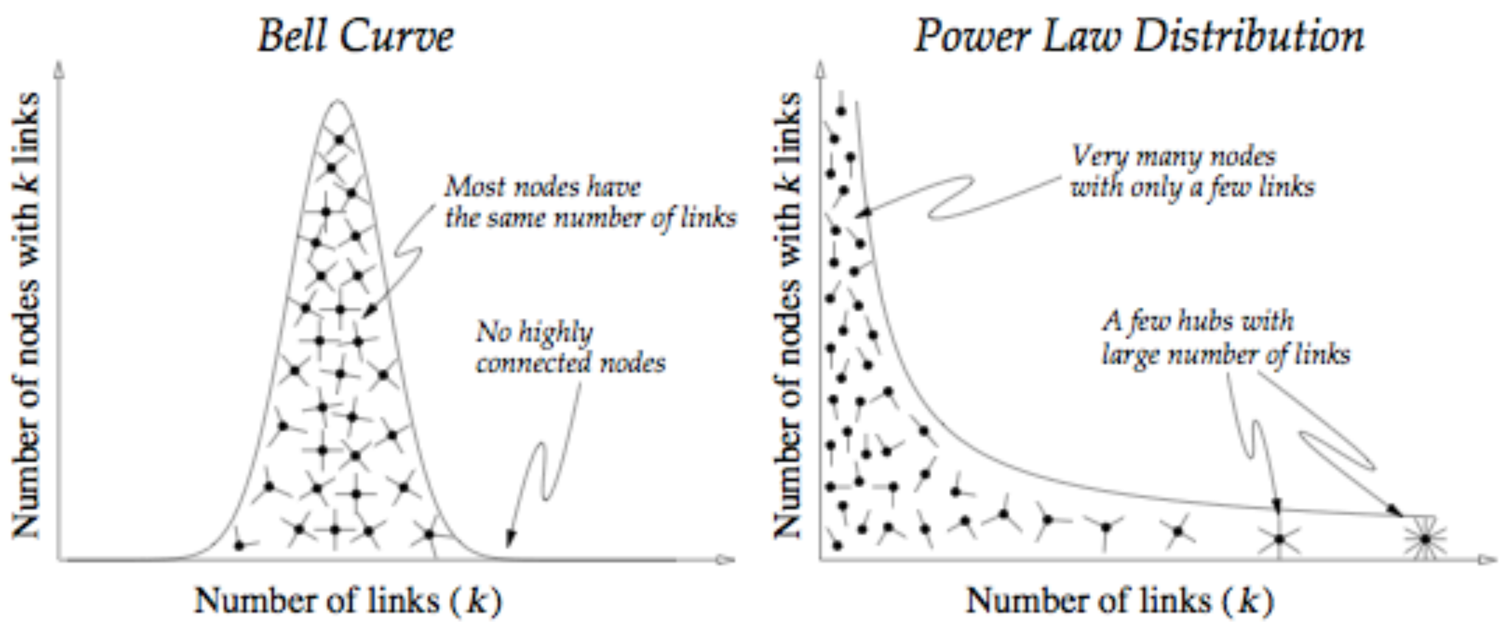
Preferential Attachment and Network Evolution

CS224W: Analysis of Networks
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

Exponential vs. Power-Law Tails

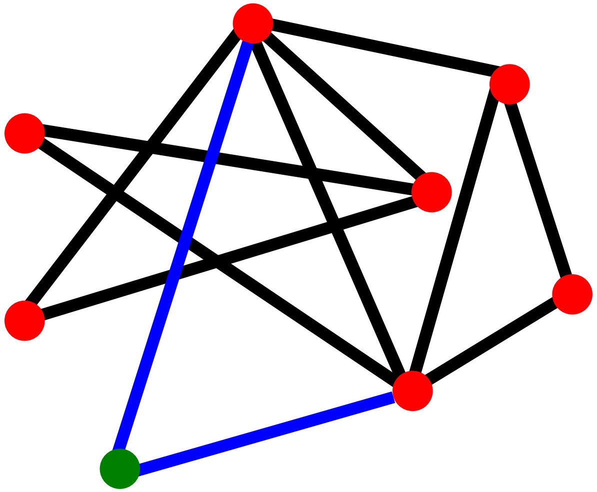


Model: G_{np} **?**

Model: Preferential attachment

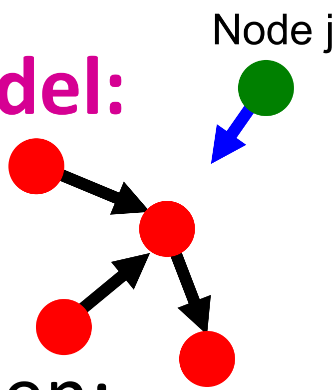
- **Preferential attachment:**
  [de Solla Price '65, Albert-Barabasi '99, Mitzenmacher '03]
  - Nodes arrive in order **1,2,…,n**
  - At step j, let d_i be the degree of node $i < j$
  - A new node j arrives and creates m out-links
  - Prob. of j linking to a previous node i is **proportional to degree d_i of node i**

