Preferential Attachment and Network Evolution

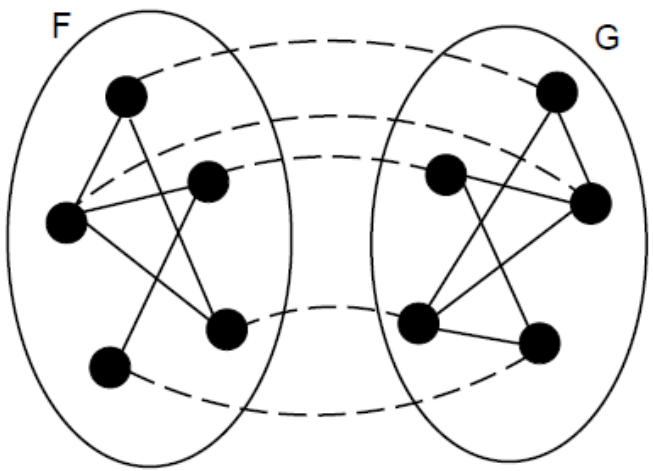
CS224W: Social and Information Network Analysis

Jure Leskovec, Stanford University

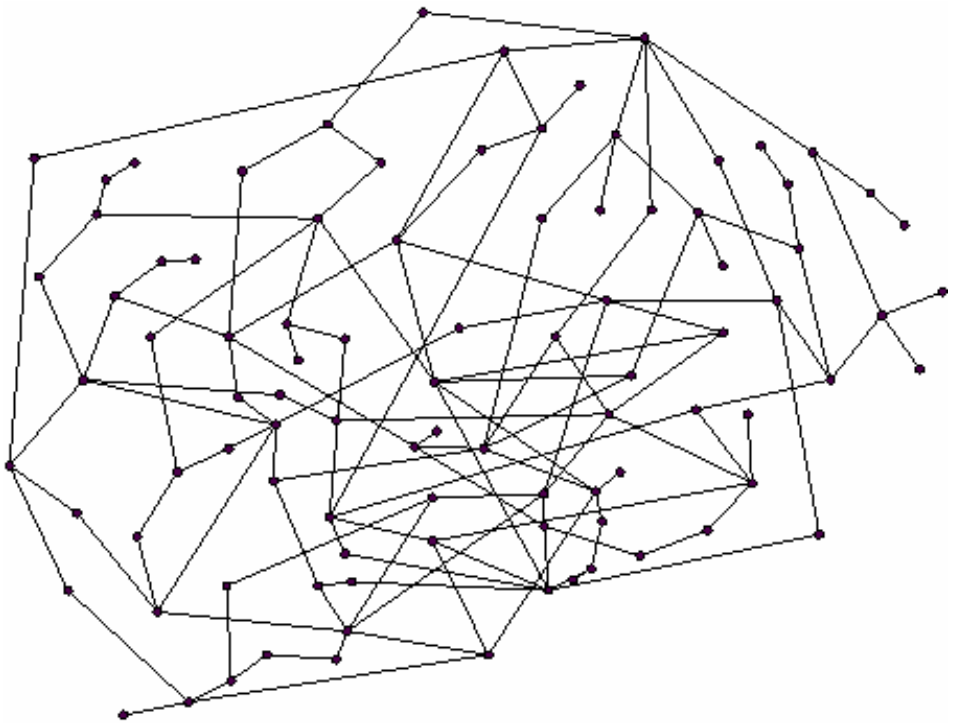
http://cs224w.stanford.edu

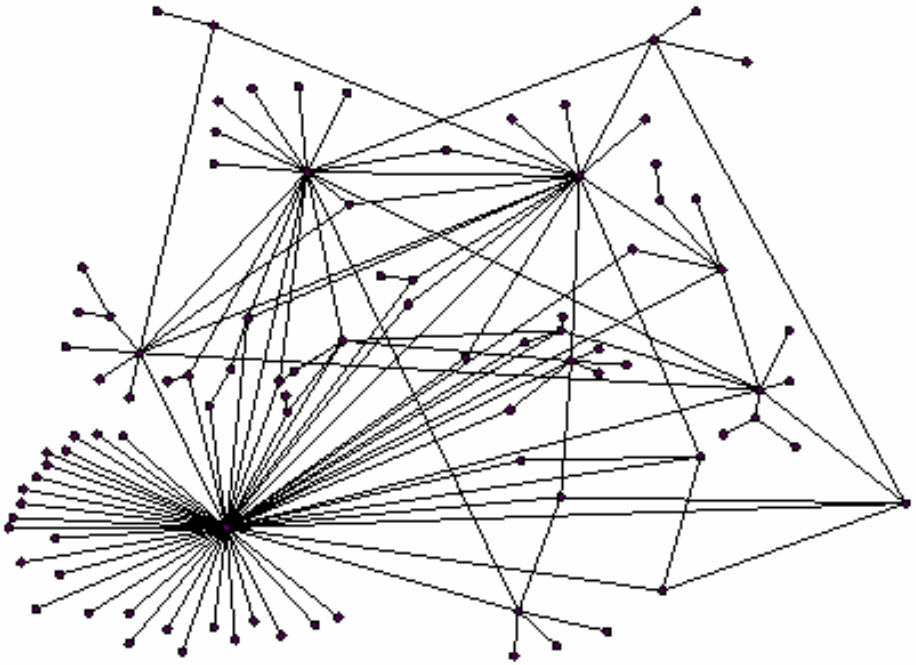
Announcement: HW3-Competition

- **Task: (HW3 is optional)**
  - Find node correspondences between two graphs
- **Incentives:**
  - **European chocolates!**
  - Fame!
  - Up to 10% extra credit
- **Due:**
  - Monday Nov 14
  - No late days!



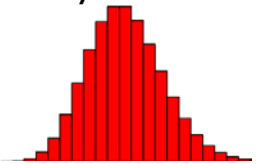
Random vs. Scale-free network





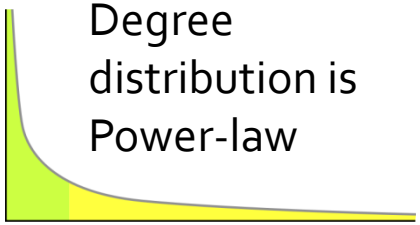
Random network

(Erdos-Renyi random graph)



Degree distribution is Binomial

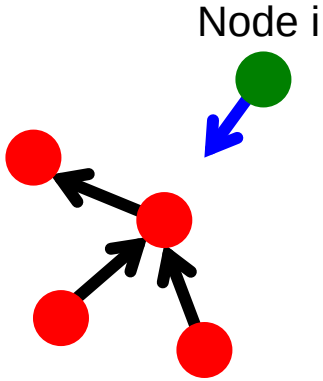
Scale-free (power-law) network



Degree distribution is Power-law

[Mitzenmacher, '03]

The Exact Model



We will analyze the following model:

- Nodes arrive in order $1,2,3,\ldots,n$
- When node i is created it makes a single link to an earlier node i chosen:
  - 1) With prob. p, i links to j chosen uniformly at random (from among all earlier nodes)
  - 2) With prob. $1-p$, node i chooses node j uniformly at random and links to a node j points to.

The Model Givens Power-Laws

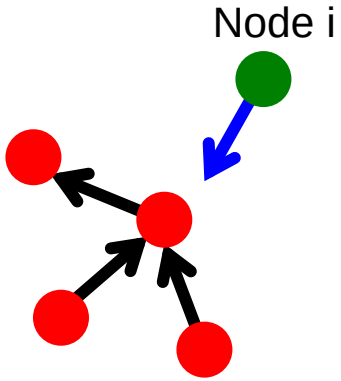
- **<u>Claim:</u>** The described model generates networks where the fraction of nodes with degree k scales as:

$$P(d_i = k) \propto k^{-(1+\frac{1}{q})}$$

where q=1-p

- **Consider deterministic and continuous** approximation to the in-degree of node *i* as a function of time *t*
  - *t* is the number of nodes that have arrived so far
  - In-degree $d_i(t)$ of node i ($i=1,2,\ldots,n$) is a **continuous quantity** and it **grows deterministically** with time t

Continuous Degree: What We Know



- **Initial condition:**
  - $d_i(t)=0$, when $t=i$ (node i just arrived)
- **Expected change of $d_i(t)$ over time:**
  - Node i gains an in-link at step $t+1$ only if a link from a newly created node $t+1$ points to it.
  - **What's the probability of this event?**
    - With prob. p node $t+1$ links **randomly**:
      - Links to our node i with prob. $1/t$
    - With prob. $1-p$ node $t+1$ links **preferentially**:
      - Links to our node i with prob. $d_i(t)/t$
  - **So: Prob. node $t+1$ links to i is:** $\mathbf{p}\frac{1}{t}+(1-p)\frac{d_i(t)}{t}$

What is the rate of growth of d_i?

- **Expected change of $d_i(t)$ over time**
- $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$
- $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ — Divide by $p+q\,d_i(t)$
- $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$ — integrate
- $\frac{1}{q}\ln\big(p + qd_i(t)\big) = \ln t + c$ — Let $A=e^c$ and exponentiate
- $qd_i(t) + p = A\,t^q \;\Rightarrow d_i(t) = \frac{1}{q}(At^q - p)$

What is the constant A?

$$d_i(t) = \frac{1}{q}(At^q - p)$$

What is the constant A?

- **We know:** $d_i(i) = 0$
- **So:** $d_i(i) = \frac{1}{q}(Ai^q - p) = 0$
- $\Rightarrow A = \frac{p}{i^q}$
- $\Rightarrow d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right)$

Degree Distribution

- What is F(d) the fraction of nodes that has degree at <u>least</u> d at time t?
  - How many nodes i have degree > t?
    - $d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right) > d$
    - then: $\mathrm{i} < \mathrm{t}\left(\frac{q}{p}d - 1\right)^{-\frac{1}{q}}$
- There are t nodes total at time t so F(d):

$$F(d) = \left[\frac{q}{p}d + 1\right]^{-\frac{1}{q}}$$

Degree Distribution

- **What is the fraction of nodes with degree exactly d?**
  - Take derivative of F(d):
    - F(d) is CDF, so F'(d) is the PDF

$$F'(d) = \frac{1}{p}\left[\frac{q}{p}d + 1\right]^{-1-\frac{1}{q}} \Rightarrow \alpha = 1 + \frac{1}{q}$$

Preferential attachment: Reflections

- Two changes from the G_{np}
  - Groth + Preferential attachment
- Do we need both? **Yes!**
  - **Add growth to G_{np}**

    (assume 1 edge is added at each step)
    - X_j = degree of node j at the end
    - $X_j(u) = 1$ if node u links to j, else 0
    - $X_j = X_j(j+1) + X_j(j+2) + \cdots + X_j(n)$
    - $E[X_j(u)] = P[u \text{ links to } j] = \frac{1}{u-1}$
    - $E[X_j] = \sum_{u=j}^{n} \frac{1}{u-1} = H_{n-1} - H_j$
    - $E[X_j] = \log(n-1) - \log(j) = \log\left(\frac{n-1}{j}\right) \neq \left(\frac{n}{j}\right)^{\alpha}$

H_n…n^{th} harmonic number:

$$H_n = \sum_{k=1}^{n} \frac{1}{k} \approx \log(n)$$

H_n…n^{th} harmonic number:

$$H_n = \sum_{k=1}^{n} \frac{1}{k} \approx \log(n)$$

Preferential attachment: Good news

- **Preferential attachment gives power-law degrees**
- Intuitively reasonable process
- Can tune p to get the observed exponent
  - On the web, $P[\text{node has degree } d] \sim d^{-2.1}$
  - $2.1 = 1+1/(1-p) \rightarrow \underline{p \sim 0.1}$

