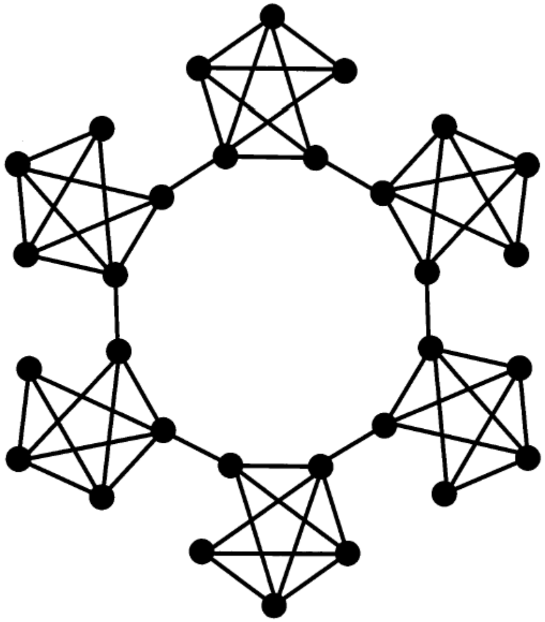
- High clustering coefficient but long paths

Small-world Graphs

- According to Watts [W99]
  - Large networks ($n >> 1$)
  - Sparse connectivity (avg degree $z << n$)
  - No central node ($k_{max} << n$)
  - Large clustering coefficient (larger than in random graphs of same size)
  - Short average paths (~$\log n$, close to those of random graphs of the same size)

The Caveman Model [W99]

- The random graph
  - edges are generated completely at random
  - low avg. path length $L \leq \log n / \log z$
  - low clustering coefficient $C \sim z/n$
- The Caveman model
  - edges follow a structure
  - high avg. path length $L \sim n/z$
  - high clustering coefficient $C \sim 1\text{-}O(1/z)$



- Can we interpolate between the two?

Mixing order with randomness

- Inspired by the work of Solmonoff and Rapoport
  - nodes that share neighbors should have higher probability to be connected
- Generate an edge between i and j with probability proportional to R_{ij}

$$R_{ij} = \begin{cases} 1 & \text{if } m_{ij} \geq z \\ \left(\frac{m_{ij}}{z}\right)^{\alpha}(1-p)+p & \text{if } 0 < m_{ij} < z \\ p & \text{if } m_{ij} = 0 \end{cases}$$

m_{ij} = number of common neighbors of i and j

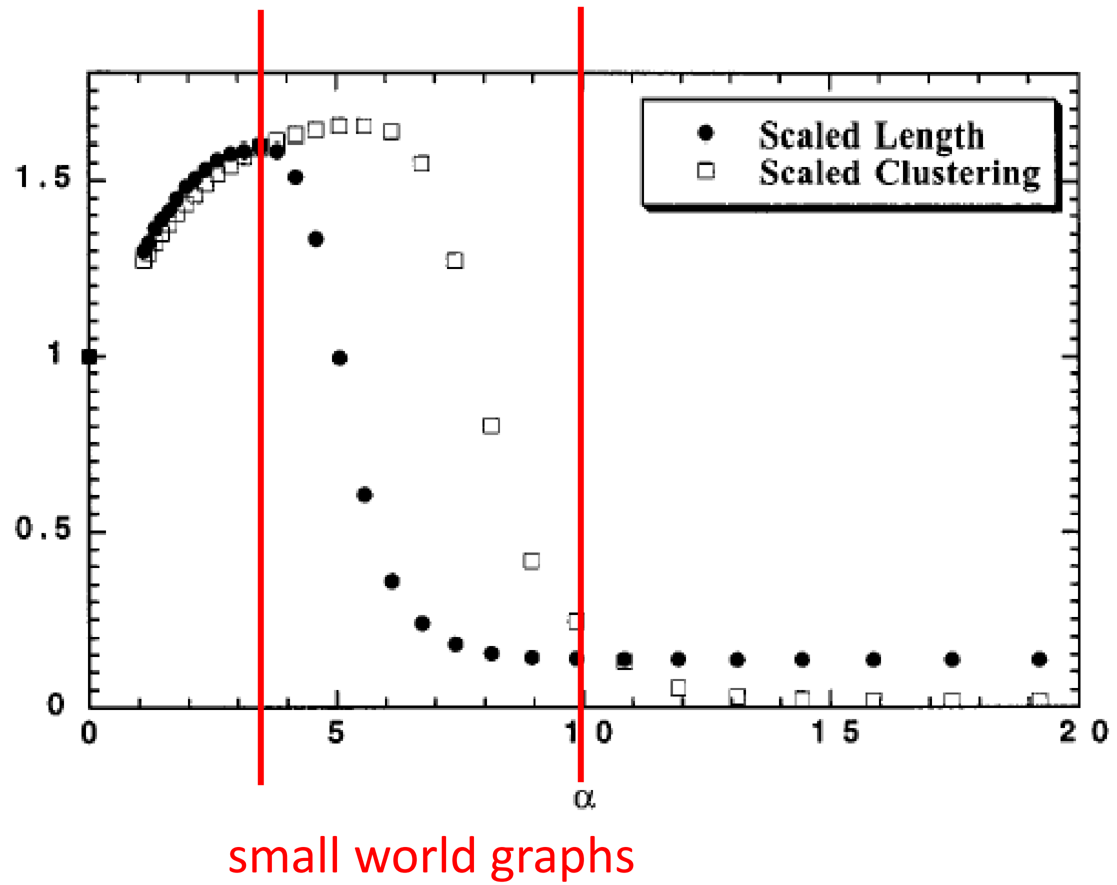
p = very small probability

- When $\alpha \to \infty$, edges are determined by common neighbors
- When $\alpha = 0$, edges are independent of common neighbors
- For intermediate values we obtain a combination of order and randomness