$$P(j \rightarrow i) = \frac{d_i}{\sum_k d_k}$$



[Mitzenmacher, '03]

The Exact Model



We will analyze the following simple model:

- Nodes arrive in order $1,2,3, \dots, n$
- When node j is created it makes a **single out-link** to an earlier node i chosen:
  - **1)** With prob. p, j links to i chosen **uniformly at random** (from among all earlier nodes)
  - **2)** With prob. $1-p$, node j chooses i uniformly at random & links **to a random node *l* *that i* points to**
    - **This is same as saying:** With prob. $1-p$, node j links to node l with prob. proportional to d_l (the in-degree of l)
  - **Our graph is directed:** Every node has out-degree 1

The Model Gives Power-Laws

- **Claim:** The described model generates networks where the fraction of nodes with in-degree $\boldsymbol{k}$ scales as:

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

where q=1-p

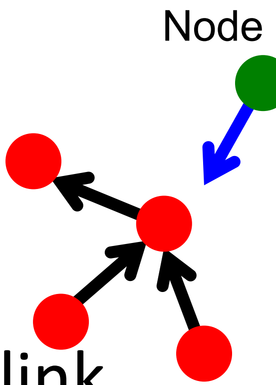
So we get power-law degree distribution with exponent:

$$\alpha = 1 + \frac{1}{1-p}$$

Continuous Approximation

- **Consider deterministic and continuous approximation** to the degree of node $\boldsymbol{i}$ as a function of time $\boldsymbol{t}$
  - $\boldsymbol{t}$ is the number of nodes that have arrived so far
  - **In-Degree $d_i(t)$** of node $\boldsymbol{i}$ ($i = 1,2, \dots, n$) is a **continuous quantity** and it **grows deterministically** as a function of time $\boldsymbol{t}$
- **Plan: Analyze $d_i(t)$ – continuous in-degree** of node $\boldsymbol{i}$ at time $\boldsymbol{t}$ $(t > i)$
  - **Note: Node i arrives to the graph at time i**

Continuous Degree: What We Know



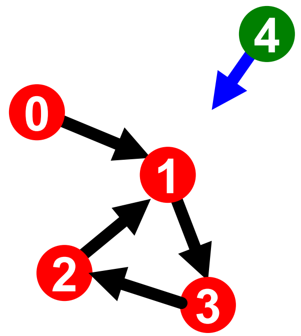
- **Initial condition:**
  - $d_i(t) = 0$, when $t = i$ **(node *i* just arrived)**
- **Expected change of $d_i(t)$ over time:**
  - Node i gains an in-link at step $t + 1$ only if a link from a newly created node $t + 1$ points to it
  - **What's the probability of this event?**
    - With prob. p node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

Note: each node creates exactly 1 edge. So after *t* nodes/steps there are *t* edges in total.

Continuous Degree

- At $t = 4$ node $i = 4$ comes. **It has out-degree of 1 to deterministically share with other nodes:**

Node i	$d_i(t)$	$d_i(t+1)$
0	0	$=0 + p\frac{1}{4} + (1-p)\frac{0}{4}$
1	2	$=2 + p\frac{1}{4} + (1-p)\frac{2}{4}$
2	0	$=0 + p\frac{1}{4} + (1-p)\frac{1}{4}$
3	1	$=1 + p\frac{1}{4} + (1-p)\frac{1}{4}$
4	/	0



- $d_i(t) - d_i(t-1) = \frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = \mathbf{p}\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
- **How does $d_i(t)$ evolve as $t \to \infty$?**

What is the rate of growth of d_i?