There are also other network formation mechanisms that generate scale-free networks:

- Random surfer model
- Forest Fire model

PA-like Link Formation

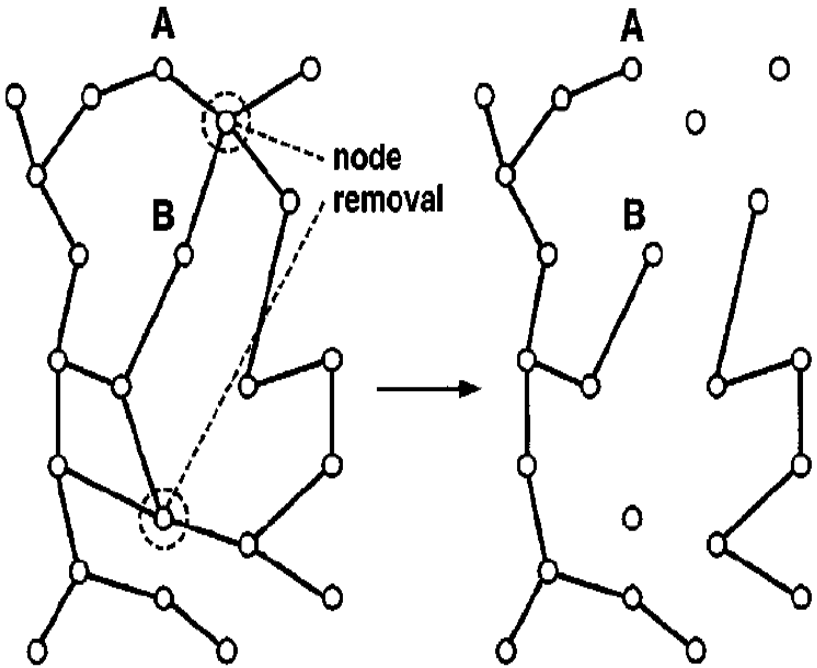
- **Copying mechanism** (directed network)
  - select a node and an edge of this node
  - attach to the endpoint of this edge
- **Walking on a network** (directed network)
  - the new node connects to a node, then to every
  - first, second, … neighbor of this node
- **Attaching to edges**
  - select an edge
  - attach to both endpoints of this edge
- **Node duplication**
  - duplicate a node with all its edges
  - randomly prune edges of new node

Preferential attachment: Bad news

- **Preferential attachment is not so good at predicting network structure**
  - Age-degree correlation
  - Links among high degree nodes
    - On the web nodes sometime avoid linking to each other
- **Further questions:**
  - What is a reasonable probabilistic model for how people sample through web-pages and link to them?
    - Short+Random walks
    - Effect of search engines – reaching pages based on number of links to them

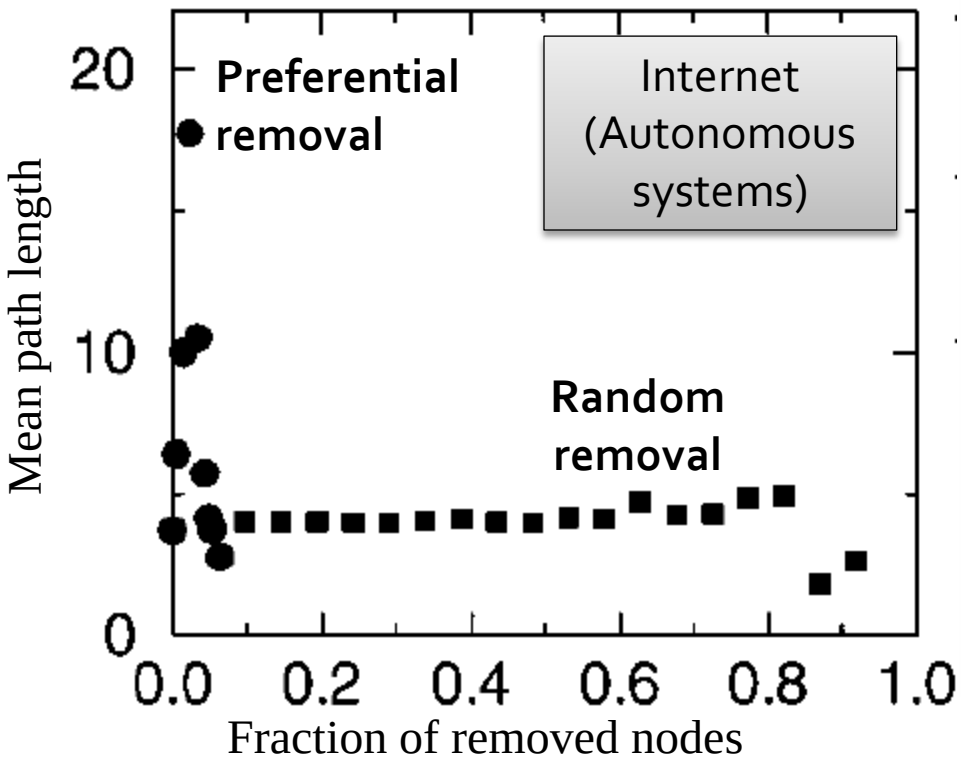
Network resilience

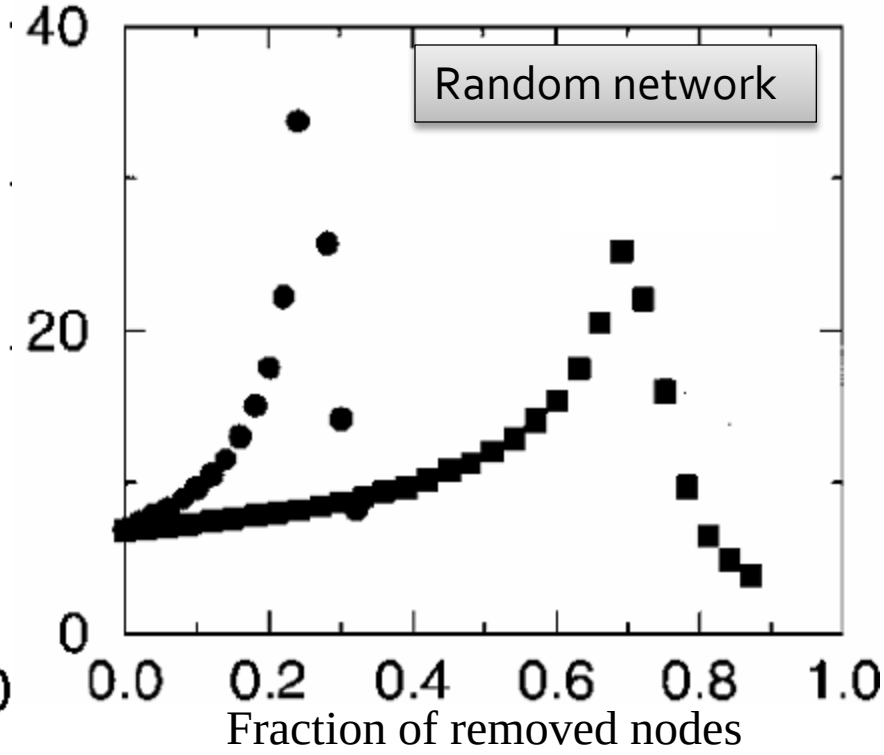
- **How does the connectivity of the network change as the vertices get removed?**
  [Albert et al. 00; Palmer et al. 01]
- **Vertices can be removed:**
  - Uniformly at random
  - In order of decreasing degree
- It is important for **epidemiology**
  - Removal of vertices corresponds to vaccination



Network Resilience

- **Real-world networks are resilient to random attacks**
  - You need to remove all web-pages of degree > 5 to disconnect the web
  - But this is a very small fraction of all web pages
- **Random network has better resilience to targeted attacks**





Evolution of Social Networks

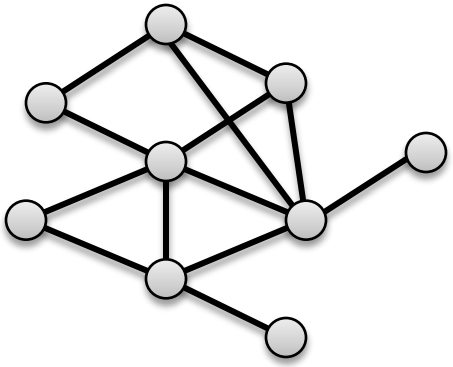
Network Evolution: Observation

- **Preferential attachment is a model of a growing network**
- **What governs network growth and evolution?**
  - P1) Node arrival process:
    - When nodes enter the network
  - P2) Edge initiation process:
    - Each node decides when to initiate an edge
  - P3) Edge destination process:
    - The node determines destination of the edge

[Leskovec et al., KDD '08]

Let's Look at the Data

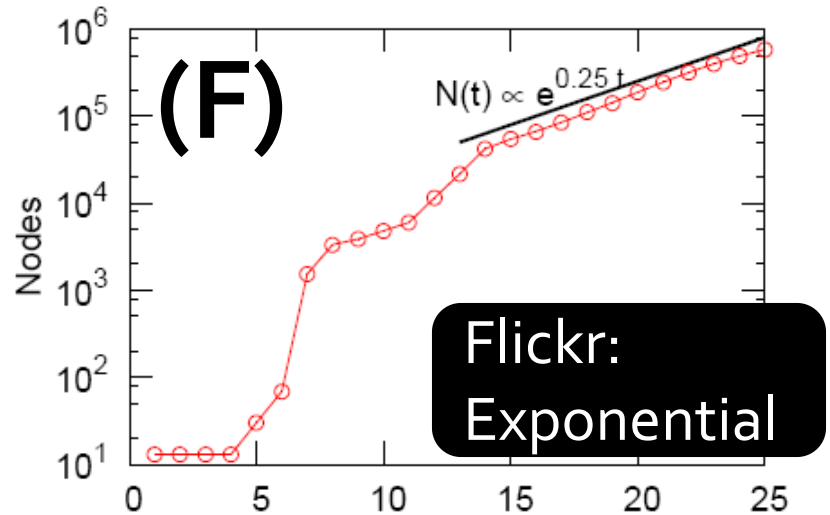
- **4 online social networks with exact edge arrival sequence**
  - For every edge *(u,v)* we know **exact time** of the appearance t_{uv}
- **Directly observe mechanisms leading to global network properties**