Algorithm

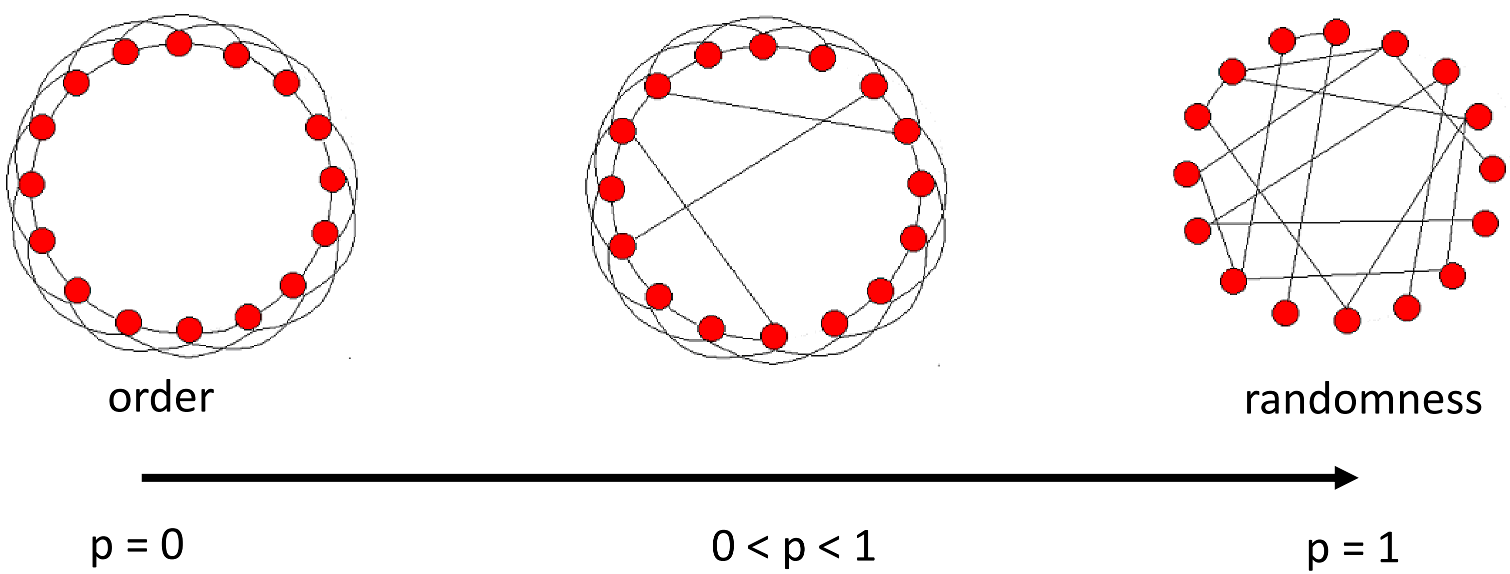
- Start with a ring
- For $i = 1 \dots n$
  - Select a vertex j with probability proportional to R_{ij} and generate an edge (i,j)
- Repeat until z edges are added to each vertex

Clustering coefficient – Avg path length

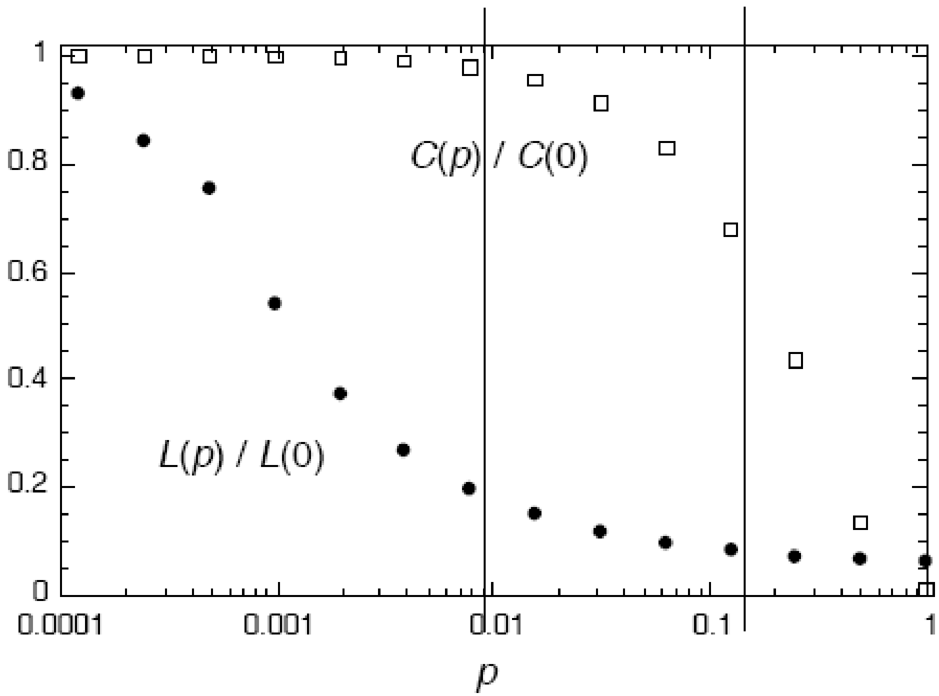


Watts and Strogatz model [WS98]

- Start with a ring, where every node is connected to the next z nodes
- With probability p, rewire every edge (or, add a shortcut) to a uniformly chosen destination.
  - Granovetter, “The strength of weak ties”



