What did we see in the last lecture?

What are we going to talk about today?

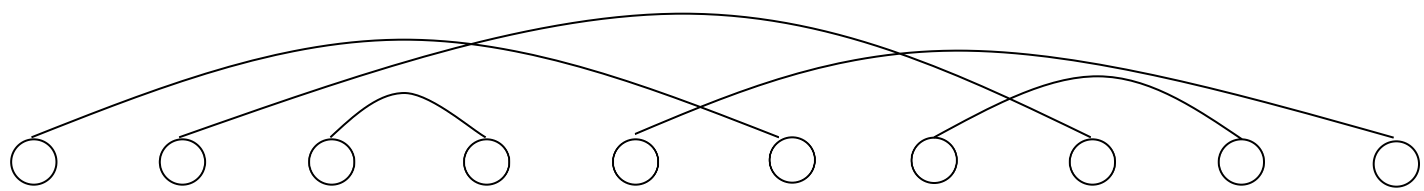
- Generative models for graphs with power-law degree distribution
- Generative models for graphs with small-world properties
- Models that capture graph evolution over time

The Linearized Chord Diagram (LCD) model

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

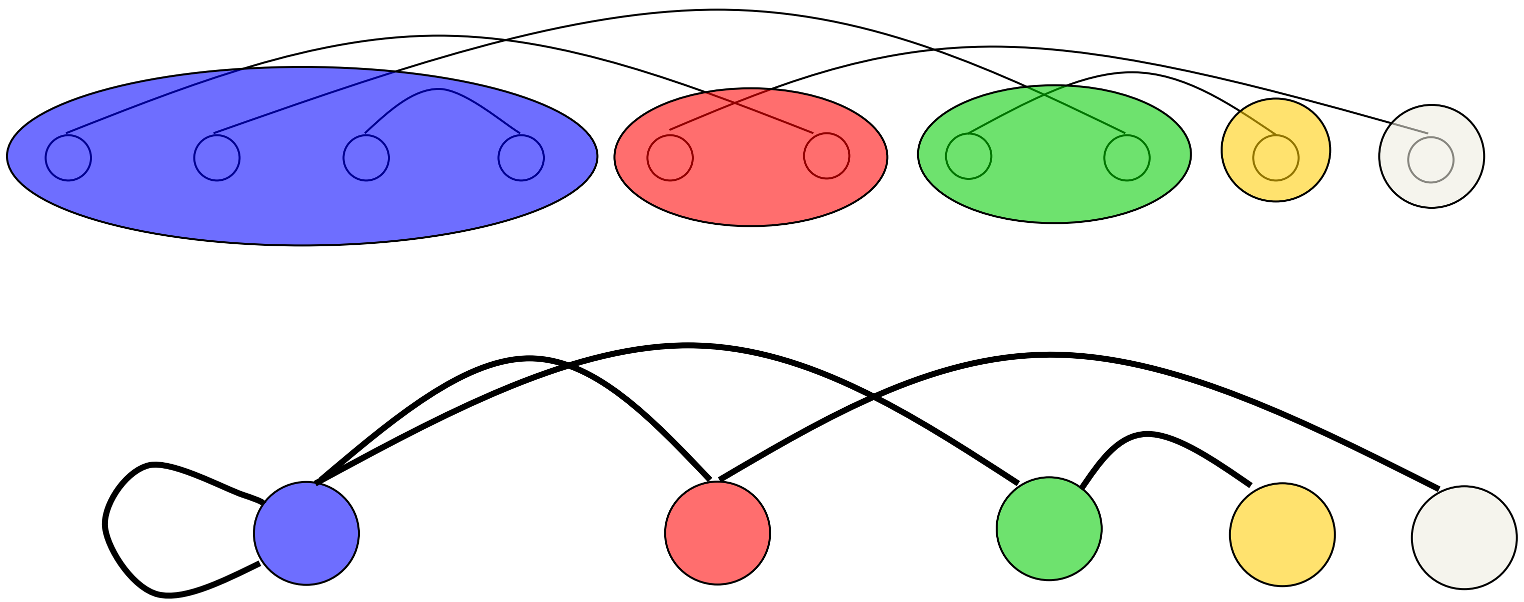
Linearized Chord Diagram

- Generate a random matching of the nodes.



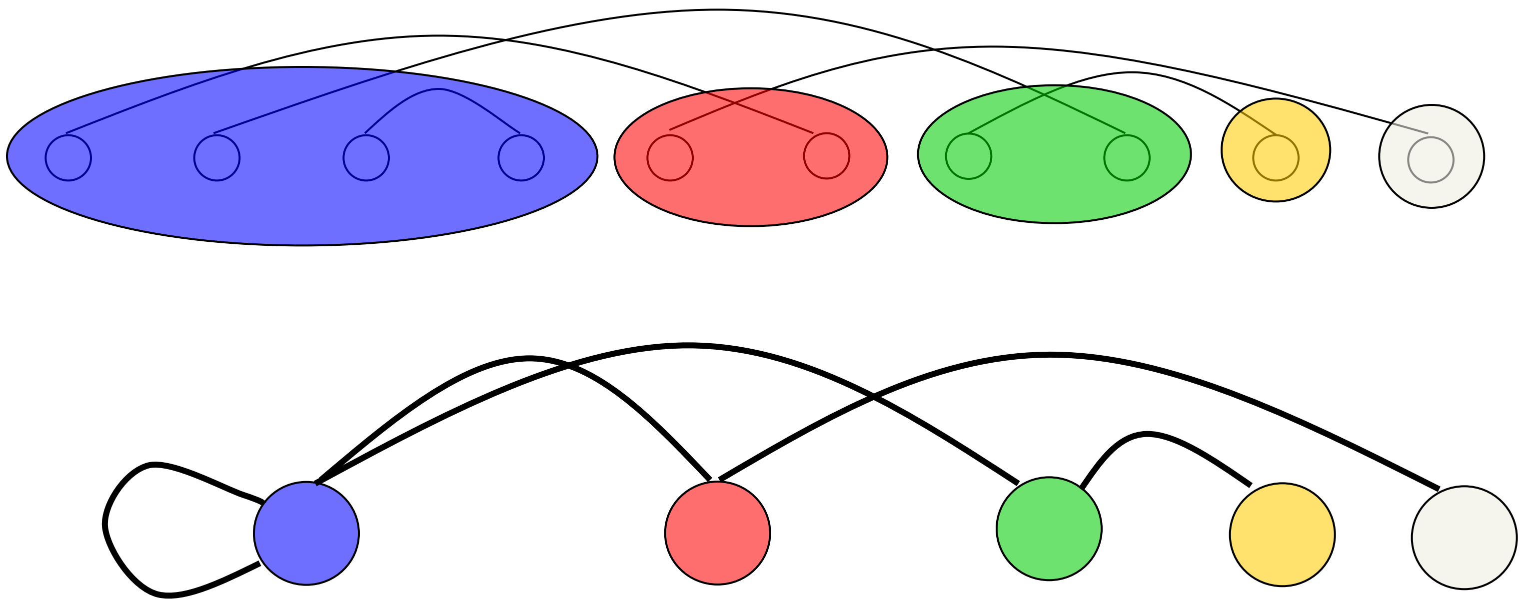
Linearized Chord Diagram

- 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.