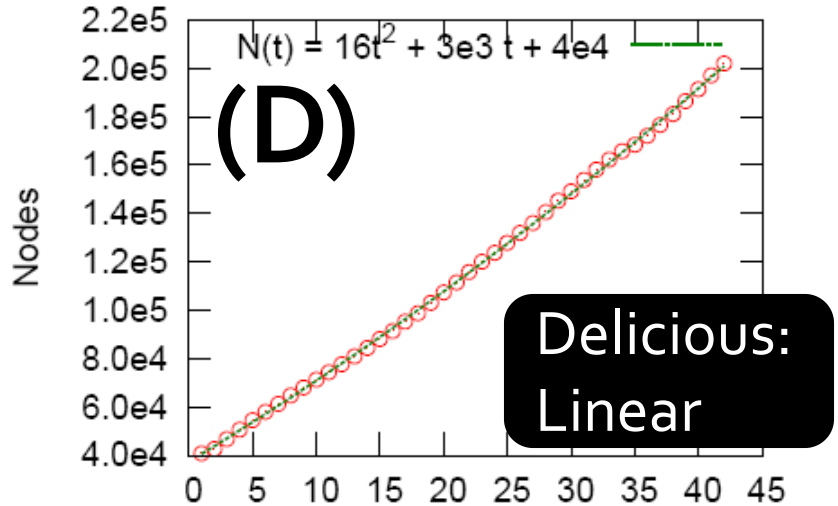


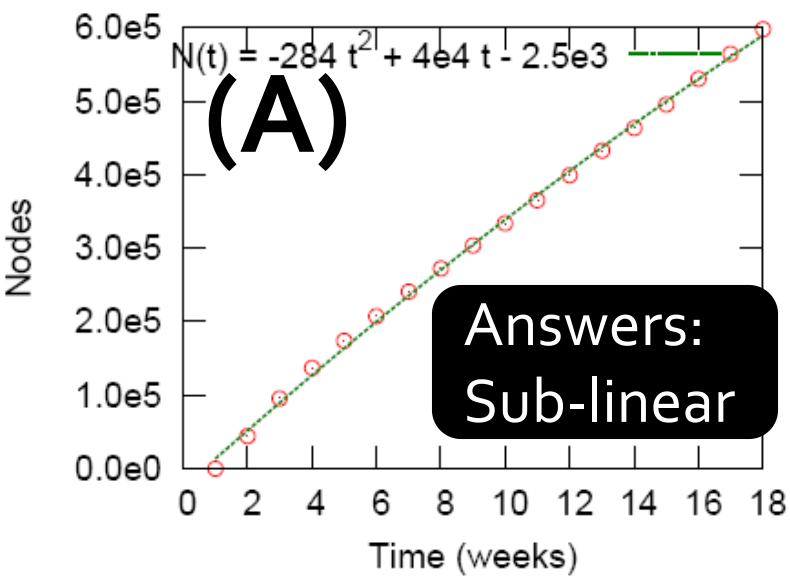
and so on for millions...

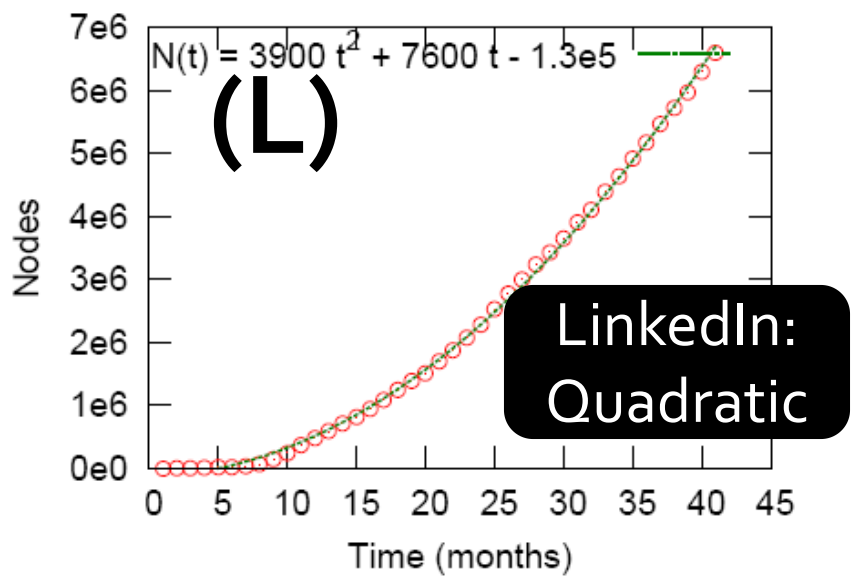
	Network	T	N	E
(F)	FLICKR (03/2003–09/2005)	621	584,207	3,554,130
(D)	DELICIOUS (05/2006–02/2007)	292	203,234	430,707
(A)	ANSWERS (03/2007–06/2007)	121	598,314	1,834,217
(L)	LINKEDIN (05/2003–10/2006)	1294	7,550,955	30,682,028

P1) When are New Nodes Arriving?



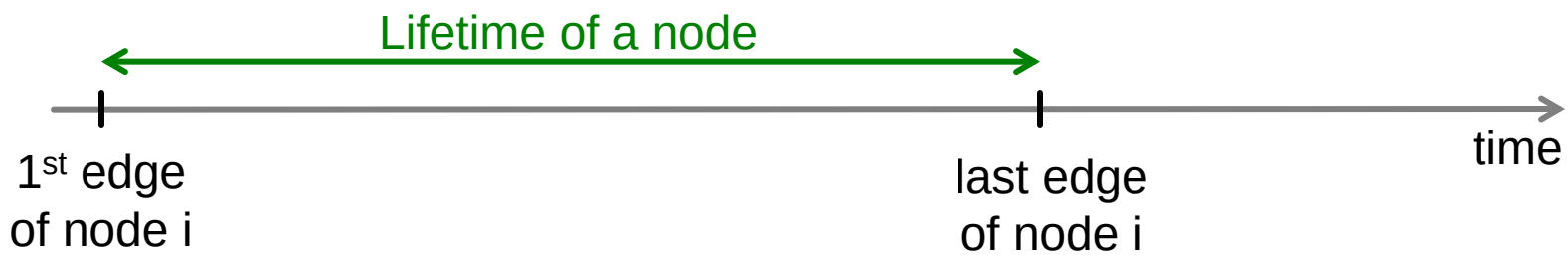




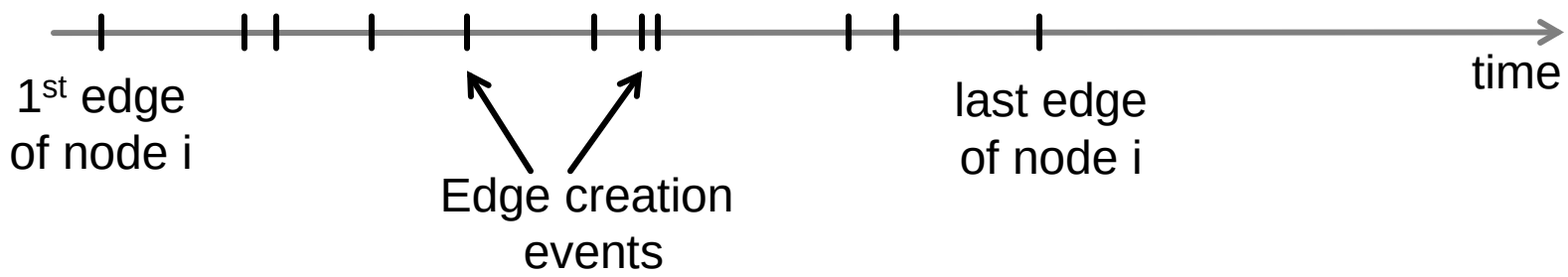


P2) When Do Nodes Create Edges?

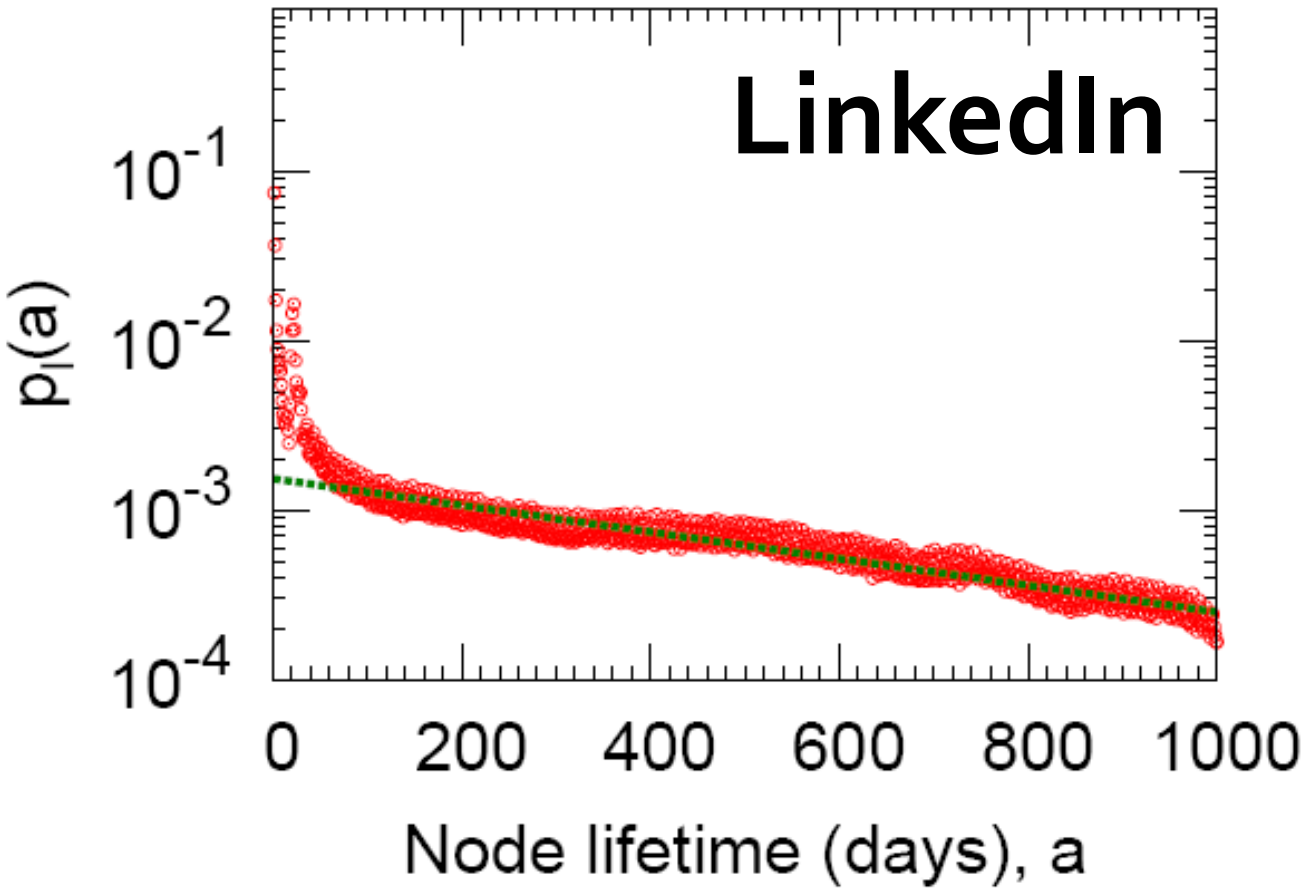
- **How long do nodes live?**
  - Node life-time is the time between the 1st and the last edge of a node



- **How do nodes "wake up" to create links?**



P2) What is Node Lifetime?