Clustering Coefficient – Characteristic Path Length



log-scale in p

When p = 0, C = 3(k-2)/4(k-1) ~ ¾
L = n/k

For small p, C ~ ¾
L ~ logn

Graph Theory Results

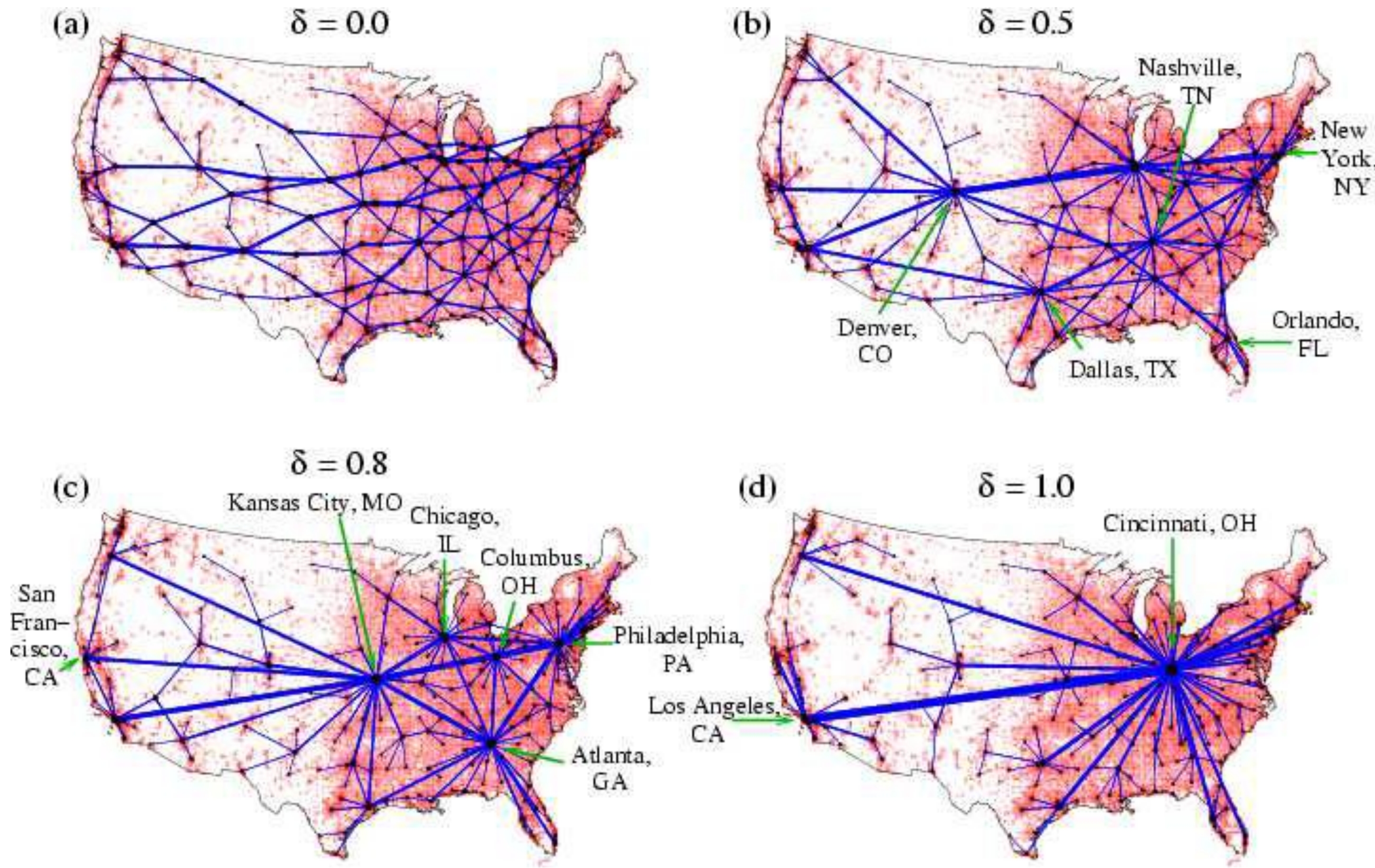
- Graph theorist failed to be impressed. Most of these results were known.

Optimized graphs

- Suppose you are building an airline network, how would you set up the routes?
- Optimization criteria
  - Minimize the cost of routes
  - Minimize the travel time of passengers
    - Distance travelled
    - Number of hops
  - Take city populations into account.

Use δ to control the tradeoff between the two

Experiment with US flights



Evolution of graphs

- So far we looked at the properties of graph snapshots. What if we have the history of a graph?
  - e.g., citation networks, internet graphs

Measuring preferential attachment

- Is it the case that the rich get richer?

- Look at the network for an interval [t,t+dt]
- For node i, present at time t, we compute

$$D_i = \frac{dk_i}{dk}$$

  - dk_i = increase in the degree
  - dk = number of edges added
- Fraction of edges added to nodes of degree k

$$f(k) = \sum_{i:k_i=k} D_i$$

- Cumulative: fraction of edges added to nodes of degree at most k

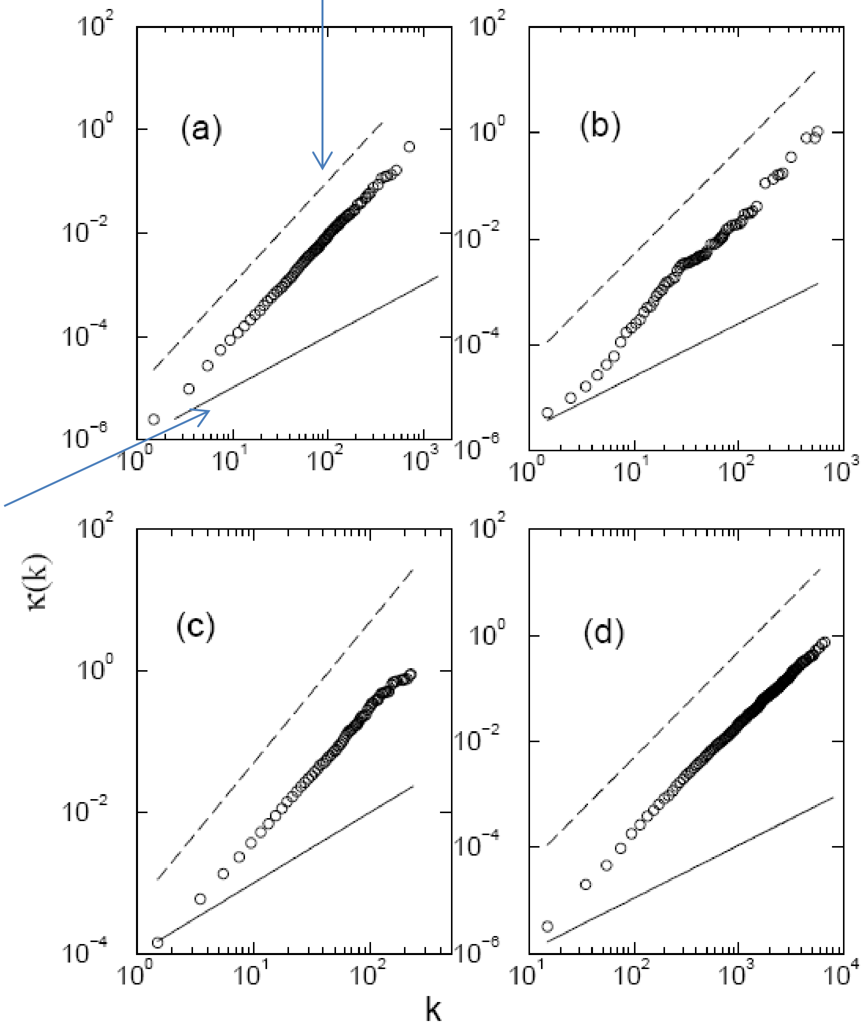
$$F(k) = \sum_{j=1}^{k} f(j)$$

Measuring preferential attachment

- We plot F(k) as a function of k. If preferential attachment exists we expect that $F(k) \sim k^b$
  - actually, it has to be $b \sim 1$

Linear preferential attachment

No preferential attachment



(a) citation network
(b) Internet
(c) scientific collaboration network
(d) actor collaboration network