- **Expected change of $d_i(t)$:**
  - $\underbrace{d_i(t+1) - d_i(t)} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t}$
  - $\frac{\mathrm{d}d_i(t)}{\mathrm{d}t} = p\frac{1}{t} + (1-p)\frac{d_i(t)}{t} = \frac{p+qd_i(t)}{t}$ $\quad q = (1-p)$
  - $\frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \frac{1}{t}\mathrm{d}t$ $\quad$ Divide by $p + q\,d_i(t)$
  - $\int \frac{1}{p+qd_i(t)}\mathrm{d}d_i(t) = \int \frac{1}{t}\mathrm{d}t$ $\quad$ integrate
  - $\frac{1}{q}\ln\left(p + qd_i\,(t)\right) = \ln t + c$ $\quad$ Exponentiate and let $A = e^c$
  - $p + qd_i(t) = e^{qc}\ t^q \ \Rightarrow \boldsymbol{d_i(t) = \frac{1}{q}((At)^q - p)}$ $\quad$ **A=?**

What is the constant A?

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

What is the value of constant A?

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

Observation: Old nodes (small i values) have higher in-degrees $d_i(t)$

Degree Distribution

- **What is $F(k)$, the fraction of nodes that has degree <u>less than</u> k at time t?**
  - **How many nodes have degree < k?**
    - $d_i(t) = \frac{p}{q}\left(\left(\frac{t}{i}\right)^q - 1\right) < k$
    - Solve for i and obtain: $i > t\left(\frac{q}{p}k + 1\right)^{-\frac{1}{q}}$
- **There are t nodes total at time t so the fraction $F(k)$ is:**

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

Degree Distribution

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

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

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

q.e.d.

Preferential attachment: Good news

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

Preferential Attachment: Bad News

- **Preferential attachment is not so good at predicting network structure**
  - **Age-degree correlation**
    - **Solution:** Node fitness (virtual degree)
  - **Links among high degree nodes:**
    - On the web nodes sometimes avoid linking to each other
- **Further questions:**
  - **What is a reasonable model for how people sample network nodes and link to them?**
    - Short random walks

Many models lead to Power-Laws

- **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 and attach to both endpoints of this edge
- **Node duplication**
  - Duplicate a node with all its edges
  - Randomly prune edges of new node

Preferential attachment: Reflections

Extra!

- **Two changes from the G_{np}**
  - The network grows
  - Preferential attachment
- **Do we need both? Yes!**
  - **Add growth to G_{np}** (assume $p = 1/n$):
    - $X_j =$ degree of node j at the end
    - $X_j(u) = 1$ if u links to j, else 0
    - $X_j = X_j(j+1) + X_j(j+2) + \cdots + X_j(n)$
    - $E[X_j(u)] = P[u \text{ links to } j] = 1/(u-1)$
    - $E[X_j] = \sum_{j+1}^{n} \frac{1}{u-1} = \frac{1}{j} + \frac{1}{j+1} + \cdots + \frac{1}{n-1} = H_{n-1} - H_j$
    - $E[X_j] \approx \log(n-1) - \log(j) = \log((n-1)/j)$ *NOT* $\left(\frac{n}{j}\right)^{\alpha}$

$H_n \ldots n$-th harmonic number: $= \sum_{k=1}^{n} \frac{1}{k}.$

Distances in Preferential Attachment

Extra!

$$\bar{h} = \begin{cases} const & \alpha = 2 \\ \dfrac{\log\log n}{\log(\alpha-1)} & 2 < \alpha < 3 \\ \dfrac{\log n}{\log\log n} & \alpha = 3 \\ \log n & \alpha > 3 \end{cases}$$

Ultra small world (for $2 < \alpha < 3$); Small world (for $\alpha > 3$)

Avg. path length | Degree exponent

Size of the biggest hub is of order *O(N)*. Most nodes can be connected within two steps, thus the average path length will be independent of the network size.