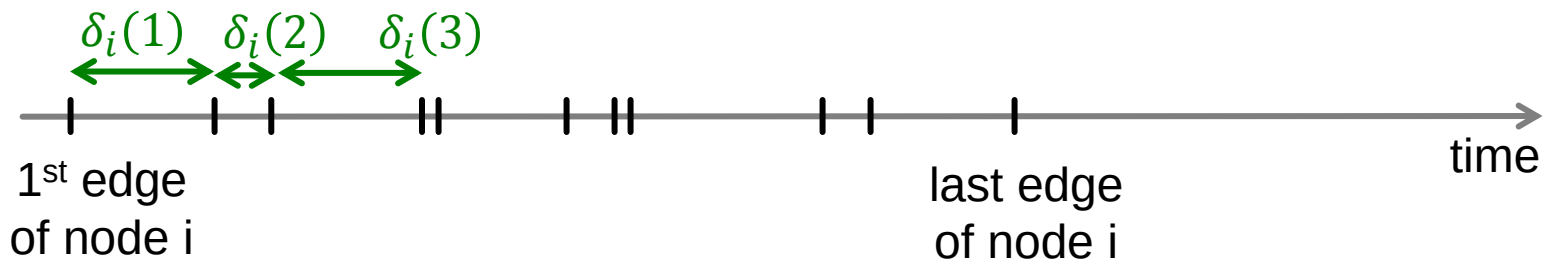
- **Lifetime *a***: time between node's first and last edge

Node lifetime is **exponentially distributed**:

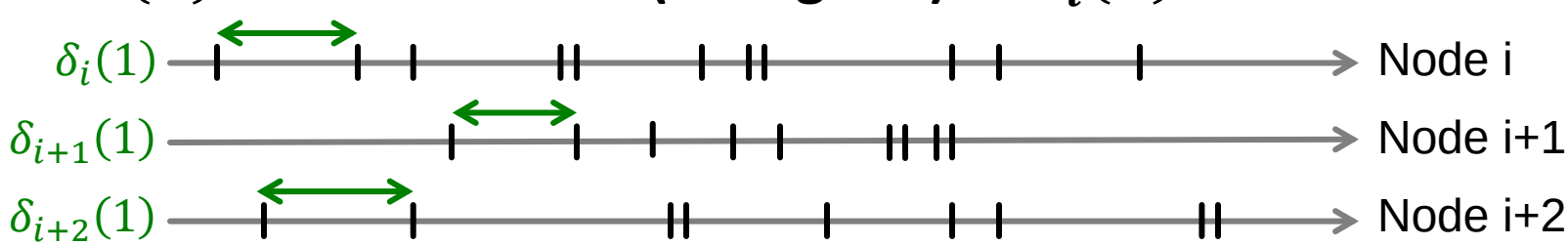
$$p_l(a) = \lambda e^{-\lambda a}$$

P2) When do Nodes Create Edges?

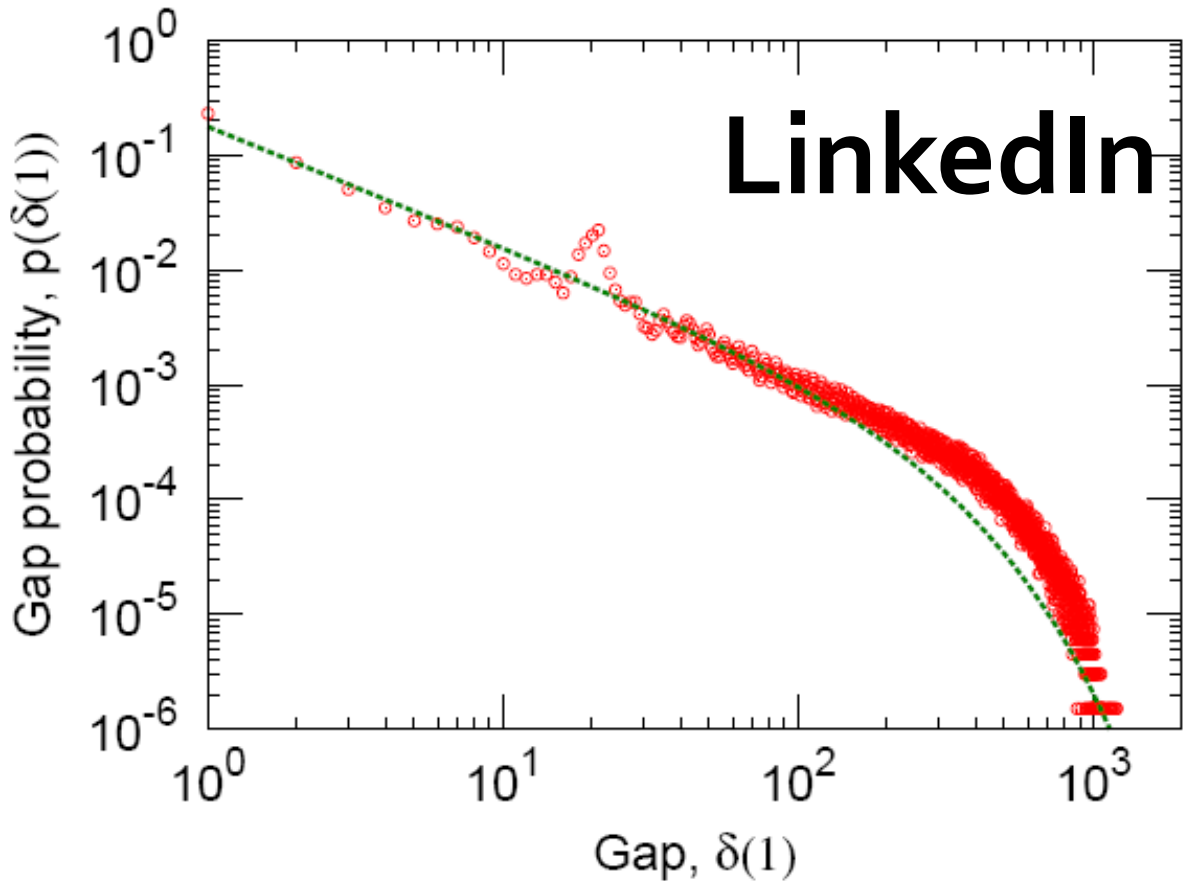
- **How do nodes "wake up" to create edges?**
  - **Edge gap $\boldsymbol{\delta_i(d)}$**: time between d^{th} and $d+1^{st}$ edge of node i:
    - Let $t_i(d)$ be the creation time of d-th edge of node i
    - $\delta_i(d) = t_i(d+1) - t_i(d)$



    - **$\boldsymbol{\delta(d)}$ is a distribution (histogram) of $\boldsymbol{\delta_i(d)}$ over all nodes i**



P2) When Do Nodes Create Edges?



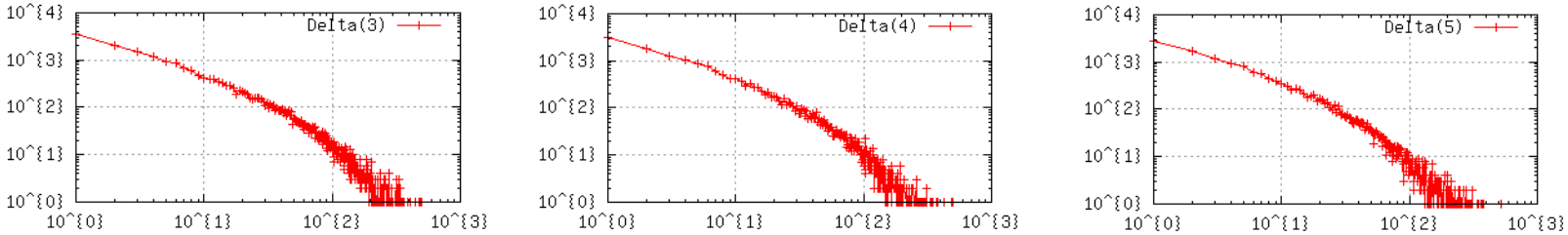
Edge gap δ(d): inter-arrival time between d^{th} and $d+1^{st}$ edge

For every *d* we get a separate histogram

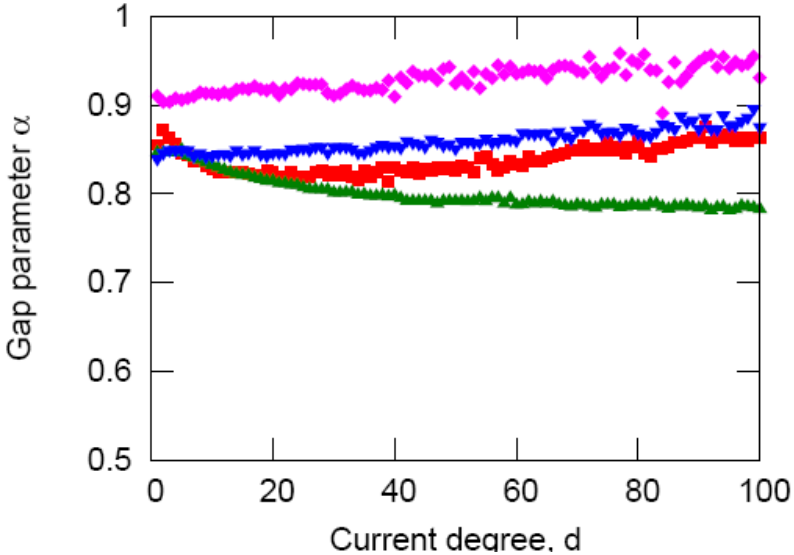
$$p_g(\delta(1)) \propto \delta(1)^{-\alpha} e^{-\beta}$$

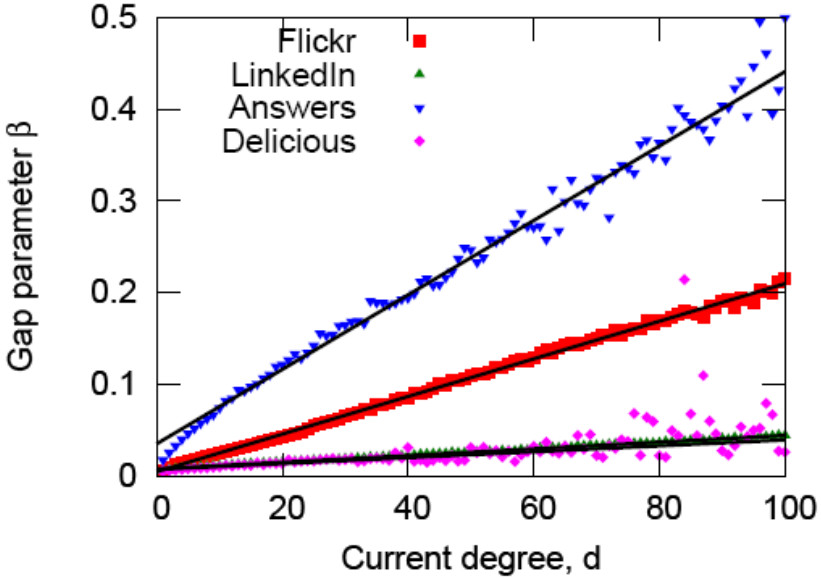
P2) How Do α & β Evolve With d?