Network models and temporal evolution

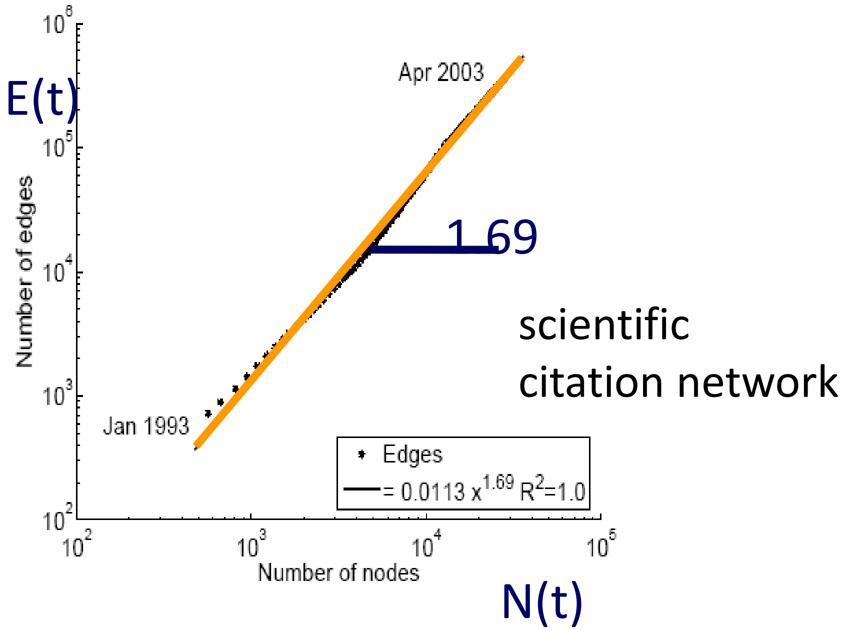
- For most of the existing models it is assumed that
  - number of edges grows linearly with the number of nodes
  - the diameter grows at rate logn, or loglogn
- What about real graphs?
  - Leskovec, Kleinberg, Faloutsos 2005

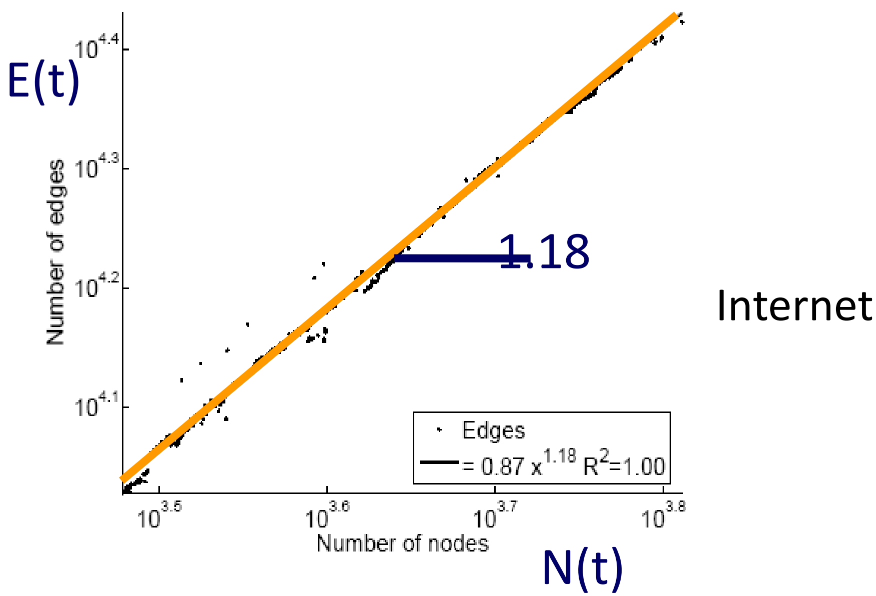
Densification laws

- In real-life networks the average degree increases! – networks become denser!