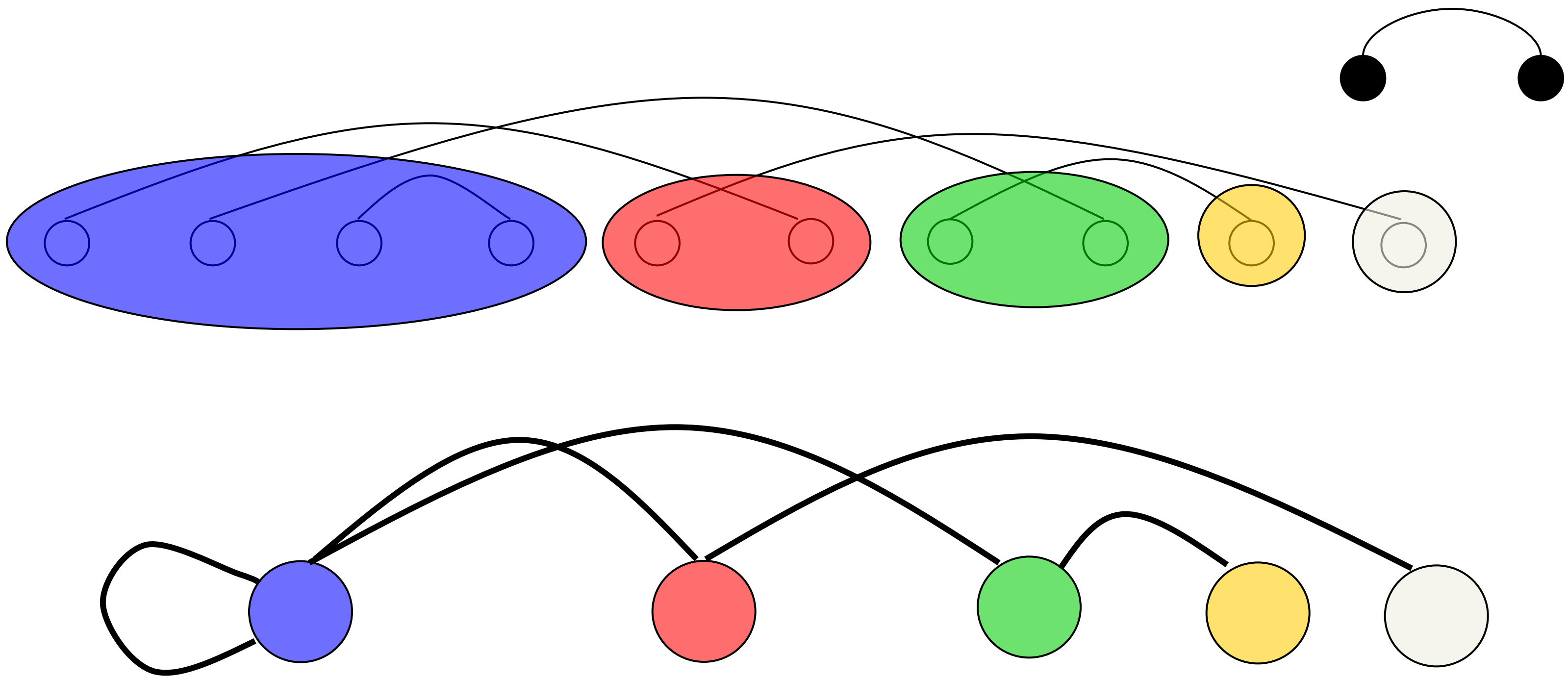
Linearized Chord Diagram

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



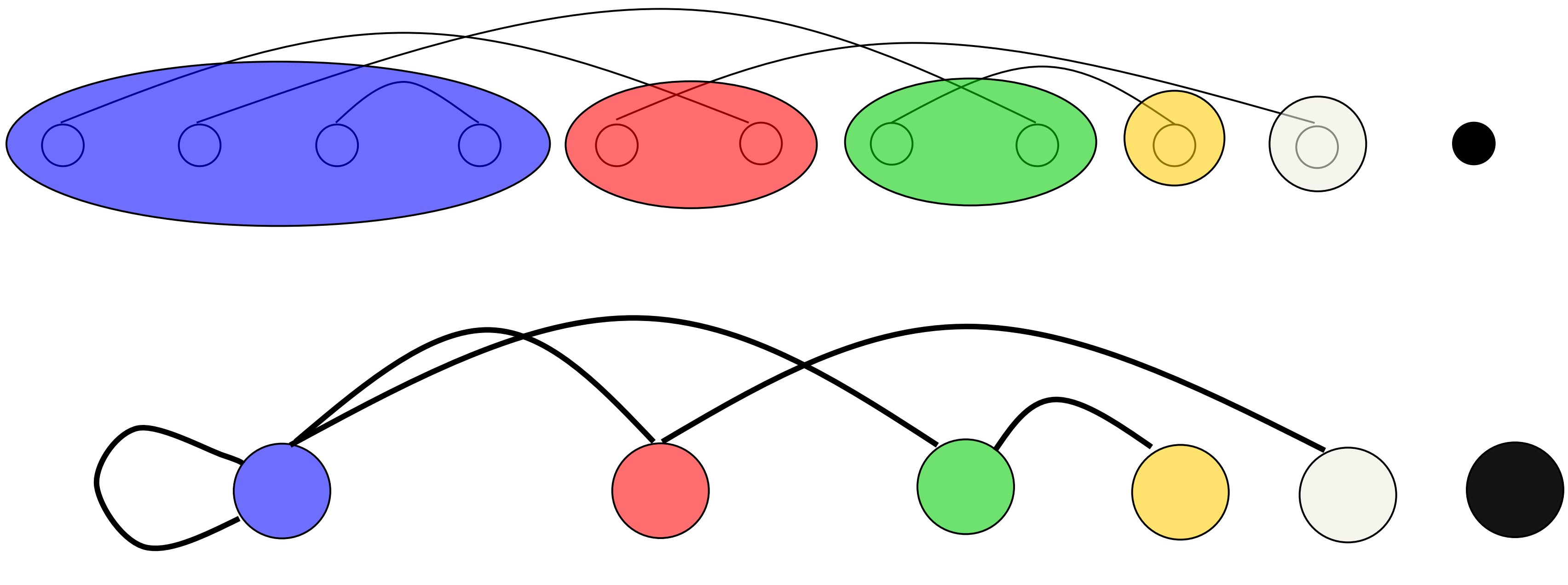
Linearized Chord Diagram

- 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



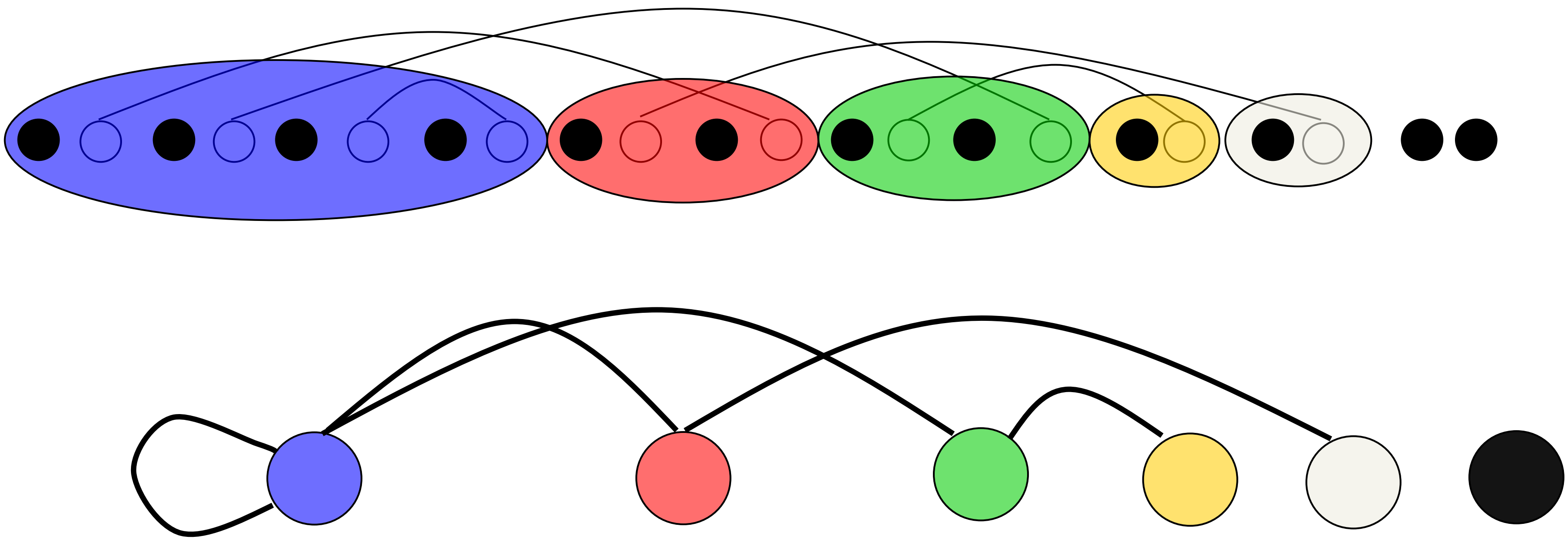
Linearized Chord Diagram

- A new right endpoint creates a new graph node



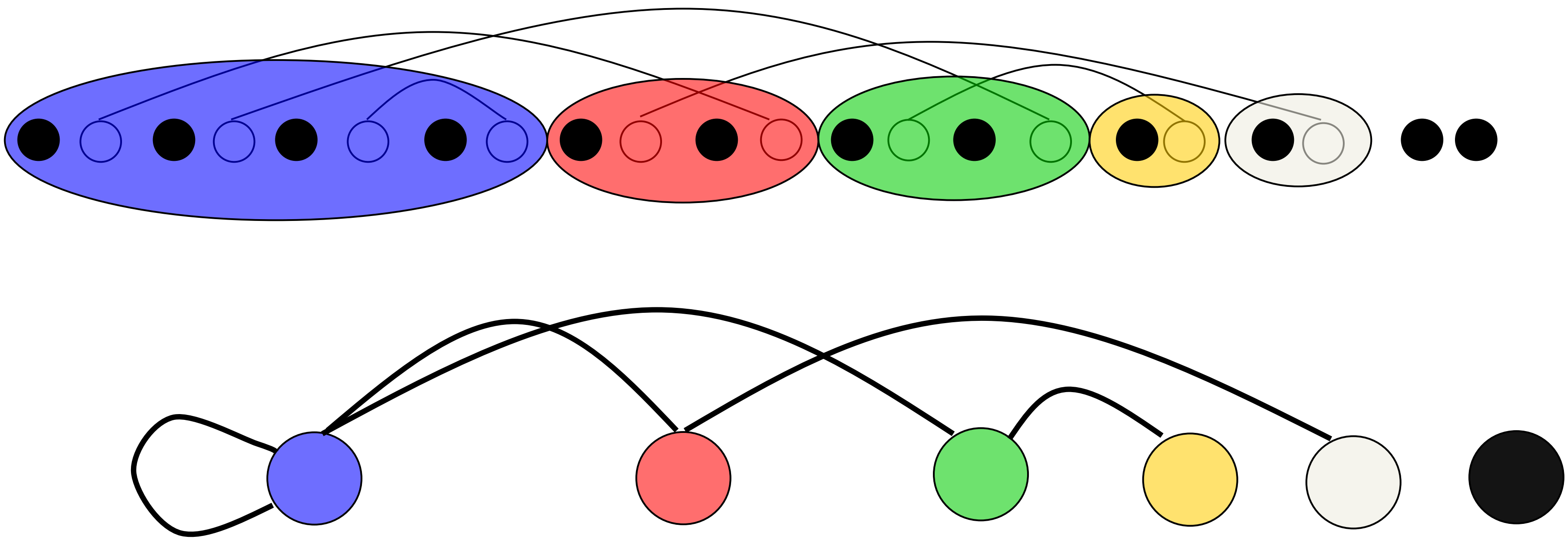
Linearized Chord Diagram

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