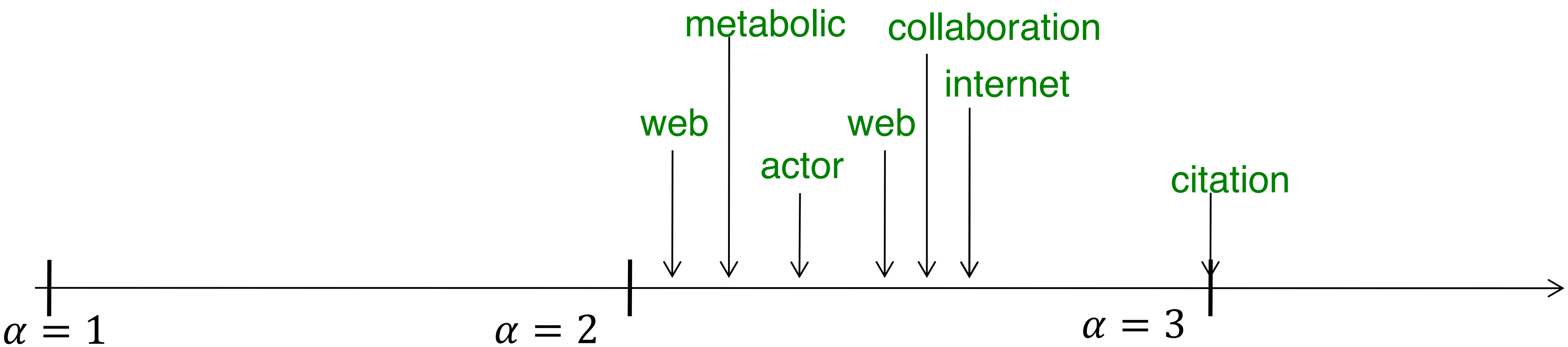
The average path length increases slower than logarithmically. In G_{np} all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network vast majority of the paths go through the few high degree hubs, reducing the distances between nodes.

Some models produce $\alpha = 3$. This was first derived by Bollobas et al. for the network diameter in the context of a dynamical model, but it holds for the average path length as well.

The second moment of the distribution is finite, thus in many ways the network behaves as a random network. Hence the average path length follows the result that we derived for the random network model earlier.

Summary: Scale-Free Networks

Extra!



Second moment $\langle k^2 \rangle$ diverges | $\langle k^2 \rangle$ finite

Average $\langle k \rangle$ diverges | Average $\langle k \rangle$ finite

Ultra small world behavior | Small world

Regime full of anomalies… | The scale-free behavior is relevant | Behaves like a random network

Evolution of Social Networks

Network Evolution: Observation

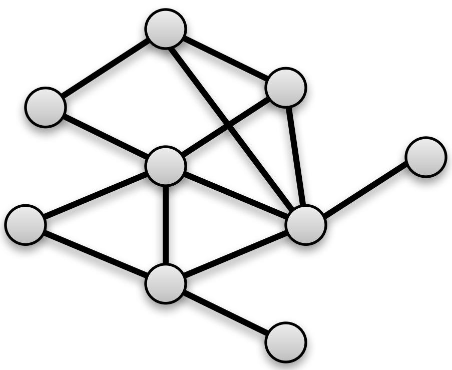
- **Preferential attachment is a model of a growing network**
- **Can we find a more realistic model?**
- **What governs network growth & evolution?**
  - **P1) Node arrival process:**
    - When nodes enter the network
  - **P2) Edge initiation process:**
    - Each node decides when to initiate an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

[Leskovec, Backstrom, Kumar, Tomkins, 2008]