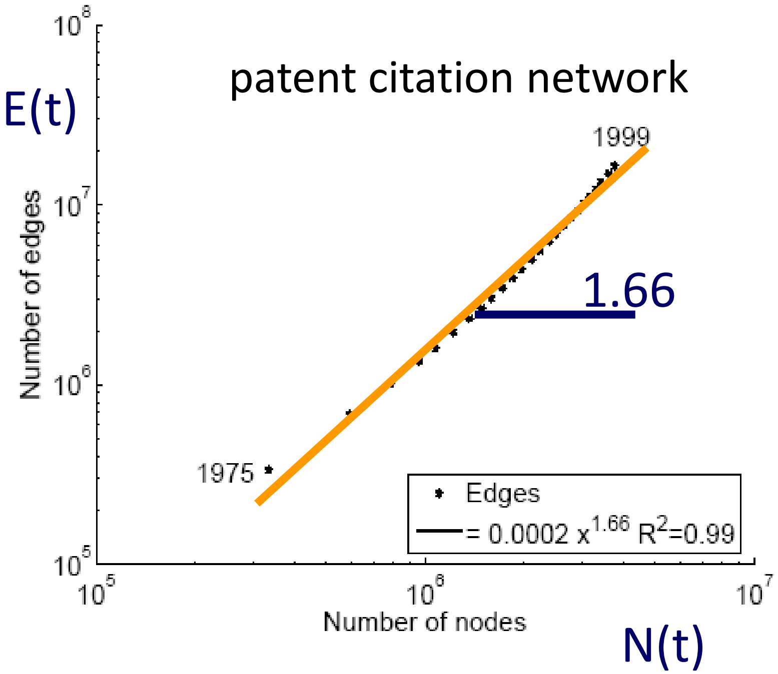
$$\boxed{E(t) \propto N(t)^a}$$

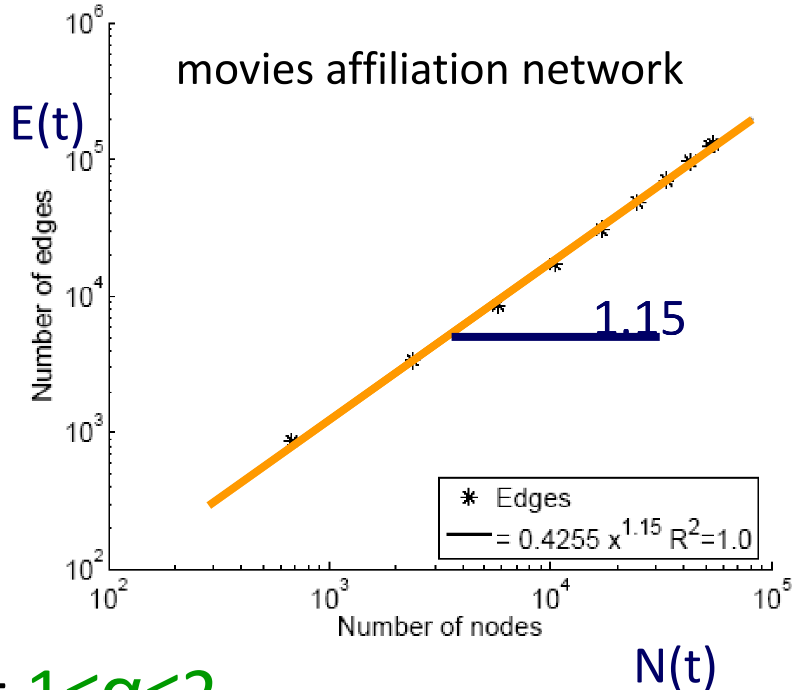
α = densification exponent





More examples

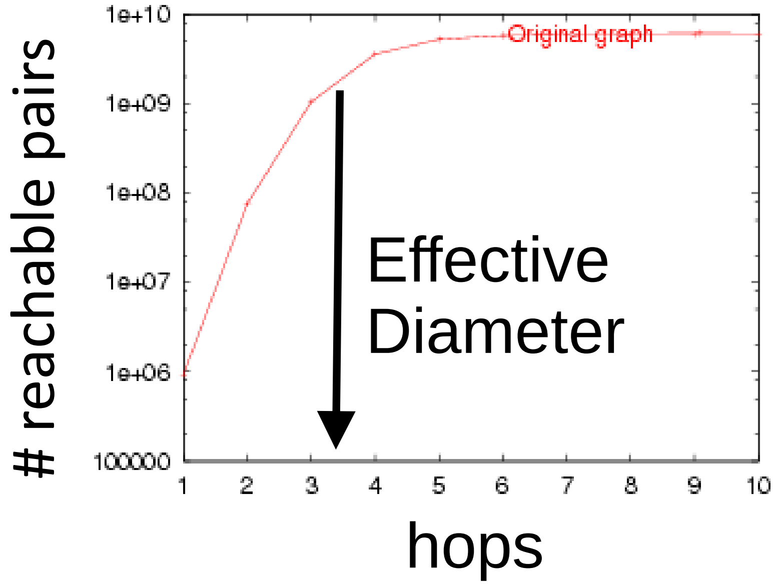




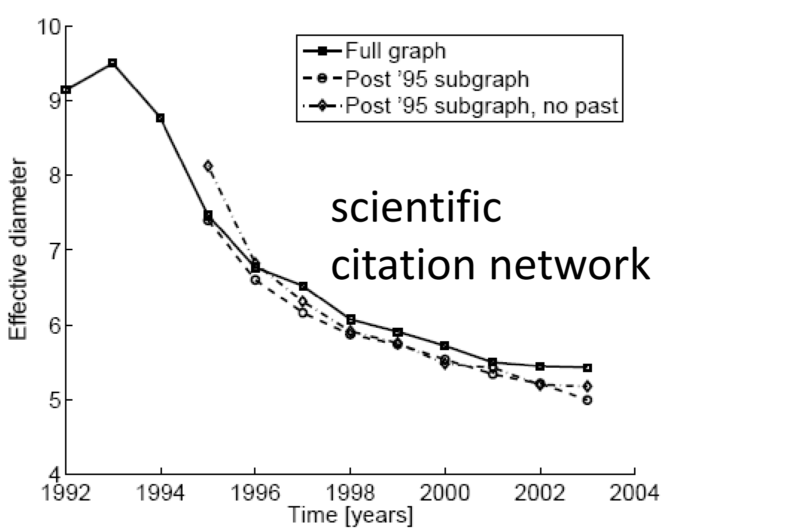
- The densification exponent $1 \le \alpha \le 2$
  - $\alpha = 1$: linear growth – constant out degree
  - $\alpha = 2$: quadratic growth - clique

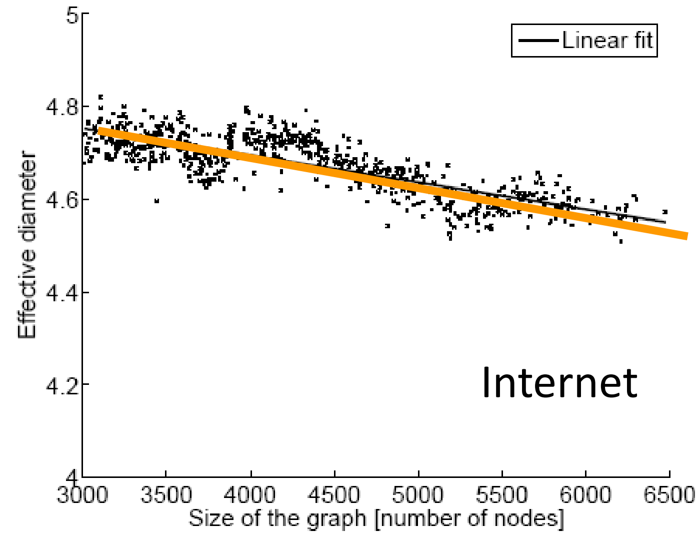
What about diameter?

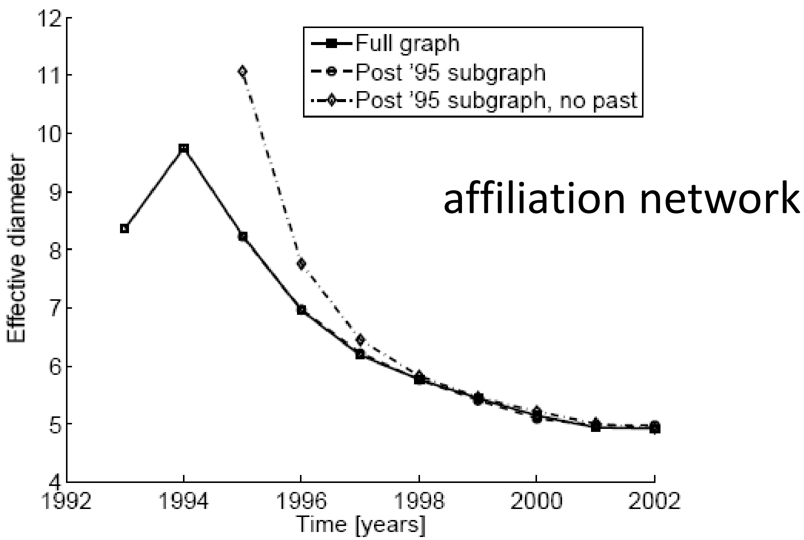
- Effective diameter: the interpolated value where 90% of node pairs are reachable