- **How do α and β change as a function of d?**



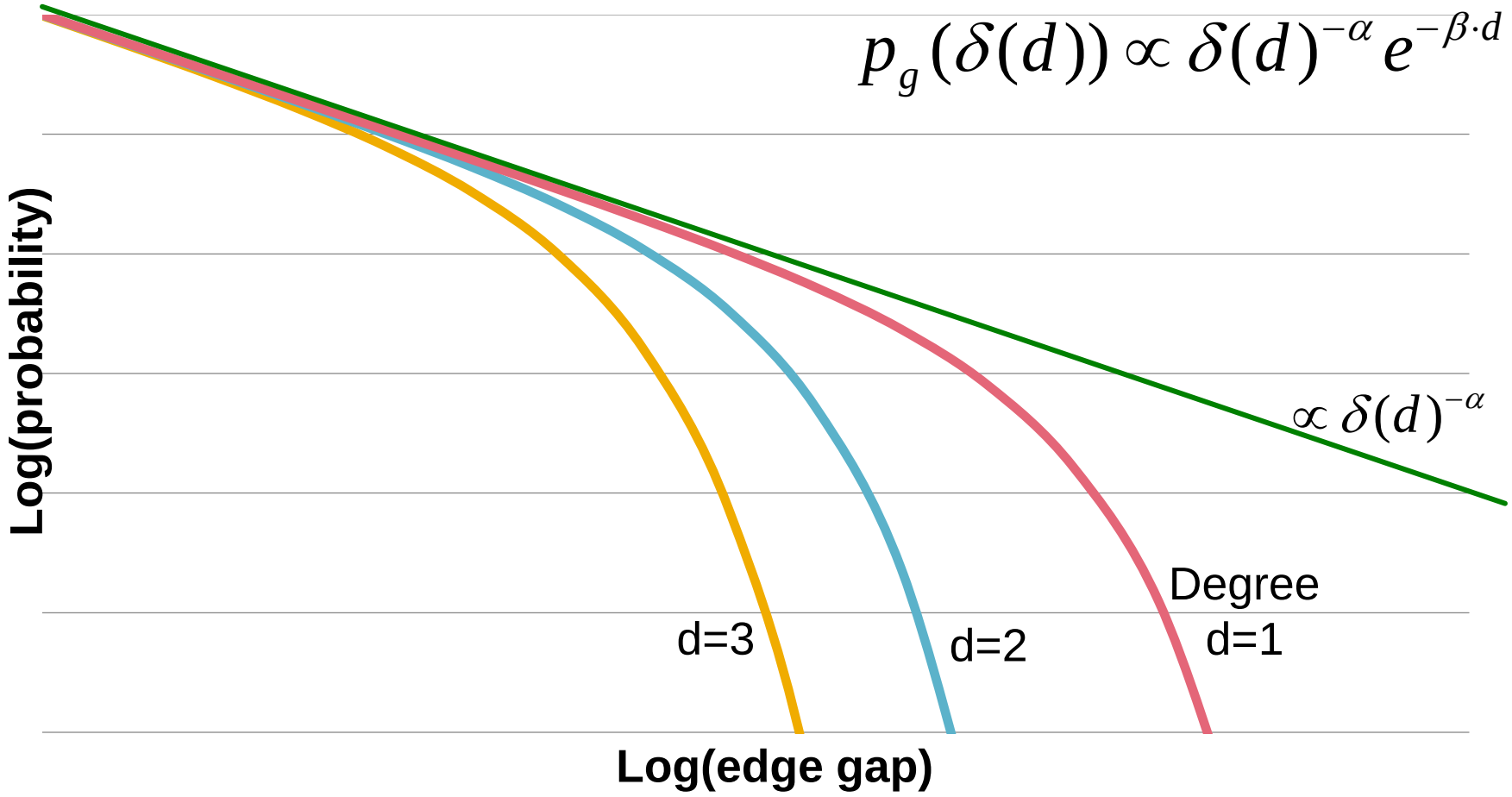
Fit to each plot of **δ(d)**: $p_g(\delta(d)) \propto \delta(d)^{-\alpha(d)} e^{-\beta(d)}$





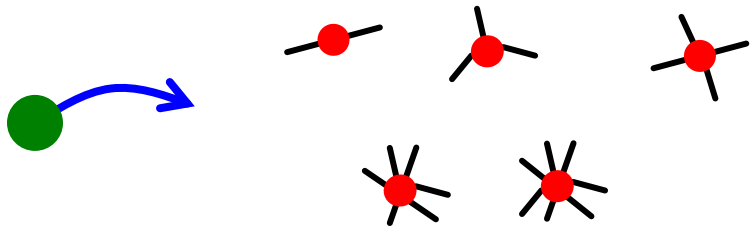
P2) Evolution of Edge Gaps

- α is const, β linear in d – gaps get smaller with d

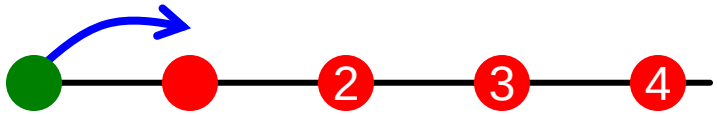


P3) How to Select Destination?

- Source node *i* wakes up and creates an edge
- How does *i* select a target node *j*?
  - What is the degree of the target *j*?
    - **Do preferential attachment really hold?**



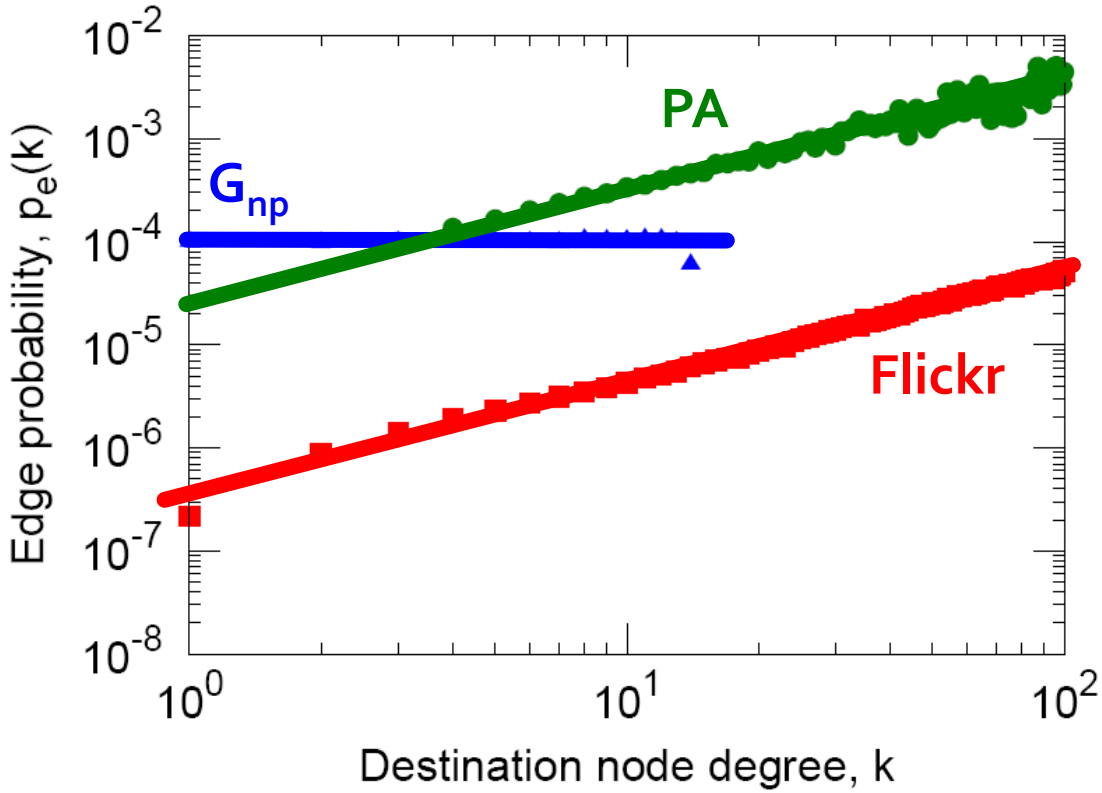
  - How many hops away is the target *j*?
    - **Are edges attaching locally?**



[Leskovec et al., KDD '08]

Edge Attachment Degree Bias

- Are edges more likely to connect to higher degree nodes?



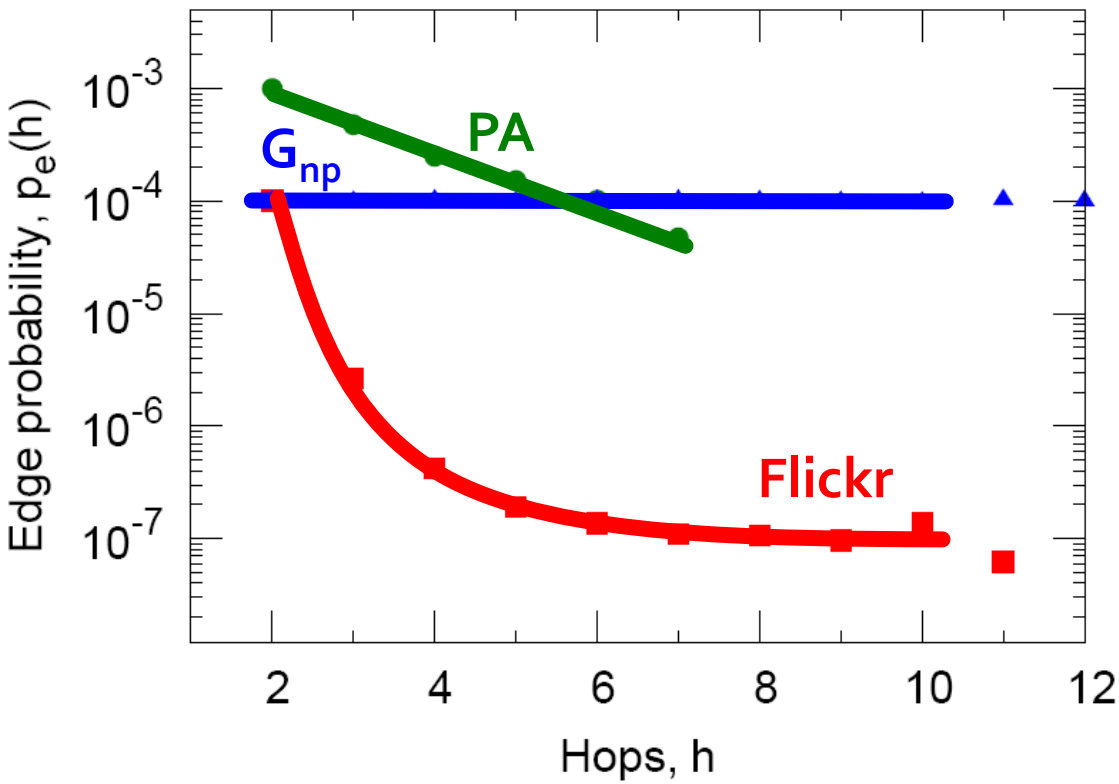
$$p_e(k) \propto k^{\tau}$$

Network	τ
G_{np}	0
PA	1
Flickr	1
Delicious	1
Answers	0.9
LinkedIn	0.6

[Leskovec et al., KDD '08]

How "far" is the Target Node?

- Just before the edge *(u,w)* is placed how many hops are between *u* and *w*?



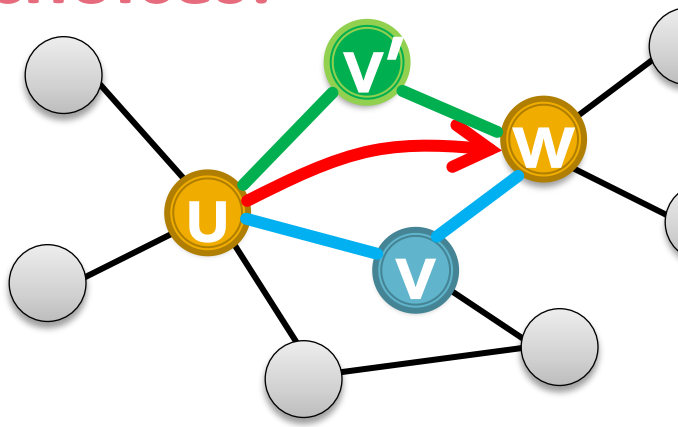
Fraction of triad closing edges	
Network	**% Δ**
Flickr	66%
Delicious	28%
Answers	23%
LinkedIn	50%

Real edges are local!
Most of them close **triangles**!