[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 creation t_{uv}

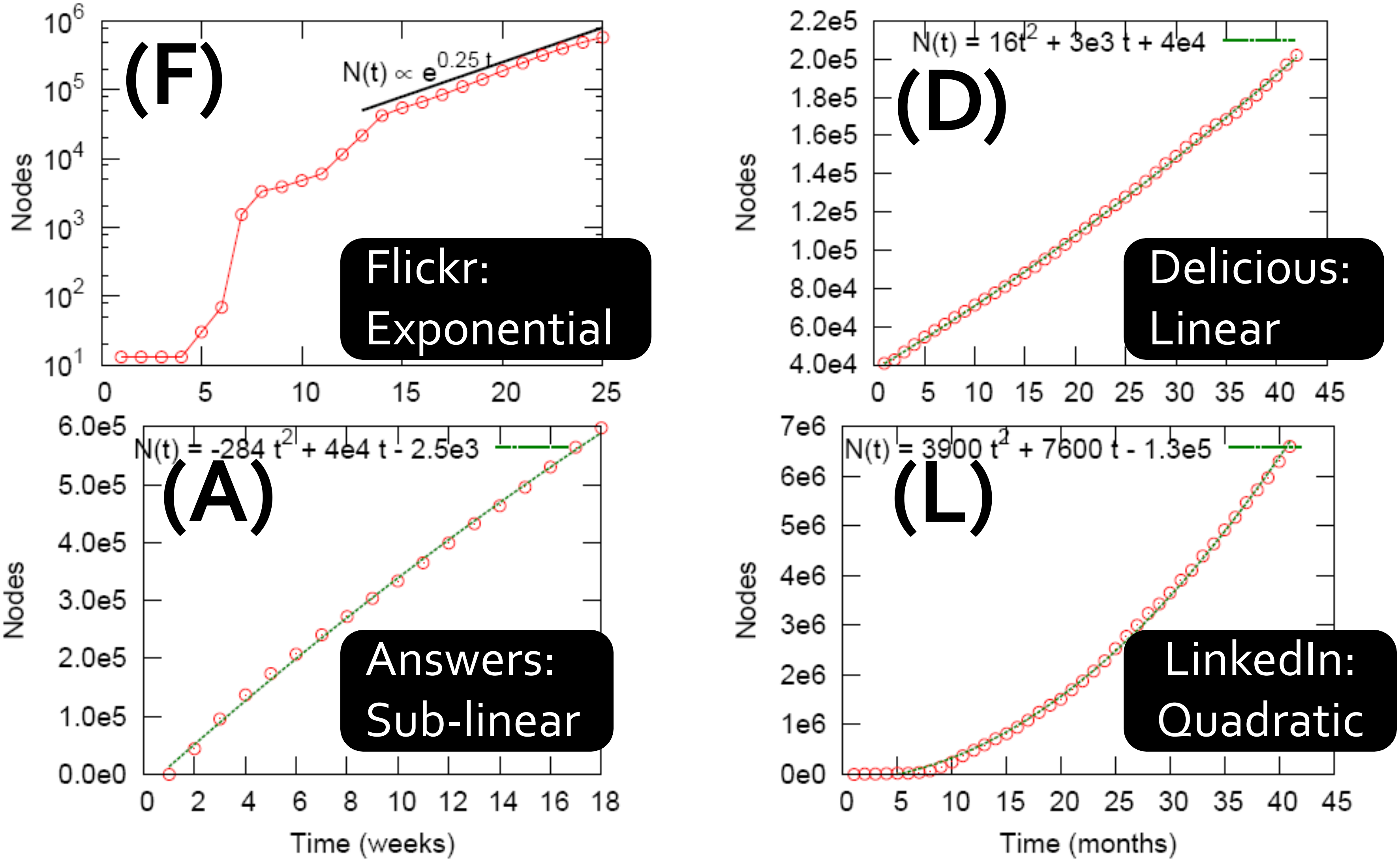


and so on for millions…

- **Directly observe mechanisms leading to global network properties**

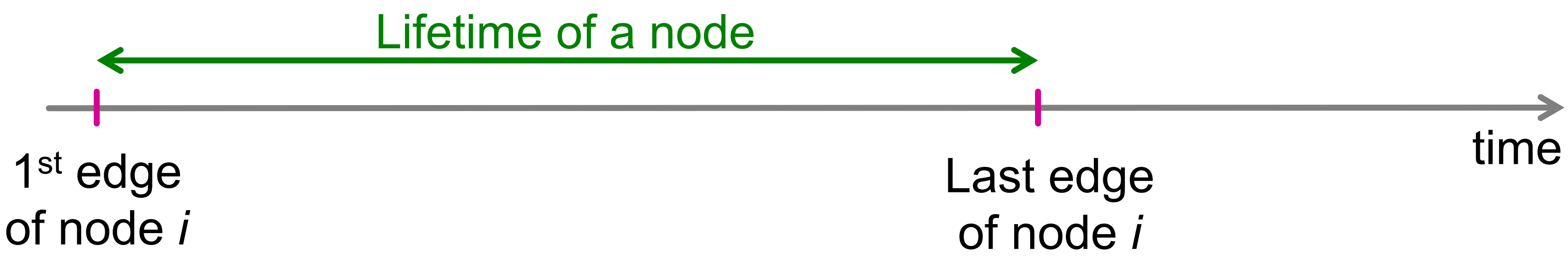
	Network	T	N	E
(F)	Flickr (03/2003–09/2005)	621	584,207	3,554,130
(D)	Delicious (05/2006–02/2007)	292	203,234	430,707
(A)	Answers (03/2007–06/2007)	121	598,314	1,834,217
(L)	LinkedIn (05/2003–10/2006)	1294	7,550,955	30,682,028

P1) When are New Nodes Arriving?