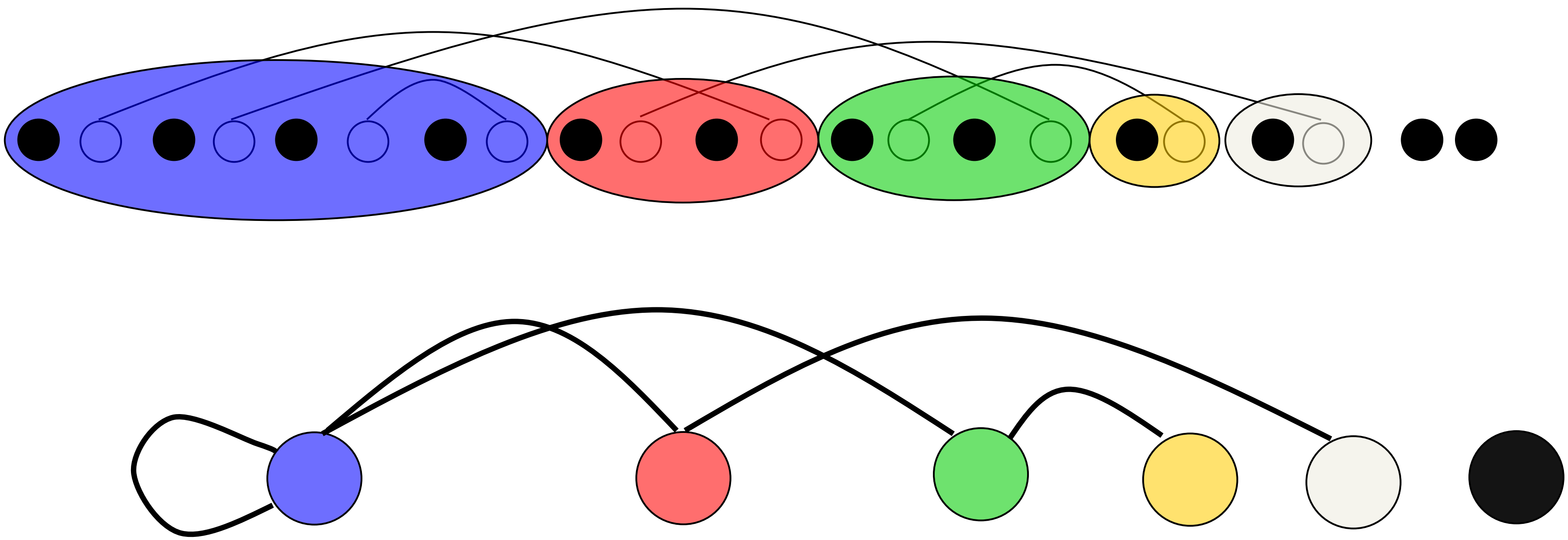
Linearized Chord Diagram

- 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
  - $d_i = 4$, $t = 6$, $d_i/(2t-1) = 4/11$

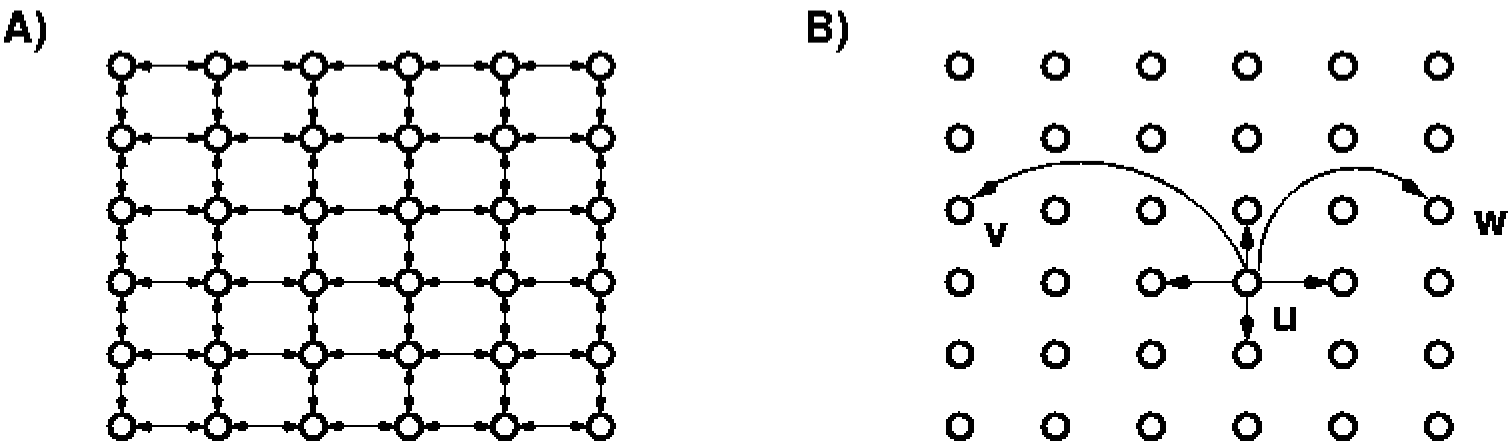


Milgram's experiment revisited

- What did Milgram's experiment show?
  - (a) There are short paths in large networks that connect individuals
  - (b) People are able to find these short paths using a simple, greedy, decentralized algorithm
- Small world models take care of (a)
- Kleinberg: what about (b)?