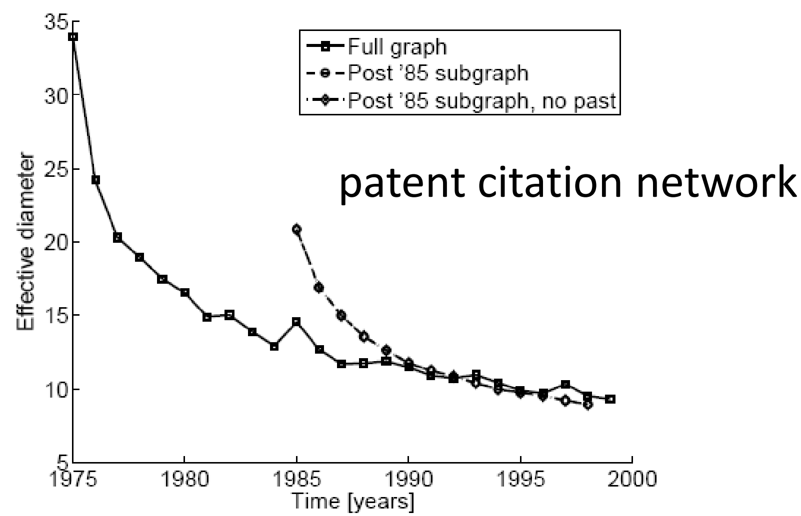


Diameter shrinks









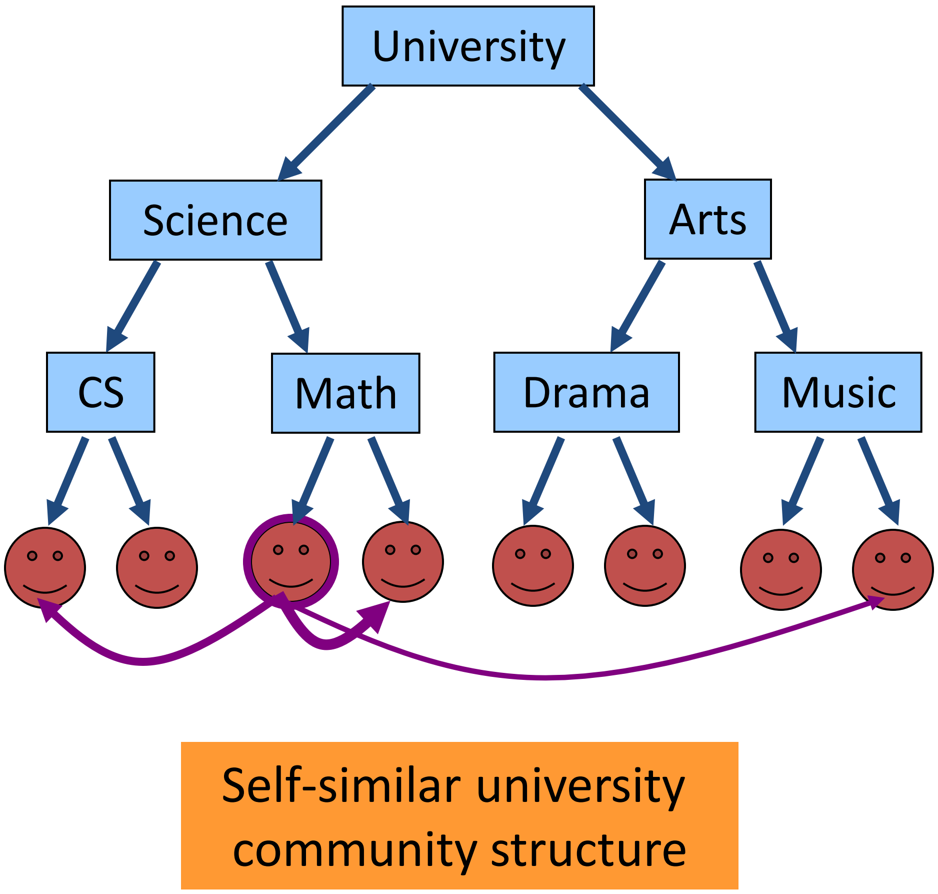
Densification – Possible Explanation

- Existing graph generation models do not capture the Densification Power Law and Shrinking diameters
- Can we find a simple model of local behavior, which naturally leads to observed phenomena?

- Two proposed models
  - Community Guided Attachment – obeys Densification
  - Forest Fire model – obeys Densification, Shrinking diameter (and Power Law degree distribution)

Community structure

- Let's assume the community structure
- One expects many within-group friendships and fewer cross-group ones
- How hard is it to cross communities?



Fundamental Assumption

- If the cross-community linking probability of nodes at tree-distance h is scale-free
- We propose cross-community linking probability:

$$\boxed{f(h) = c^{-h}}$$

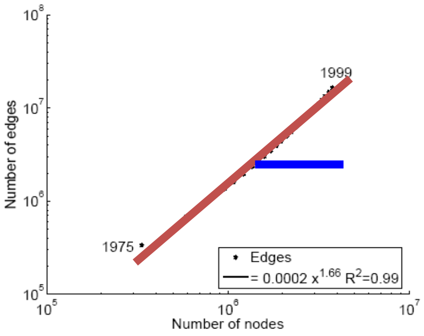
where: $c \geq 1$ … the Difficulty constant

h … tree-distance

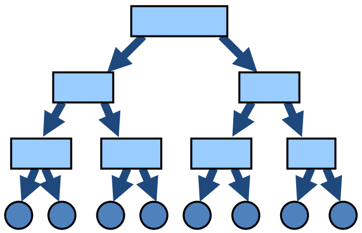
Densification Power Law

- Theorem: The Community Guided Attachment leads to Densification Power Law with exponent

$$\boxed{a = 2 - \log_b(c)}$$



- α … densification exponent $E(t) \propto N(t)^a$
- b … community structure branching factor
- c … difficulty constant



Difficulty Constant

- Theorem: $$\boxed{a = 2 - \log_b(c)}$$
- Gives any non-integer Densification exponent
- If c = 1: easy to cross communities
  - Then: α = 2, quadratic growth of edges – near clique
- If c = b: hard to cross communities
  - Then: α = 1, linear growth of edges – constant out-degree

Room for Improvement

- Community Guided Attachment explains Densification Power Law
- Issues:
  - Requires explicit Community structure
  - Does not obey Shrinking Diameters
- The ”Forrest Fire” model

“Forest Fire” model – Wish List

- We want:
  - no explicit Community structure
  - Shrinking diameters
  - and:
    - “Rich get richer” attachment process, to get heavy-tailed in-degrees
    - “Copying” model, to lead to communities
    - Community Guided Attachment, to produce Densification Power Law

“Forest Fire” model – Intuition

- How do authors identify references?
  1. Find first paper and cite it
  2. Follow a few citations, make citations
  3. Continue recursively
  4. From time to time use bibliographic tools (e.g. CiteSeer) and chase back-links