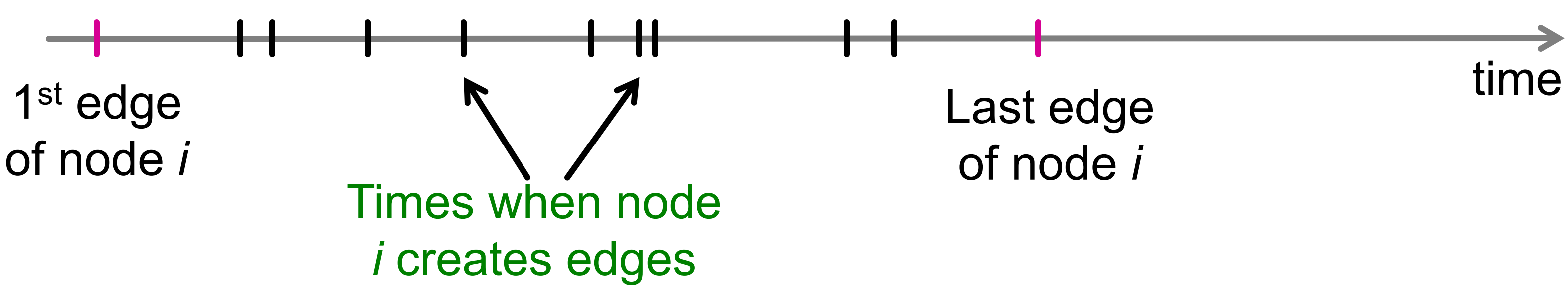


P2) When Do Nodes Create Edges?

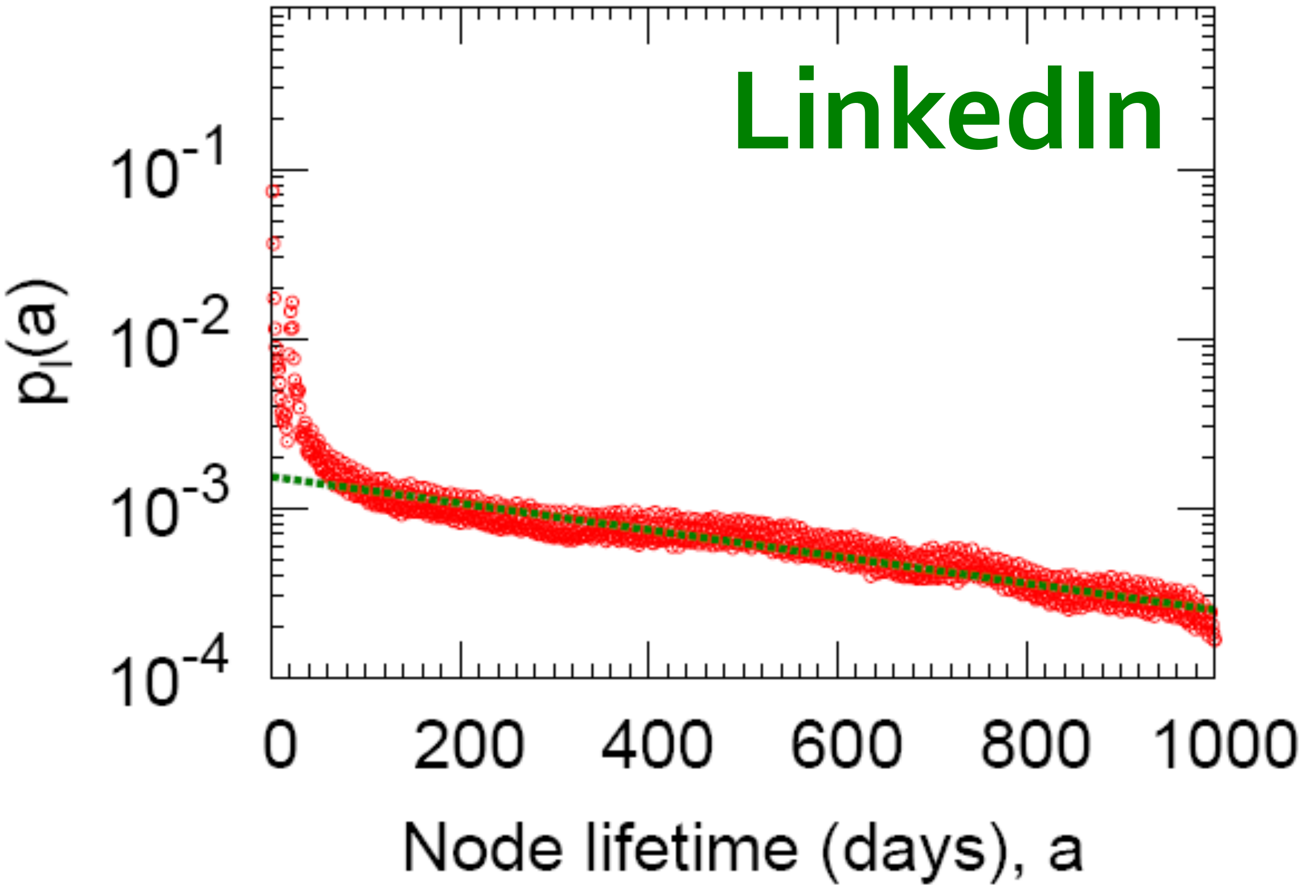
- **How long do nodes live?**
  - Node life-time is the time between the 1st and the last edge of a node



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



P2) What is Node Lifetime?