How to Close the Triangles?

- **Focus only on triad-closing edges**
- **New triad-closing edge *(u,w)* appears next**
- **Model this as 2 independent choices:**
  1. *u* choses neighbor *v*
  2. *v* choses neighbor *w* and connect *u* to *w*
  - E.g.: Under Random-Random:
    - $p(u,w) = \frac{1}{5} \cdot \frac{1}{2} + \frac{1}{5} \cdot 1 = \frac{3}{10}$
- Under a particular pair of "strategies":

  Likelihood of the graph = $\prod_{(u,w)\in E} p(u,w)$



Triad Closing Strategies

- **Improvement over the baseline:**
  - Baseline: Pick a random node 2 hops away

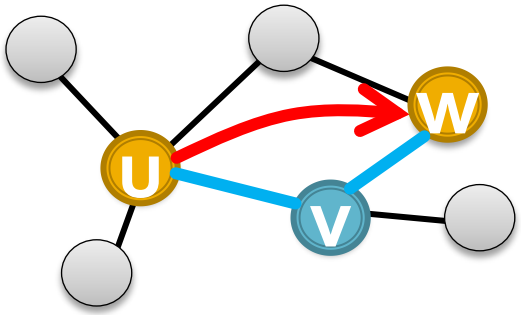
Strategy to select v (1st node)

Select w (2nd node)

FLICKR	random	$\deg^{0.2}$	com	$\text{last}^{-0.4}$	$\text{comlast}^{-0.4}$
random	13.6	13.9	14.3	16.1	15.7
$\deg^{0.1}$	13.5	14.2	13.7	16.0	15.6
$\text{last}^{0.2}$	14.7	15.6	15.0	17.2	**16.9**
com	11.2	11.6	11.9	13.9	13.4
$\text{comlast}^{0.1}$	11.0	11.4	11.7	13.6	13.2

Strategies to pick a neighbor:

- **random**: uniformly at random
- **deg**: proportional to its degree
- **com**: prop. to the number of common friends
- **last**: prop. to time since last activity
- **comlast**: prop. to **com*last**



[Leskovec et al., KDD '08]

Summary of the Model

- The model of network evolution

Process	Model
P1) Node arrival	• Node arrival function is given
P2) Edge initiation	• Node lifetime is exponential • Edge gaps get smaller as the degree increases
P3) Edge destination	Pick edge destination using random-random

[Leskovec et al., KDD '08]

Analysis of the Model

- Theorem: Exponential node lifetimes and power-law with exponential cutoff edge gaps lead to power-law degree distributions

- Interesting as temporal behavior predicts structural network property

[Leskovec et al., KDD '08]

Proof sketch

- Node lifetime: $p_l(a) =$
- Node of life-time a, what is its final degree D?

- What is distribution of D as a func. of λ, α, β?

- **The 2 exp. funcs. "cancel". Power-law survives**

Evolving the Networks

- Given the model one can take an existing network continue its evolution

- **Compare true and predicted** (based on the theorem) **degree exponent:**

	FLICKR	DELICIOUS	ANSWERS	LINKEDIN
λ	0.0092	0.0052	0.019	0.0018
α	0.84	0.92	0.85	0.78
β	0.0020	0.00032	0.0038	0.00036
true	1.73	2.38	1.90	2.11
predicted	1.74	2.30	1.75	2.08

Macroscopic Evolution

- **How do networks evolve at the macro level?**
  - What are global phenomena of network growth?

- **Questions:**
  - What is the relation between the number of nodes $n(t)$ and number of edges $e(t)$ over time t?
  - How does diameter change as the network grows?
  - How does degree distribution evolve as the network grows?

[Leskovec et al., KDD 05]

Network Evolution

- $N(t)$ … nodes at time t
- $E(t)$ … edges at time t
- Suppose that

$$N(t+1) = 2 * N(t)$$

- Q: what is

$$E(t+1) = 2 * E(t)$$

- **A: over-doubled!**
  - But obeying the Densification Power Law

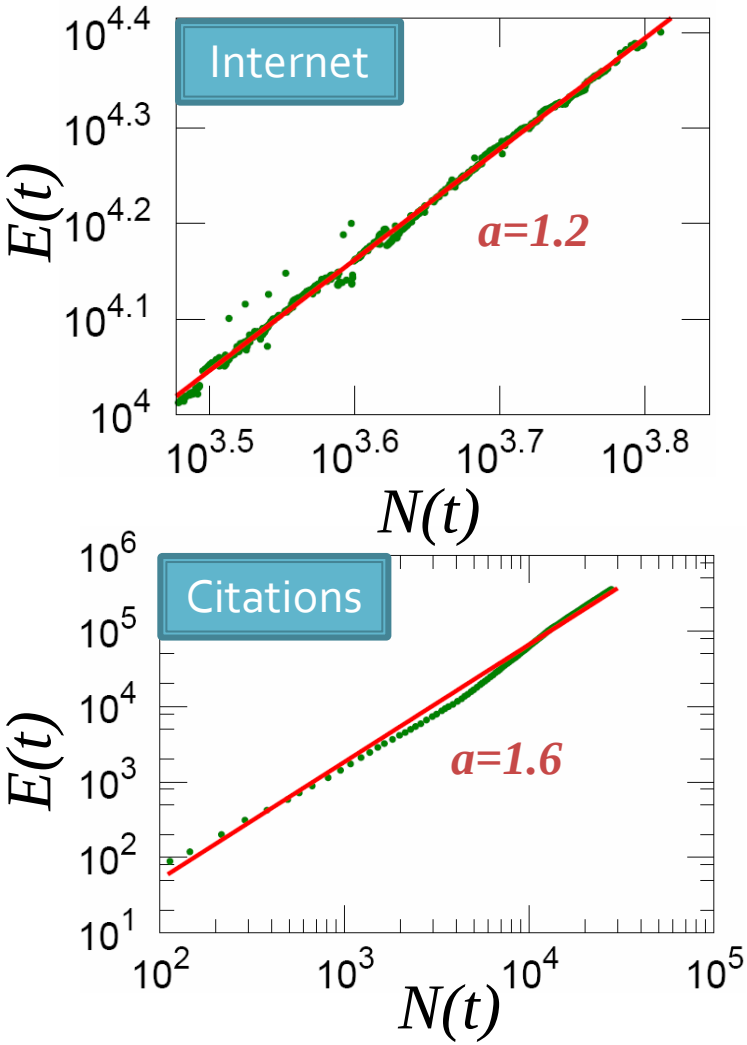
[Leskovec et al., KDD 05]

Q1) Network Evolution

- What is the relation between the number of nodes and the edges over time?
- ~~First guess: constant average degree over time~~
- Networks are denser over time
- **Densification Power Law:**

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

a … densification exponent ($1 \le a \le 2$)



[Leskovec et al. KDD 05]

Densification Power Law

- **Densification Power Law**
  - the number of edges grows faster than the number of nodes – average degree is increasing

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

or equivalently

$$\frac{\log(E(t))}{\log(N(t))} = const$$

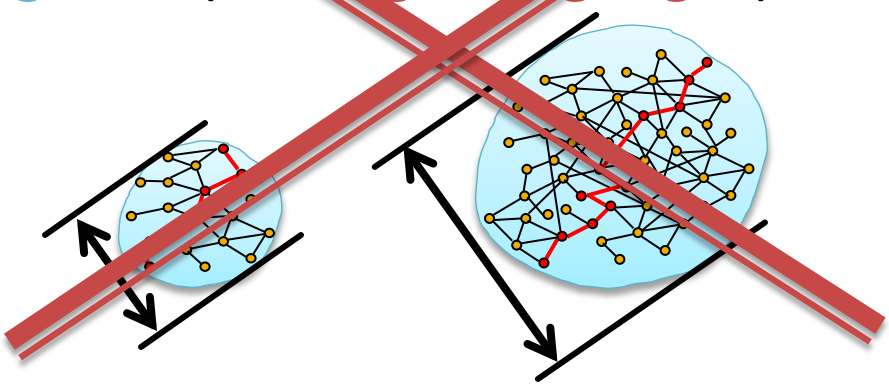
a … densification exponent: 1 ≤ a ≤ 2:

- a=1: linear growth – constant out-degree (traditionally assumed)
- a=2: quadratic growth – clique

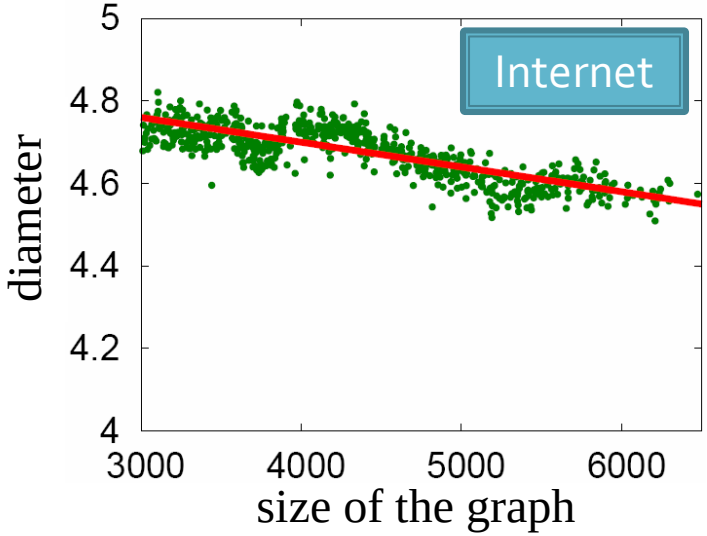
[Leskovec et al. KDD 05]

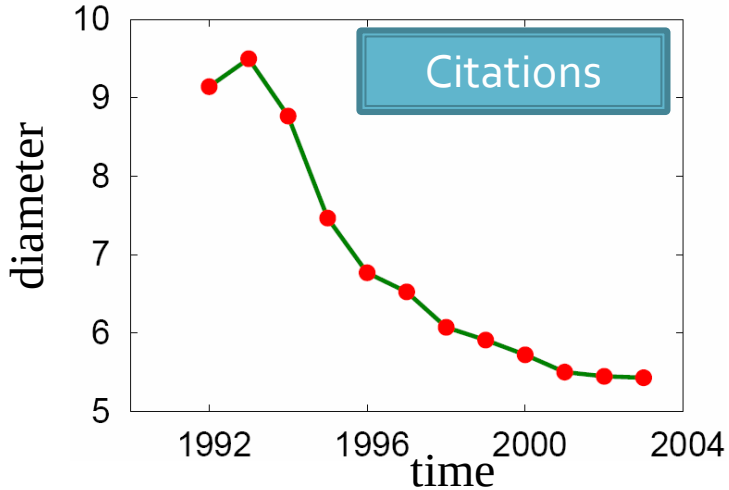
Q1) Network Evolution

- ~~Prior models and intuition say that the network diameter slowly grows (like log N, log log N)~~



- **Diameter shrinks over time**
  - as the network grows the distances between the nodes slowly **decrease**