Kleinberg’s model

- Consider a directed 2-dimensional lattice
- For each vertex u add q shortcuts
  - choose vertex v as the destination of the shortcut with probability proportional to $[d(u,v)]^{-r}$
  - when $r = 0$, we have uniform probabilities



The decentralized search algorithm

- Given a source s and a destination t, the search algorithm
  1. knows the positions of the nodes on the grid (geography information)
  2. knows the neighbors and shortcuts of the current node (local information)
  3. operates greedily, each time moving as close to t as possible (greedy operation)
  4. knows the neighbors and shortcuts of all nodes seen so far (history information)

Kleinberg results

- The search algorithm
  1. knows the positions of the nodes on the grid (geography information)
  2. knows the neighbors and shortcuts of the current node (local information)
  3. operates greedily, each time moving as close to t as possible (greedy operation)
  4. knows the neighbors and shortcuts of all nodes seen so far (history information)

- When $r=2$, an algorithm that uses only local information at each node (not 4) can reach the destination in expected time $O(\log^2 n)$.

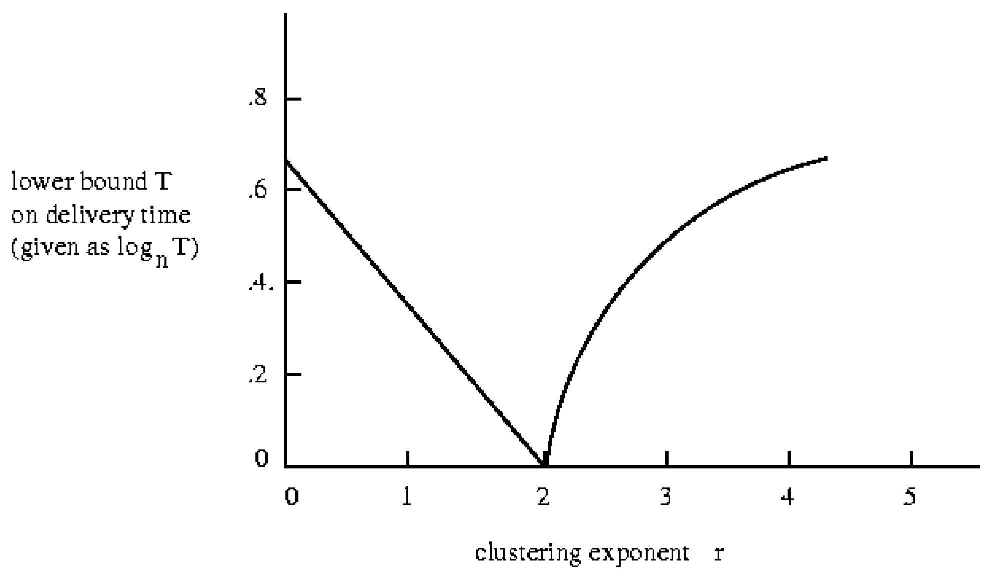
Kleinberg's results

- The search algorithm
  1. knows the positions of the nodes on the grid (geography information)
  2. knows the neighbors and shortcuts of the current node (local information)
  3. operates greedily, each time moving as close to t as possible (greedy operation)
  4. knows the neighbors and shortcuts of all nodes seen so far (history information)

- When $r<2$ a local greedy algorithm (1-4) needs expected time $\Omega(n^{(2-r)/3})$.
- When $r>2$ a local greedy algorithm (1-4) needs expected time $\Omega(n^{(r-2)/(r-1)})$.

Searching in a small world

- For r < 2, the graph has paths of logarithmic length (small world), but a greedy algorithm cannot find them
- For r > 2, the graph does not have short paths
- For r = 2 is the only case where there are short paths, and the greedy algorithm is able to find them



Generalization

- When $r=2$, an algorithm that uses only local information at each node (not 4) can reach the destination in expected time $O(\log^2 n)$.
- When $r<2$ a local greedy algorithm (1-4) needs expected time $\Omega(n^{(2-r)/3})$.
- When $r>2$ a local greedy algorithm (1-4) needs expected time $\Omega(n^{(r-2)/(r-1)})$.

- The results generalize for a d-dimensional grid. The algorithm works in expected $O(\log^2 n)$ time, when $r=d$

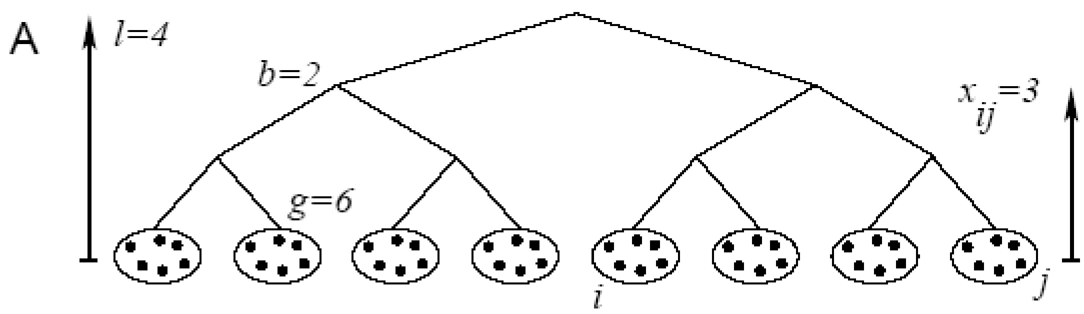
Extensions

- If there are $\log n$ shortcuts, then the search time is $O(\log n)$
  – we save the time required for finding the shortcut

- If we know the shortcuts of $\log n$ neighbors the time becomes $O(\log^{1+1/d} n)$

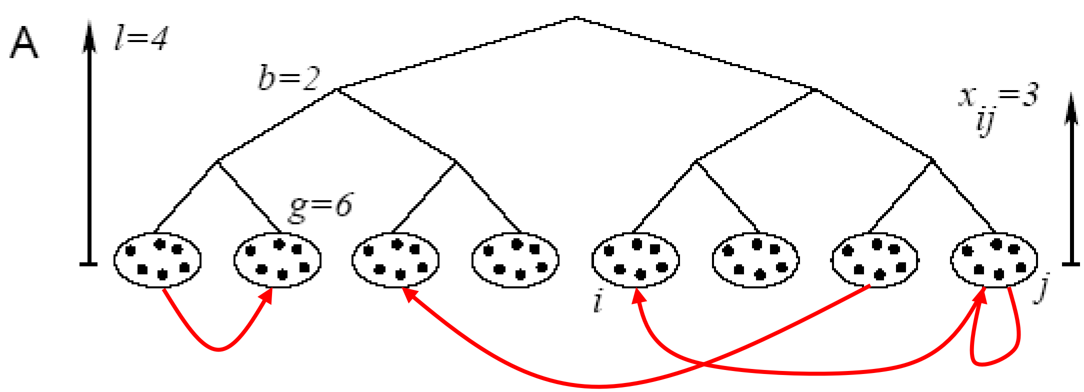
Other models

- Lattice captures geographic distance. How do we capture social distance (e.g. occupation)?
- Hierarchical organization of groups
  - distance h(i,j) = height of Least Common Ancestor



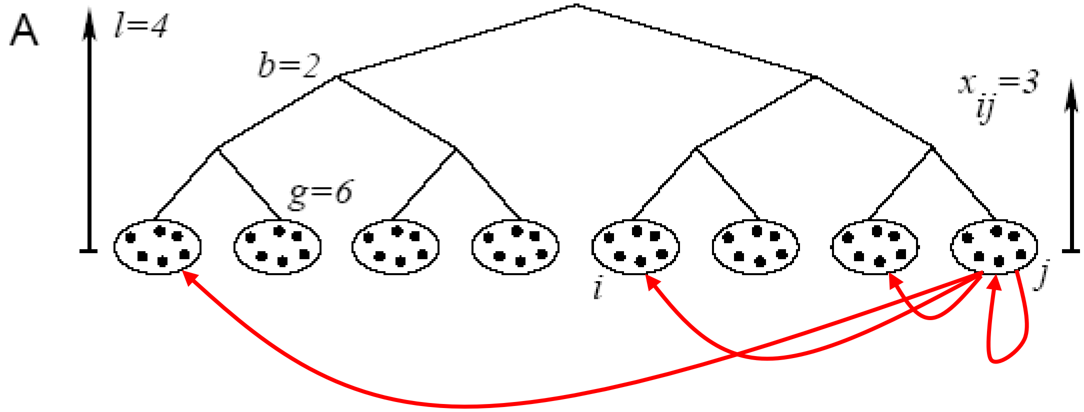
Other models

- Generate links between leaves with probability proportional to $b^{-\alpha h(i,j)}$
  - b=2 the branching factor