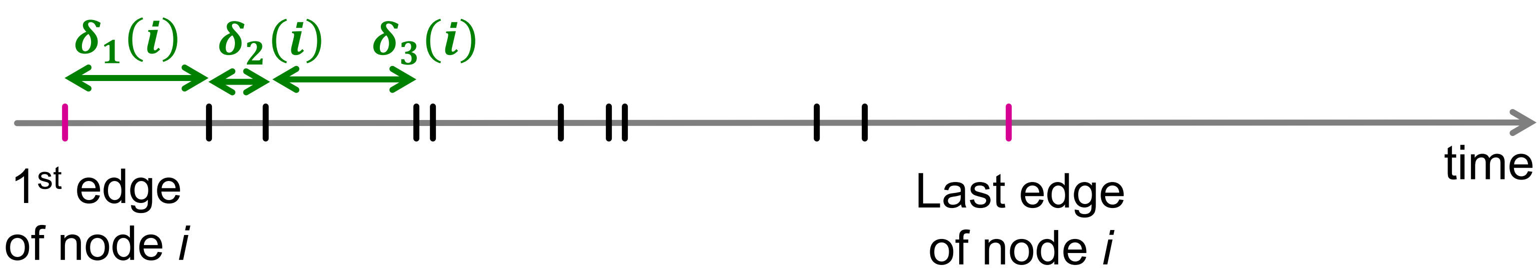
- **Lifetime *a*:** Time between node's first and last edge

Node lifetime is **exponentially distributed**:

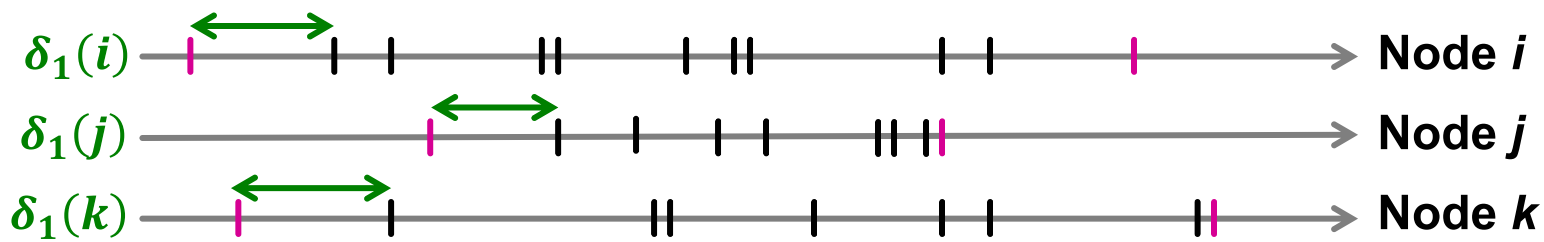
$$p_l(a) = \lambda e^{-\lambda a}$$

P2) When do Nodes Create Edges?