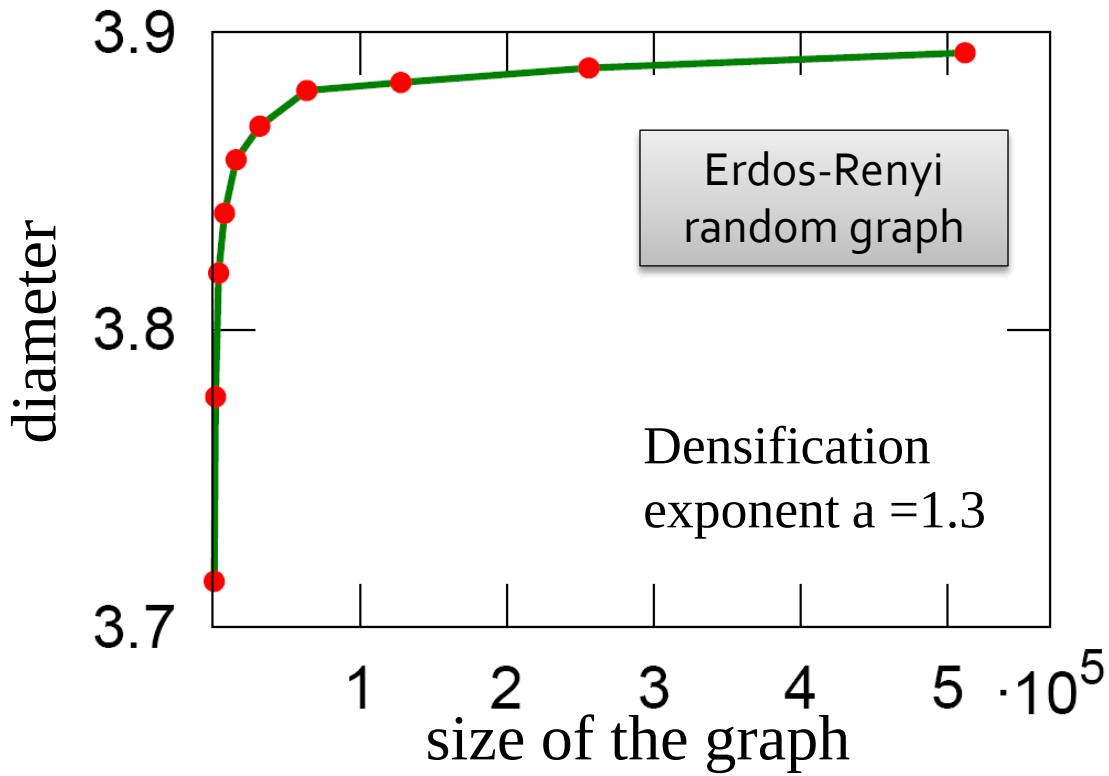




[Leskovec et al. TKDD 07]

Diameter of a Densifying G_{np}

Is shrinking diameter just a consequence of densification?



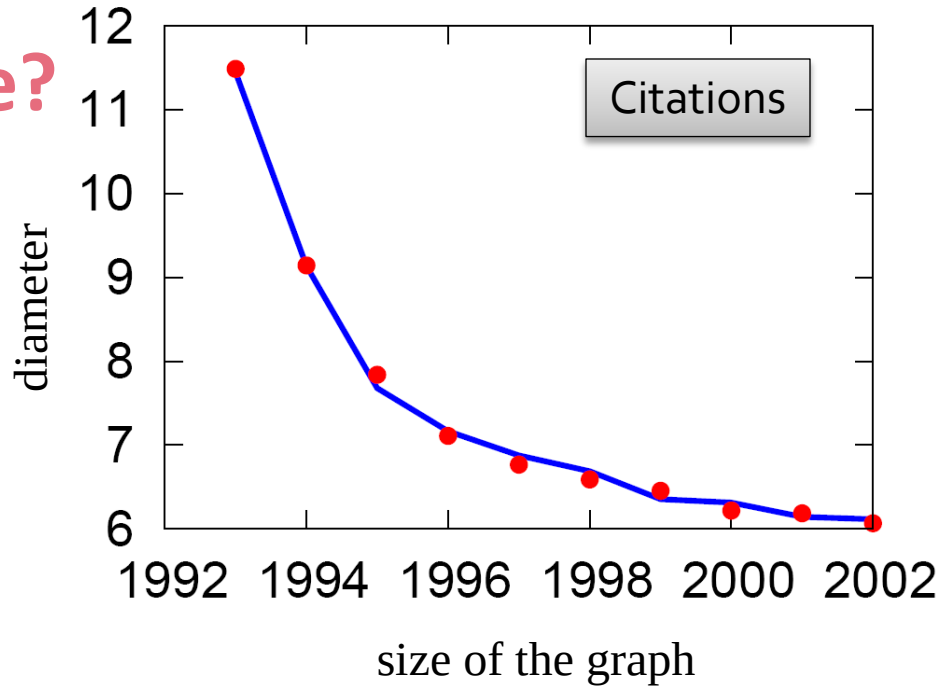
Densifying random graph has increasing diameter⇒ There is more to shrinking diameter than just densification

Diameter of a Rewired Network

Is it the degree sequence?

Compare diameter of a:

- True network (red)
- Random network with the same degree distribution (blue)



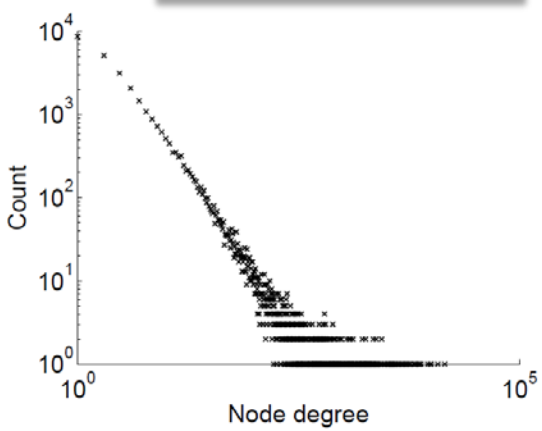
Densification + degree sequence give shrinking diameter

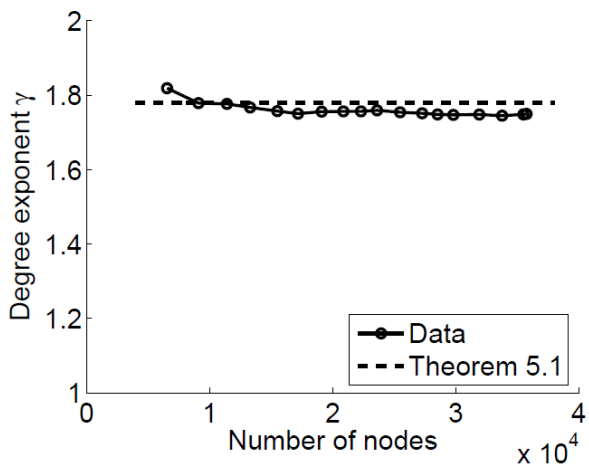
[Leskovec et al. TKDD 07]

Connecting Degrees & Densification

- **How does degree distribution evolve to allow for densification?**
- **Option 1)** Degree exponent γ_n is constant:
  - Fact 1: For degree exponent $1<\gamma<2$: $a = 2/\gamma$

Email network





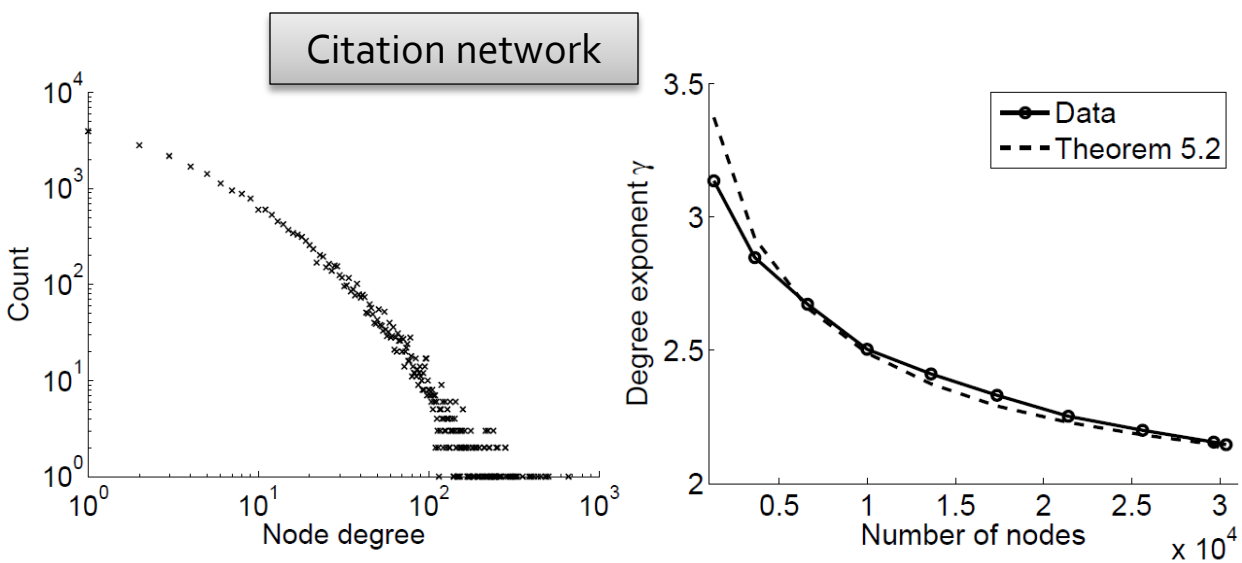
A consequence of what we learned in last class:

- Power-laws with exponents <2 have infinite expectations.
- So, to maintain constant degree exponent γ distribution with the average degree needs to grow.

[Leskovec et al. TKDD 07]

Connecting Degrees & Densification

- **How does degree distribution evolve to allow for densification?**
- **Option 2)** Exponent γ_n evolves with graph size n:
  - Fact 2: $$\gamma_n = \frac{4n^{a-1} - 1}{2n^{a-1} - 1}$$



Remember, expected degree is:

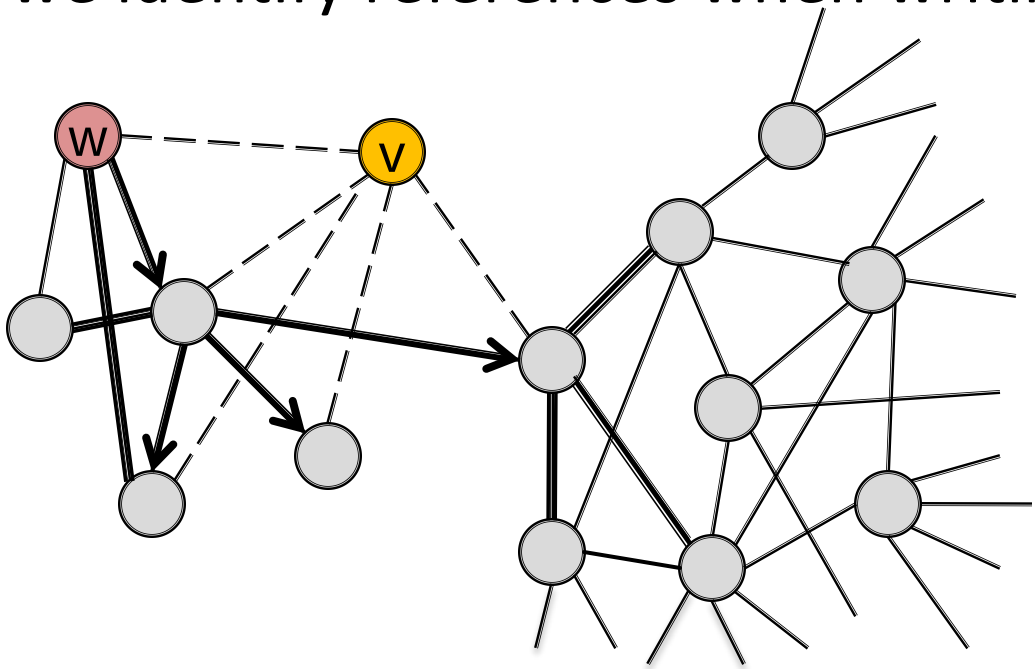
$$E[x] = \frac{\gamma - 1}{\gamma - 2} x_{min}$$

So γ has to decay as as function of graph size for the avg. degree to go up

[Leskovec et al. TKDD 07]

Forest Fire Model

- **Want to model graphs that density and have shrinking diameters**
- **Intuition:**
  - How do we meet friends at a party?
  - How do we identify references when writing papers?