Other models

- Theorem: For α=1 there is a polylogarithimic search algorithm. For α≠1 there is no decentralized algorithm with poly-log time
  - note that α=1 and the exponential dependency results in uniform probability of linking to the subtrees



Searching Power-law networks

- Kleinberg considered the case that you can fix your network as you wish. What if you cannot?
- [Adamic et al.] Instead of performing simple BFS flooding, pass the message to the neighbor with the highest degree
- Reduces the number of messages to $O(n^{(a-2)/(a-1)})$

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

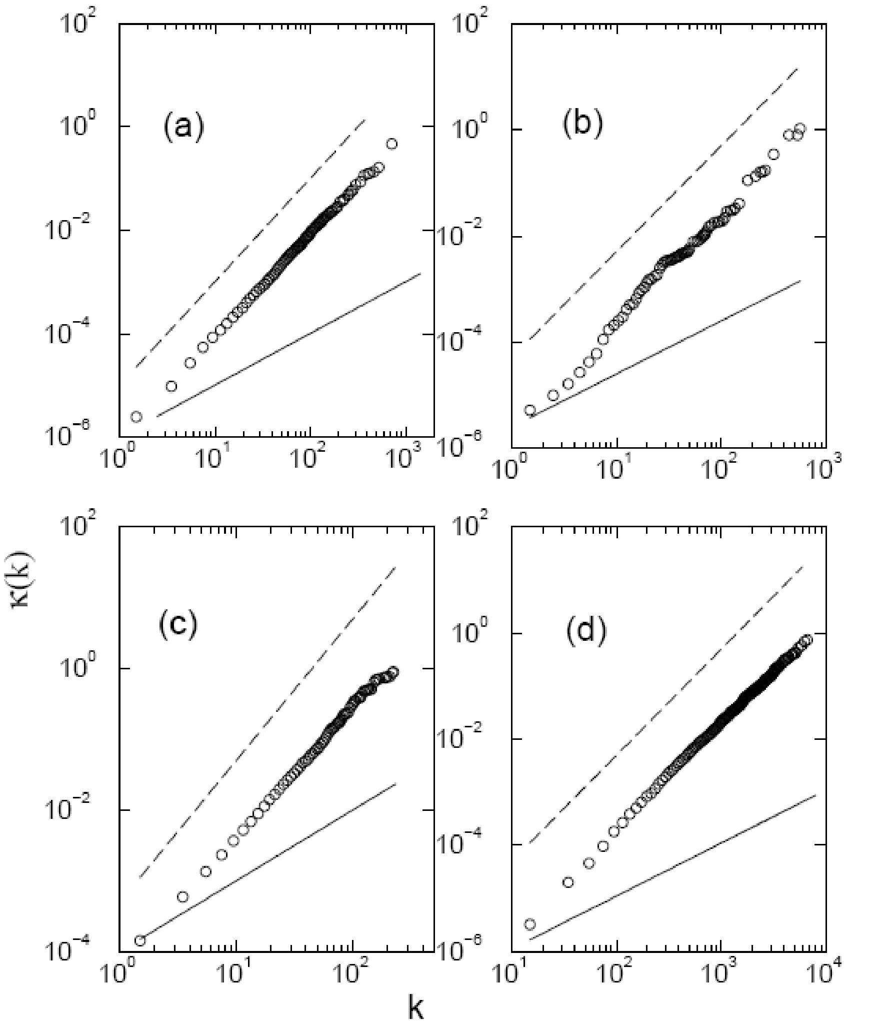
$$f(k) = \sum_{i:k_i=k} D_i$$

- 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



(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

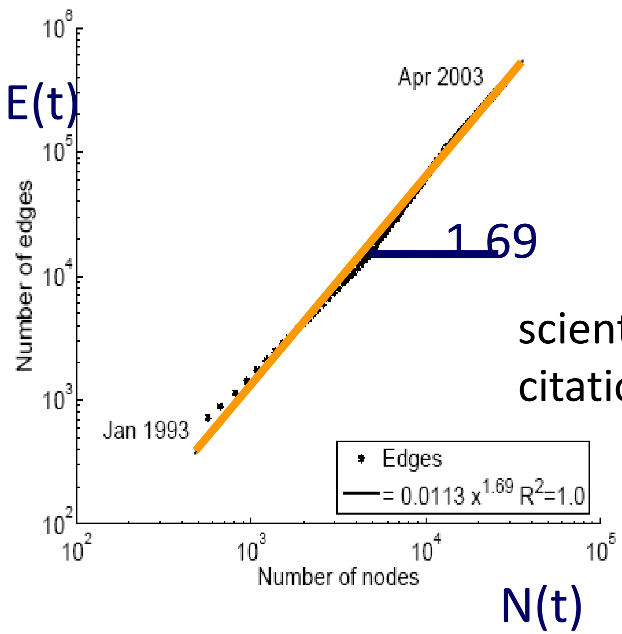
- What about real graphs?
  - Leskovec, Kleinberg, Faloutsos 2005

Densification laws

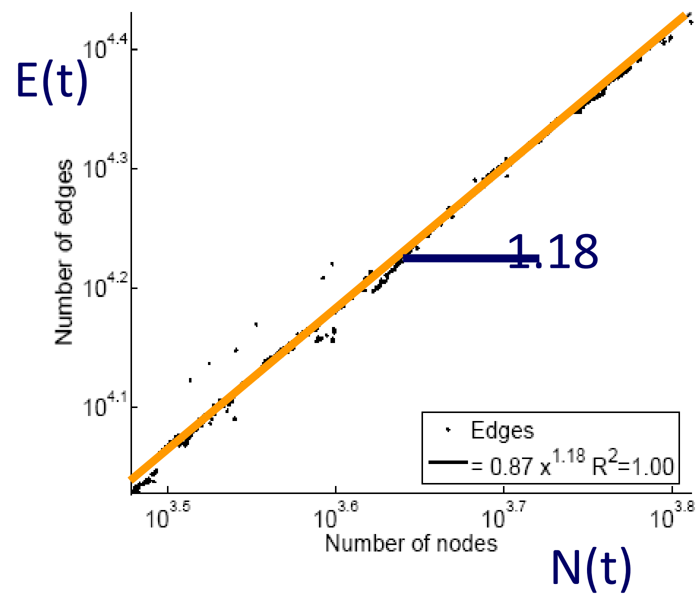
- In real-life networks the average degree increases! – networks become denser!