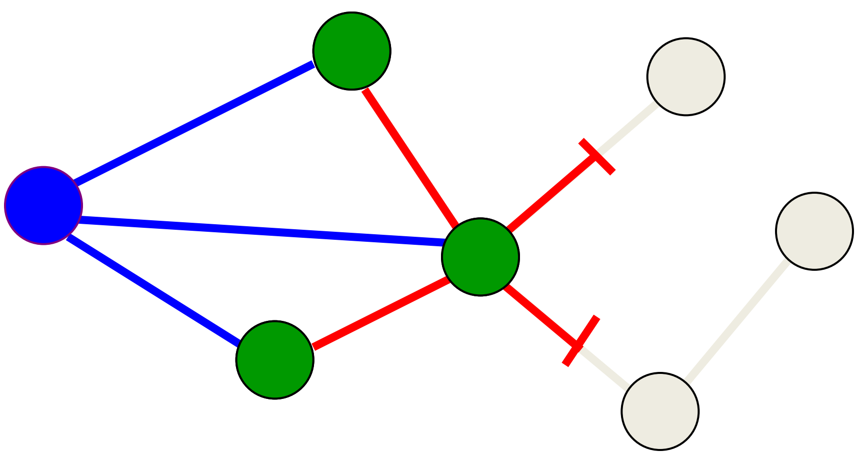
“Forest Fire” model – Intuition

- How do people make friends in a new environment?
  1. Find first a person and make friends
  2. From time to time get introduced to his friends
  3. Continue recursively

- Forest Fire model imitates exactly this process

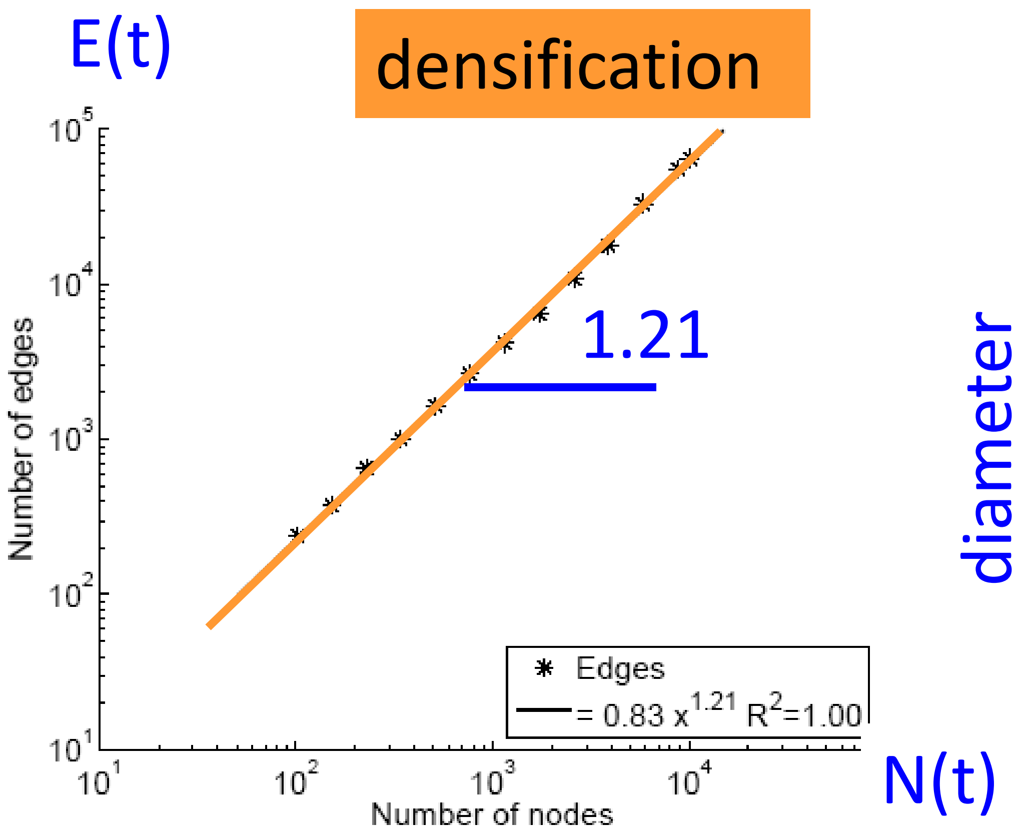
“Forest Fire” – the Model

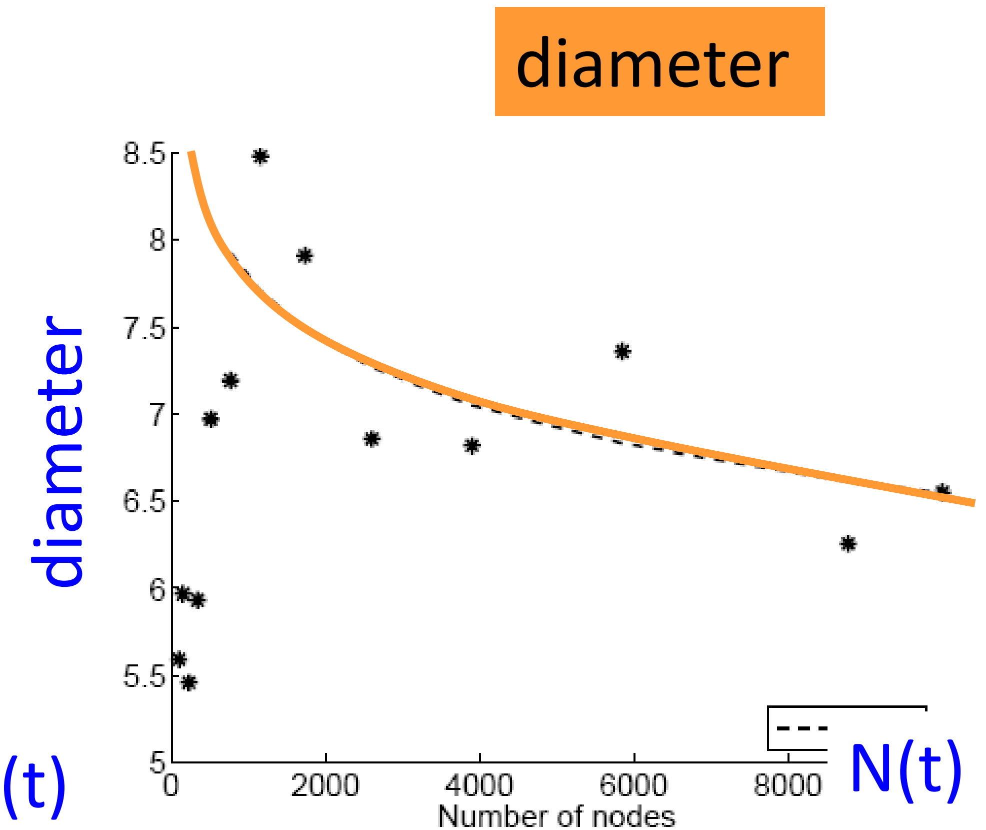
- A node arrives
- Randomly chooses an “ambassador”
- Starts burning nodes (with probability p) and adds links to burned nodes
- “Fire” spreads recursively