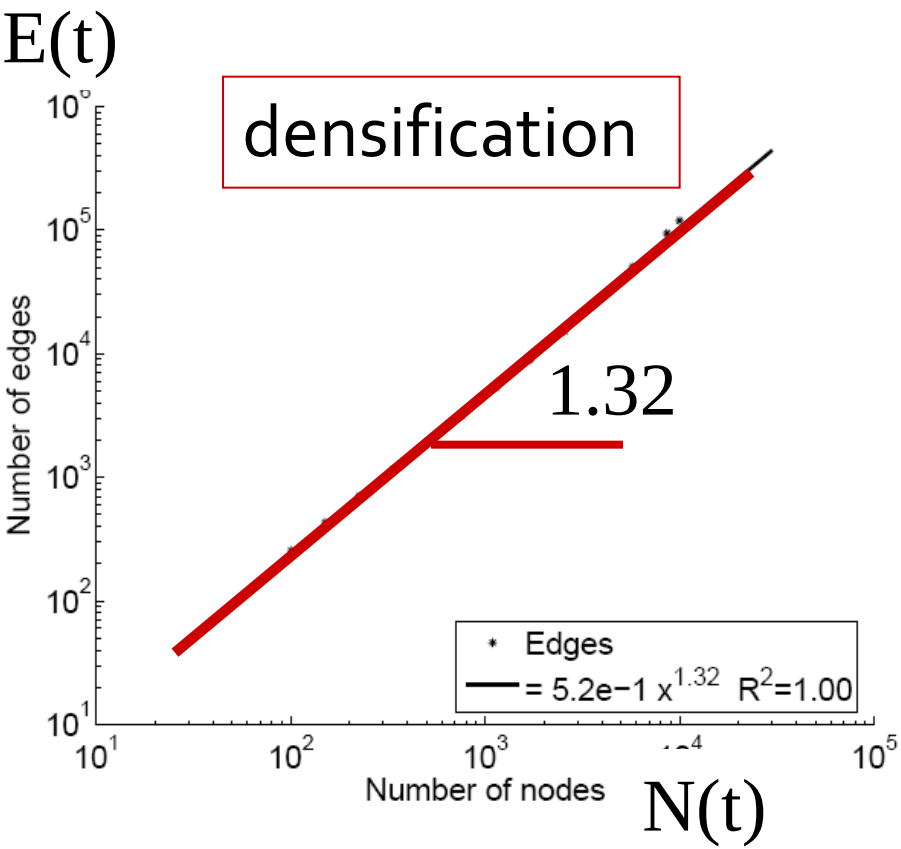
[Leskovec et al. TKDD 07]

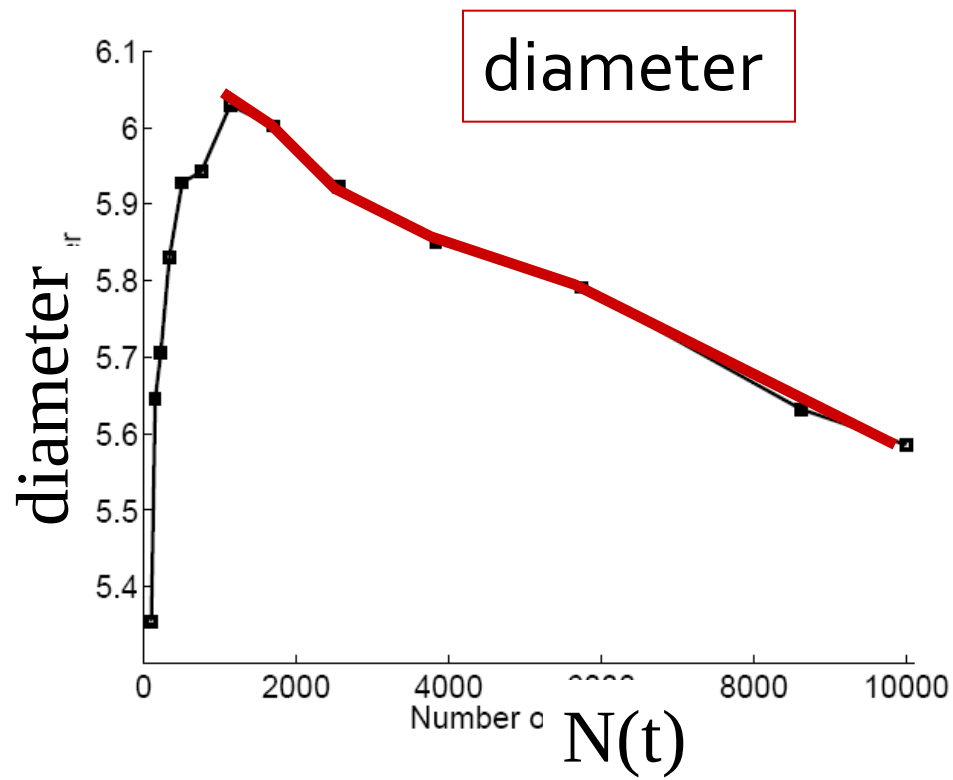
Forest Fire Model

- **The Forest Fire model has 2 parameters:**
  - **p** … forward burning probability
  - **r** … backward burning probability
- **The model:**
  - Each turn a new node *v* arrives
  - Uniformly at random chooses an "ambassador" *w*
  - Flip 2 geometric coins to determine the number of in- and out-links of *w* to follow
  - "Fire" spreads recursively until it dies
  - New node *v* links to all burned nodes

Forest Fire Model

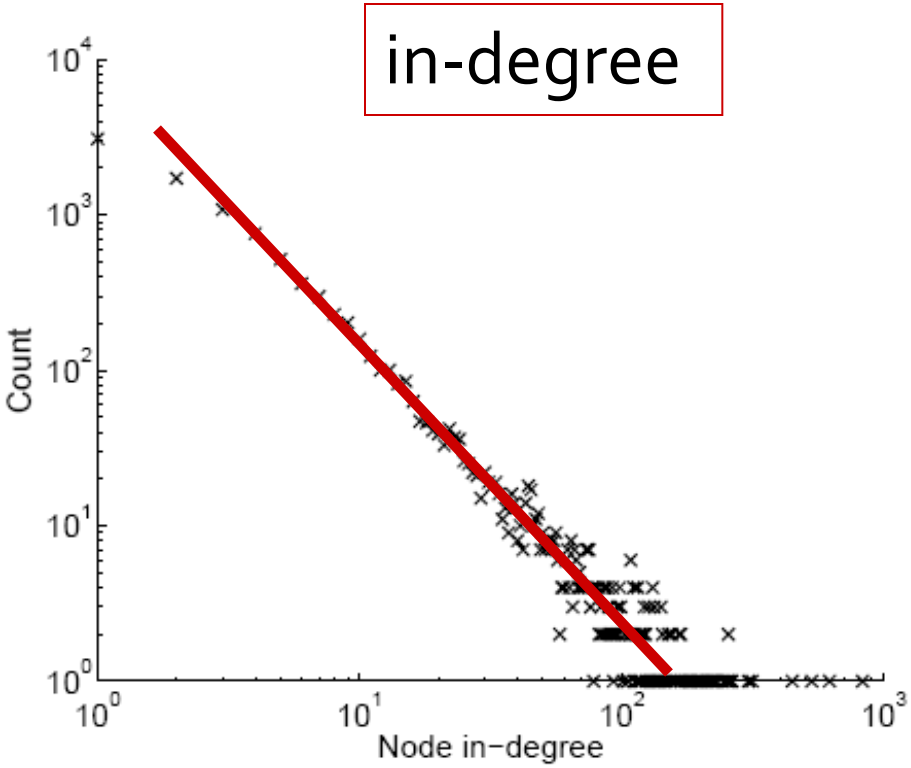
- **Forest Fire generates graphs that densify and have shrinking diameter**



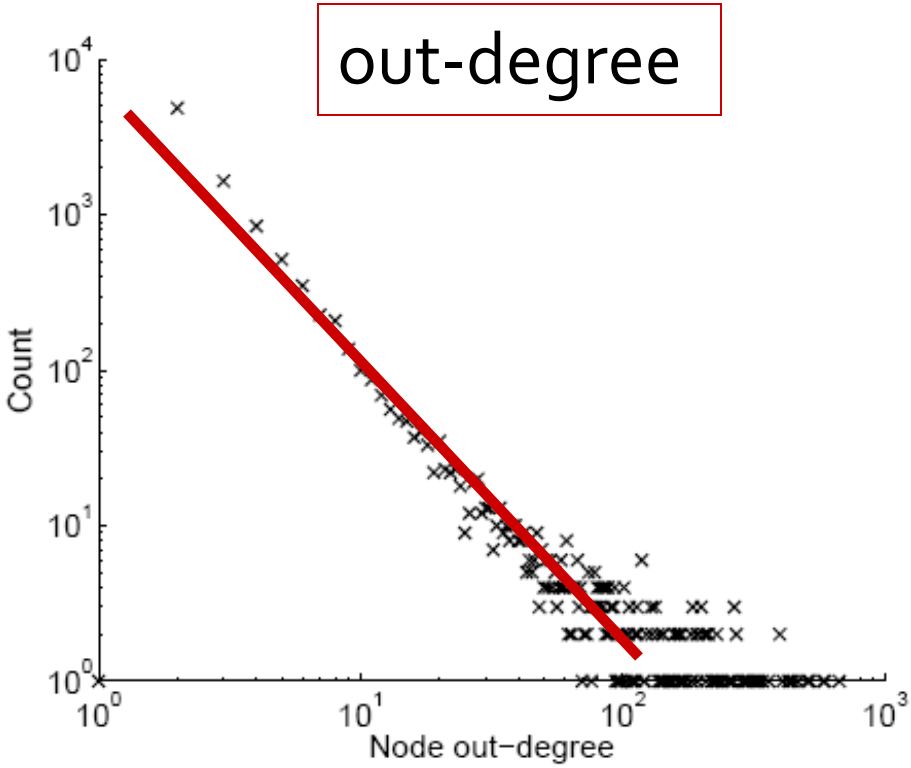


Forest Fire Model

- **Forest Fire also generates graphs with power-law degree distribution**



log count vs. log in-degree



log count vs. log out-degree

Forest Fire: Phase Transition

- Fix backward probability *r* and vary forward burning prob. *p*
- Notice a sharp transition between sparse and clique-like graphs
- **Sweet spot is very narrow**