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

α = densification exponent

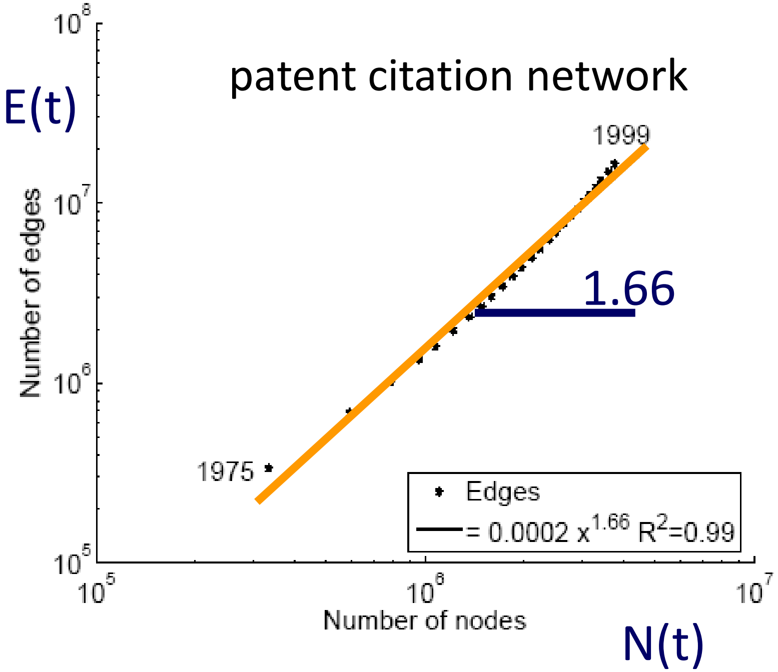


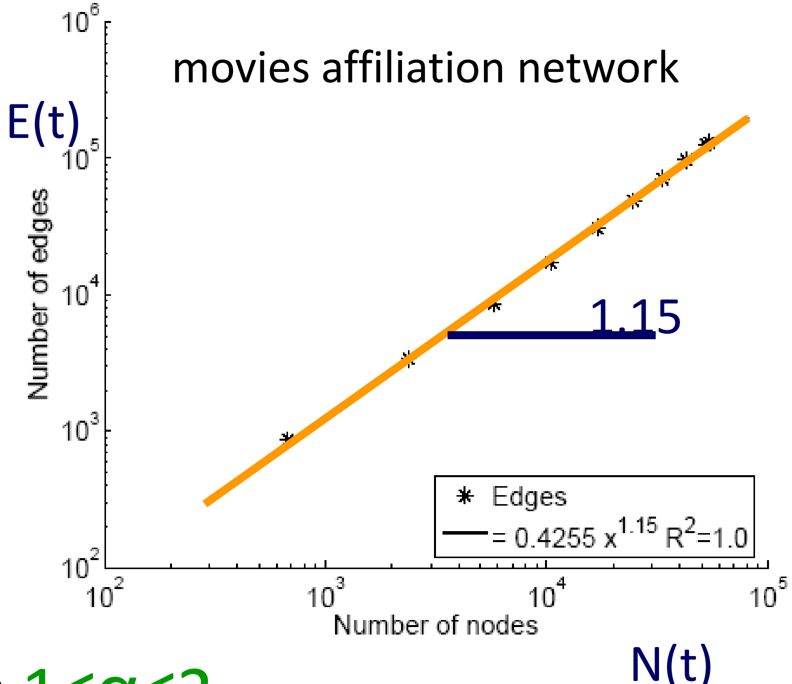
scientific
citation network



Internet

More examples

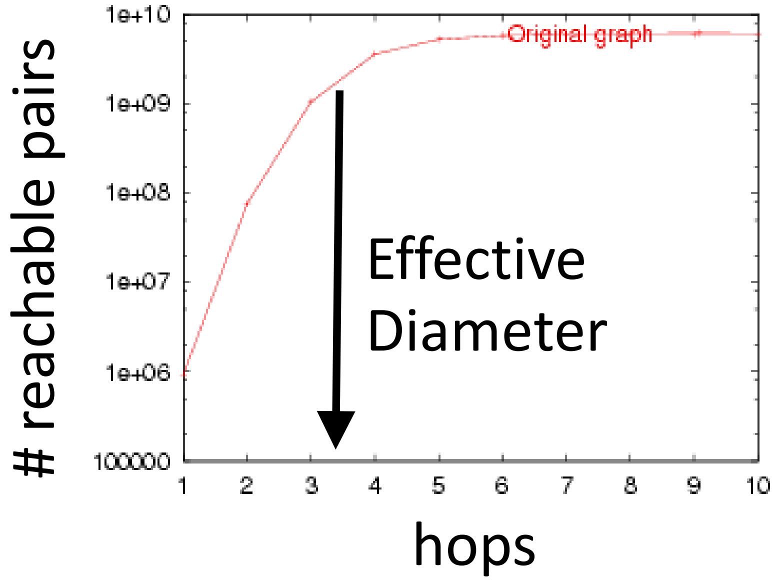




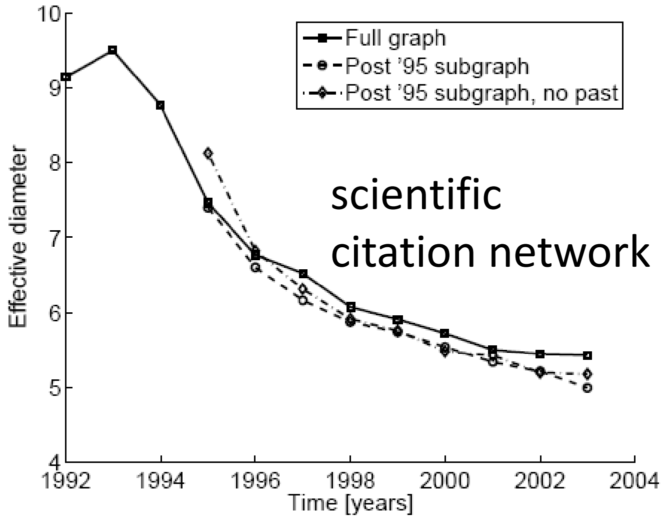
- The densification exponent $1 \leq \alpha \leq 2$
  - $\alpha = 1$: linear growth – constant out degree
  - $\alpha = 2$: quadratic growth - clique

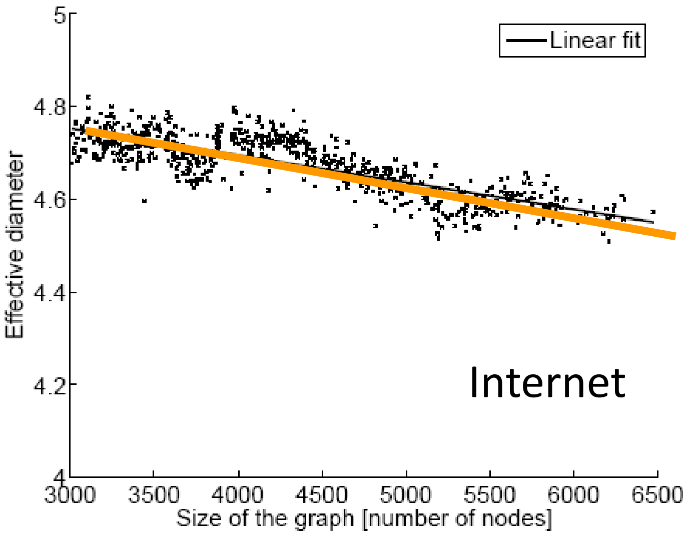
What about diameter?

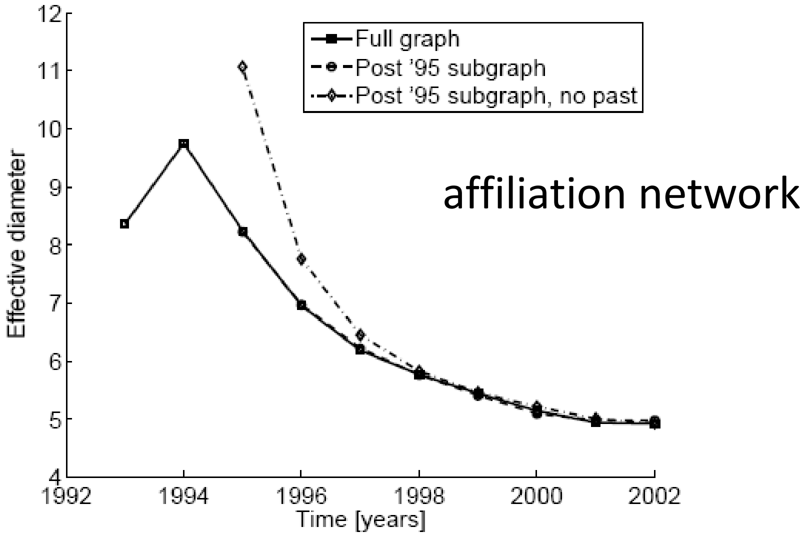
- Effective diameter: the interpolated value where 90% of node pairs are reachable