Forest Fire in Action (1)

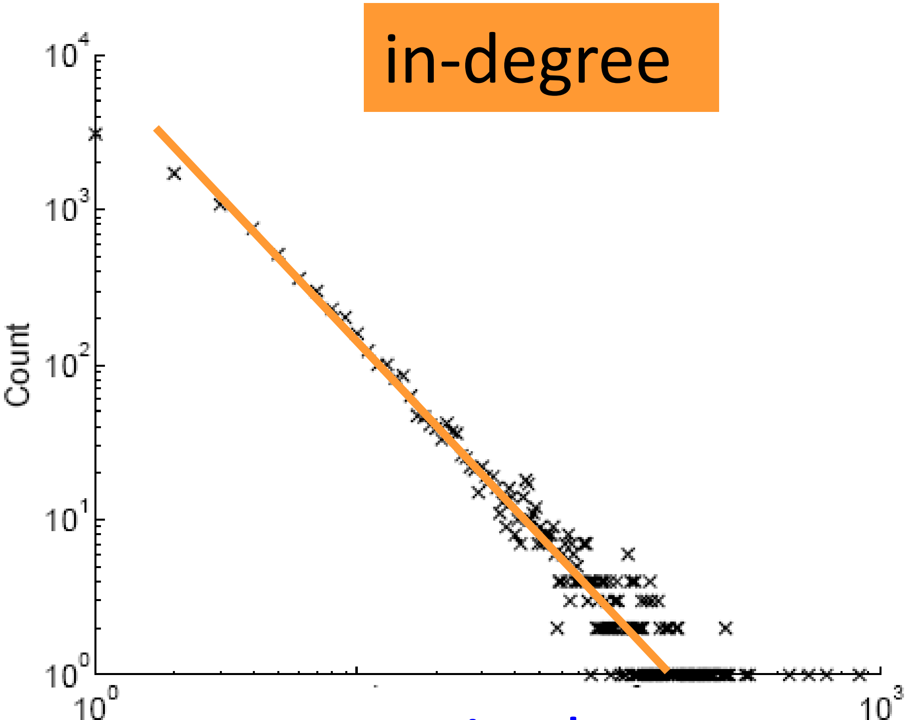
- Forest Fire generates graphs that Densify and have Shrinking Diameter



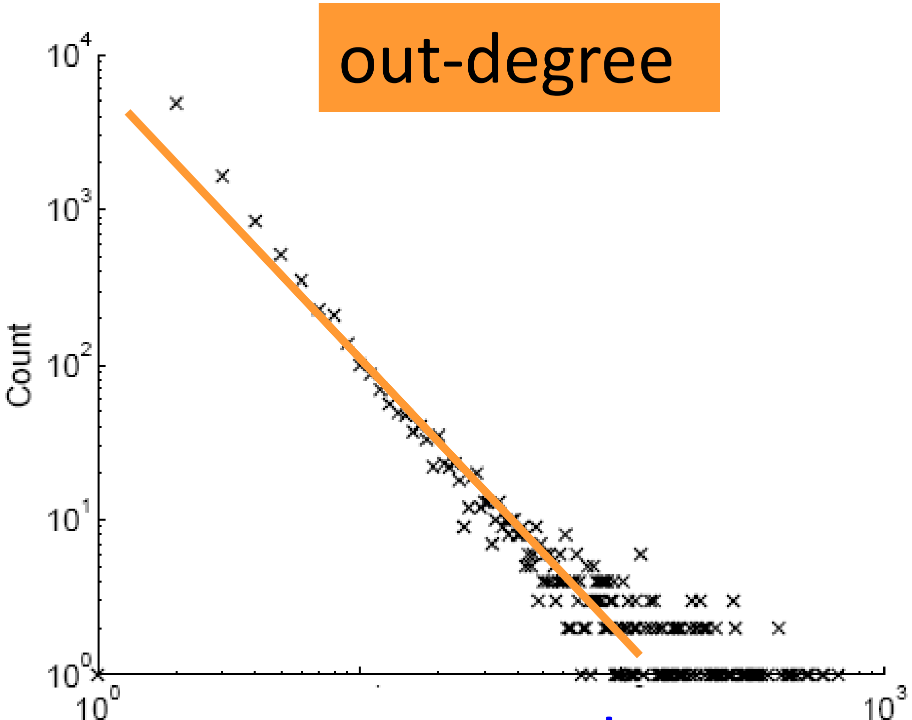


Forest Fire in Action (2)

- Forest Fire also generates graphs with heavy-tailed degree distribution



count vs. in-degree



count vs. out-degree

Forest Fire model – Justification

- Densification Power Law:
  - Similar to Community Guided Attachment
  - The probability of linking decays exponentially with the distance – Densification Power Law
- Power law out-degrees:
  - From time to time we get large fires
- Power law in-degrees:
  - The fire is more likely to reach hubs

Forest Fire model – Justification

- Communities:
  - Newcomer copies neighbors’ links
- Shrinking diameter

Acknowledgements

- Many thanks to Jure Leskovec for his slides from the KDD 2005 paper.

References

- M. E. J. Newman, The structure and function of complex networks, SIAM Reviews, 45(2): 167-256, 2003
- R. Albert and L.A. Barabasi, Statistical Mechanics of Complex Networks, Rev. Mod. Phys. 74, 47-97 (2002).
- B. Bollobas, Mathematical Results in Scale-Free random Graphs
- D.J. Watts. Networks, Dynamics and Small-World Phenomenon, American Journal of Sociology, Vol. 105, Number 2, 493-527, 1999
- Watts, D. J. and S. H. Strogatz. Collective dynamics of 'small-world' networks. Nature 393:440-42, 1998
- Michael T. Gastner and M. E. J. Newman, Optimal design of spatial distribution networks, *Phys. Rev. E* **74**, 016117 (2006).J.
- Leskovec, J. Kleinberg, C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2005.

Assignment

- In teams of 2
- Pick a scale free or a small world model
  - Scale free: Preferential Attachment, Copying model
  - Small world: Caveman model, ring rewiring
- Create different networks for different parameters
- Use Gephi to visualize the graphs, plot degree distributions, and compute clustering coefficient