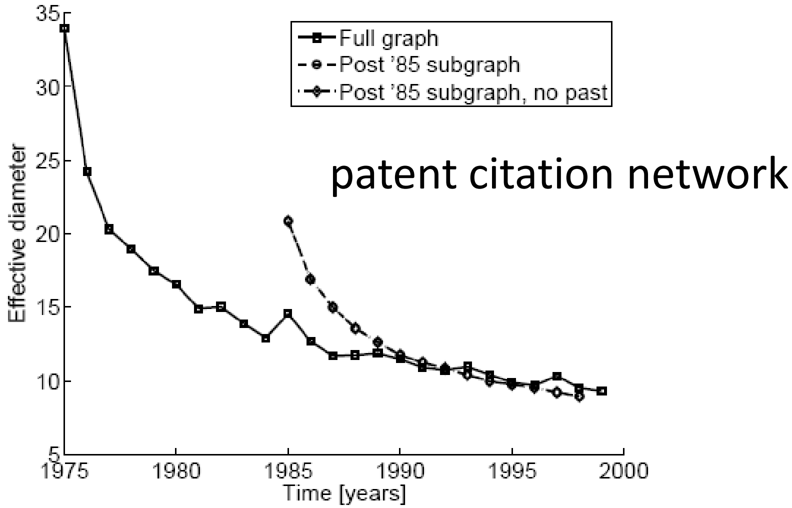


Diameter shrinks









Densification – Possible Explanation

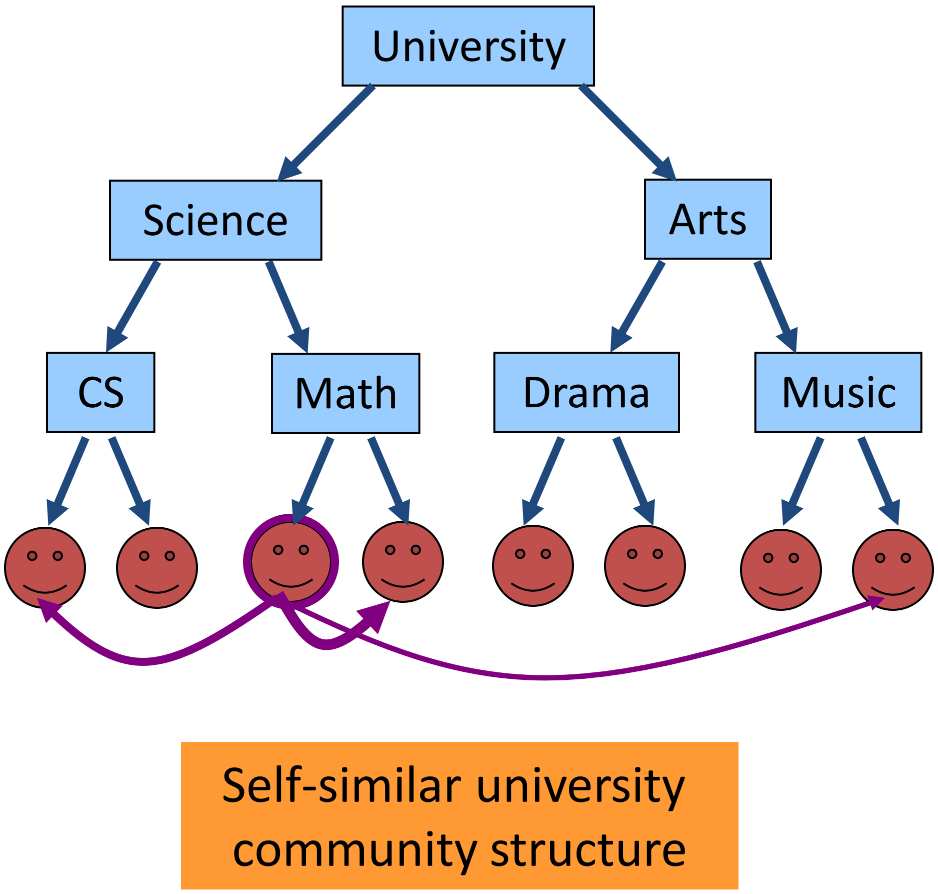
- Existing graph generation models do not capture the Densification Power Law and Shrinking diameters
- Can we find a simple model of local behavior, which naturally leads to observed phenomena?

- 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:

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

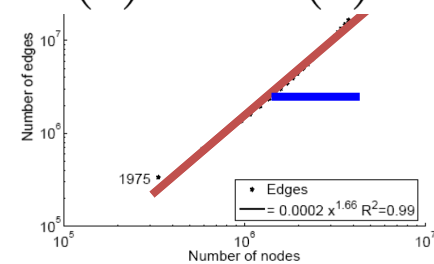
where: $c \geq 1$ … the Difficulty constant

h … tree-distance

Densification Power Law

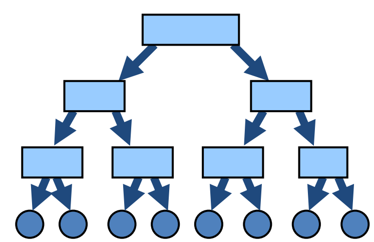
- <u>Theorem:</u> The Community Guided Attachment leads to Densification Power Law with exponent

$$E(t) \propto N(t)^a$$



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

- α … densification exponent
- 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: $\alpha = 2$, quadratic growth of edges – near clique
- If $c = b$: hard to cross communities
  - Then: $\alpha = 1$, 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

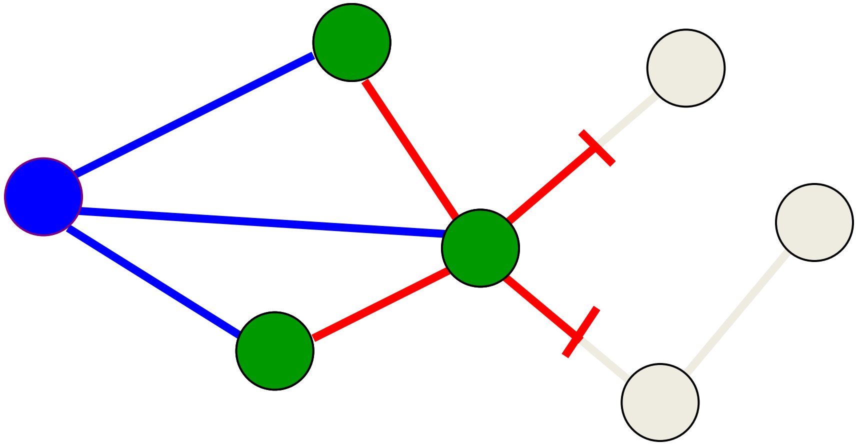
“Forest Fire” model – Intuition

- 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

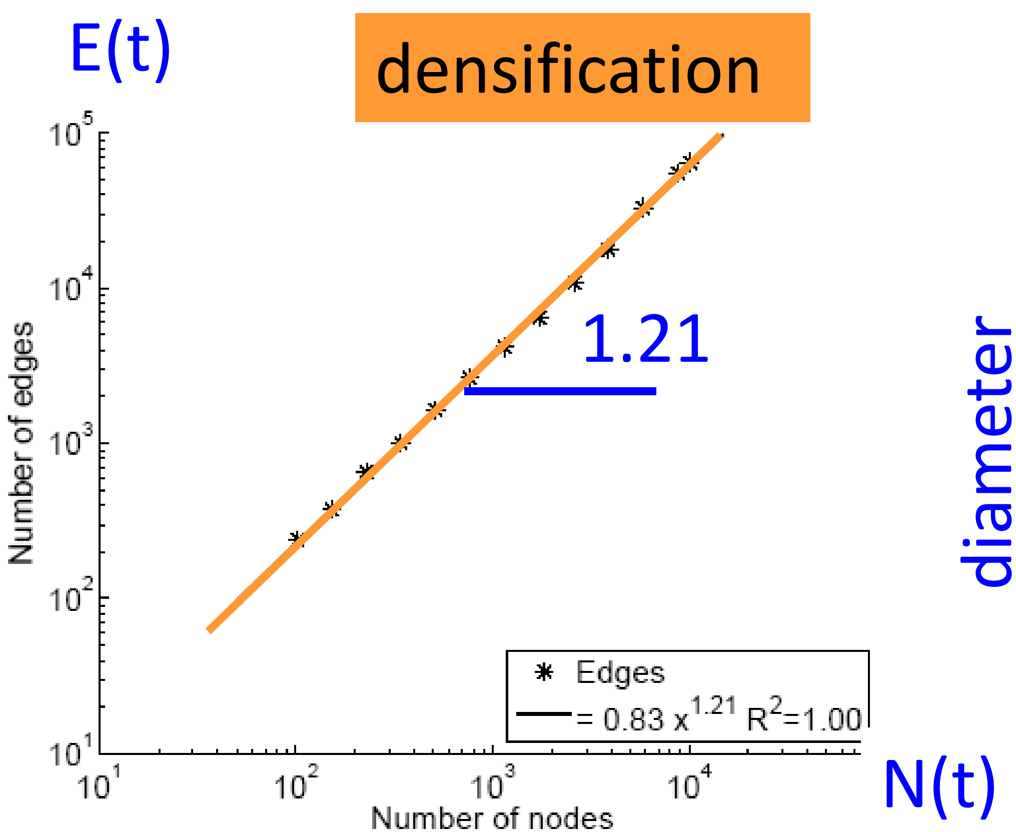
“Forest Fire” – the Model

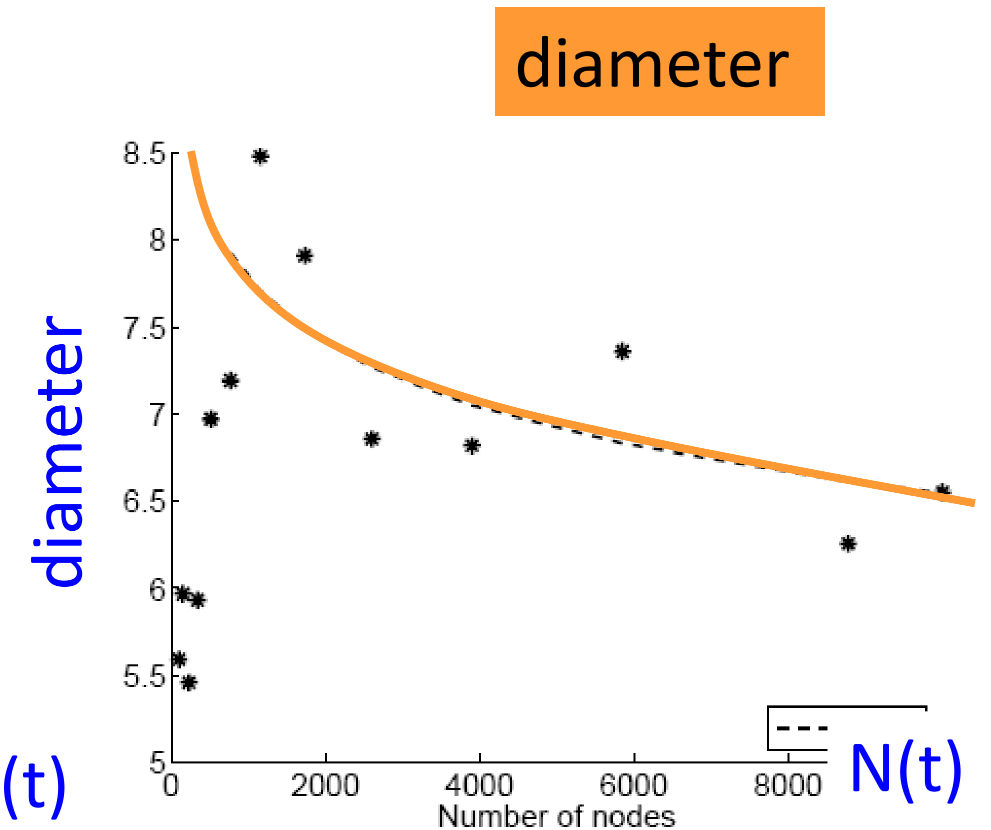
- A node arrives
- Randomly chooses an “ambassador”
- Starts burning nodes (with probability p) and adds links to burned nodes
- “Fire” spreads recursively



Forest Fire in Action (1)

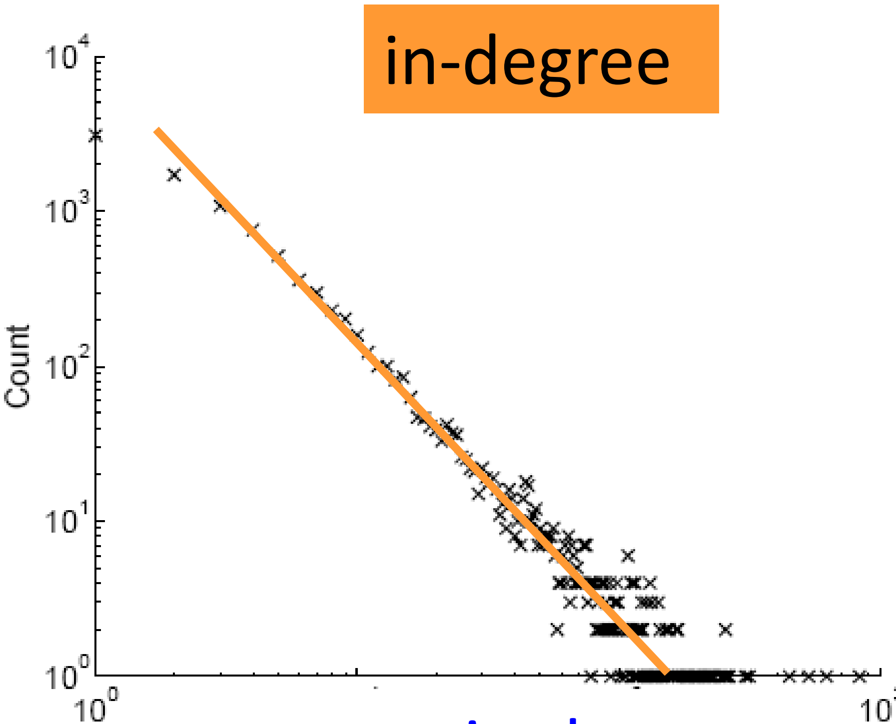
- Forest Fire generates graphs that Densify and have Shrinking Diameter



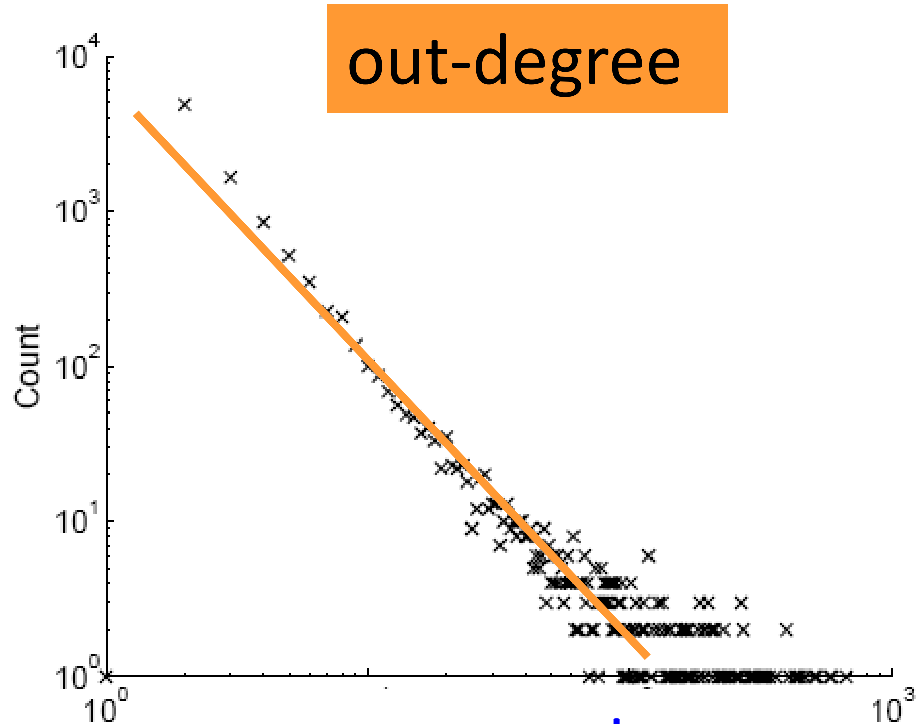


Forest Fire in Action (2)

- 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

Information-propagation models

- Epidemics
  - One of the major reasons that people started studying social networks in the first place
  - How do epidemic diseases propagate through society
- Consumer’s society
  - How trends and products propagate?
  - Major reason for studying online social networks

SIR model

- **S**: susceptible
  - A node in state **S** does not have the disease but he can, in principle, get it through someone else
- **I**: Infected
  - A node in state **I** has the disease and he can pass it on
- **R**: Recovered
  - A node is state **R** does not had the disease in the past, recovered from it and has eternal immunity

SIR model

- Any susceptible individual has uniform probability β of catching the disease per unit time
- Any infected individual can become cured at rate γ
- **Questions/problems:**
  - Given an epidemic how can we compute the parameters β and γ
  - Given a network and an epidemic, with known parameters, which are the nodes to vaccine to prevent the global explosion of the epidemic?

SIS model

- **S**: susceptible
  - A node in state **S** does not have the disease but he can, in principle, get it through someone else
- **I**: Infected
  - A node in state **I** has the disease and he can pass it on
- *A node can never get eternal immunity; once an infected node is cured he becomes susceptible again!*

The deterministic propagation model

The independent cascade model (IC)

The Linear threshold model (LT)

Problems – consumer's society

- Which nodes should I influence to buy a product so that the product becomes a trend?