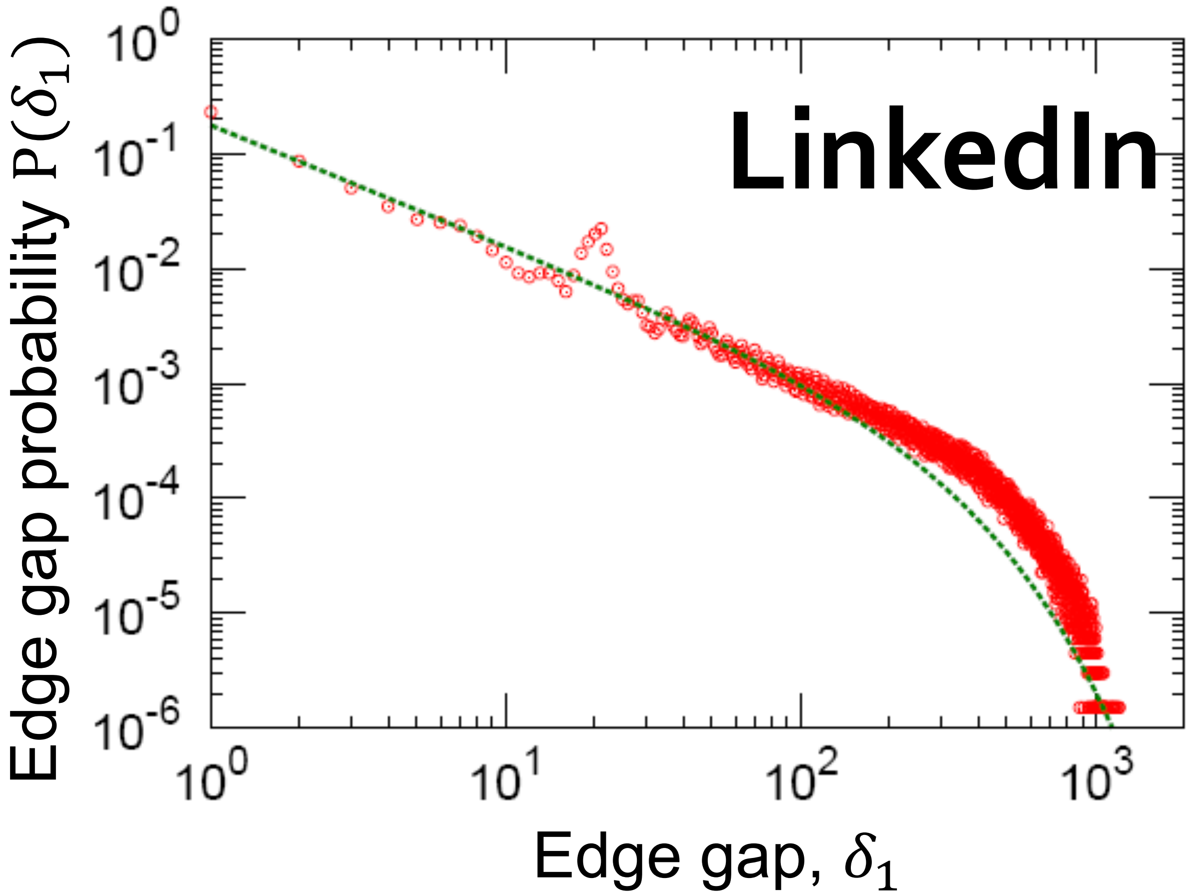
- **How do nodes "wake up" to create edges?**
  - **Edge gap $\delta_d(i)$**: time between $\boldsymbol{d^{th}}$ and $\boldsymbol{d+1^{st}}$ edge of node $\boldsymbol{i}$:
    - Let $\boldsymbol{t_d(i)}$ be the creation time of $\boldsymbol{d}$-th edge of node $\boldsymbol{i}$
    - $\boldsymbol{\delta_d(i)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



    - **δ_d is a distribution (histogram) of $\delta_d(i)$ over all nodes i**



P2) When do Nodes Create Edges?



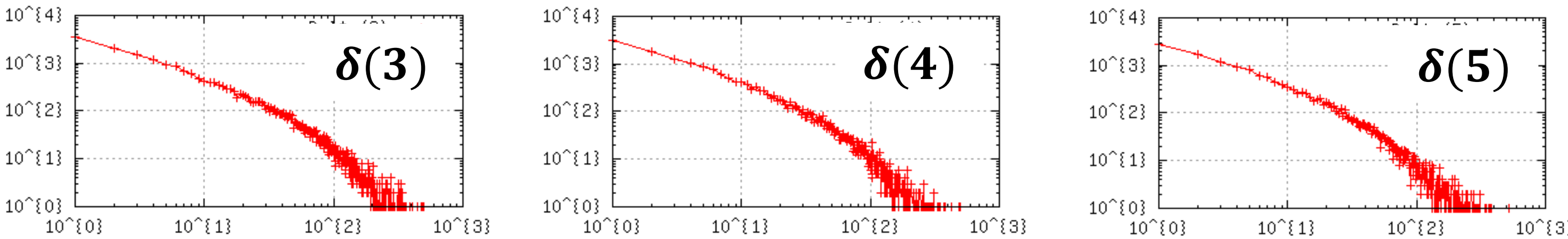
$$p_g(\delta_1) \propto \delta_1^{-\alpha} e^{-\beta \delta_1}$$

Edge gap δ_d: inter-arrival time between d^{th} and $d+1^{st}$ edge **is distributed by a power-law with exponential cut-off**

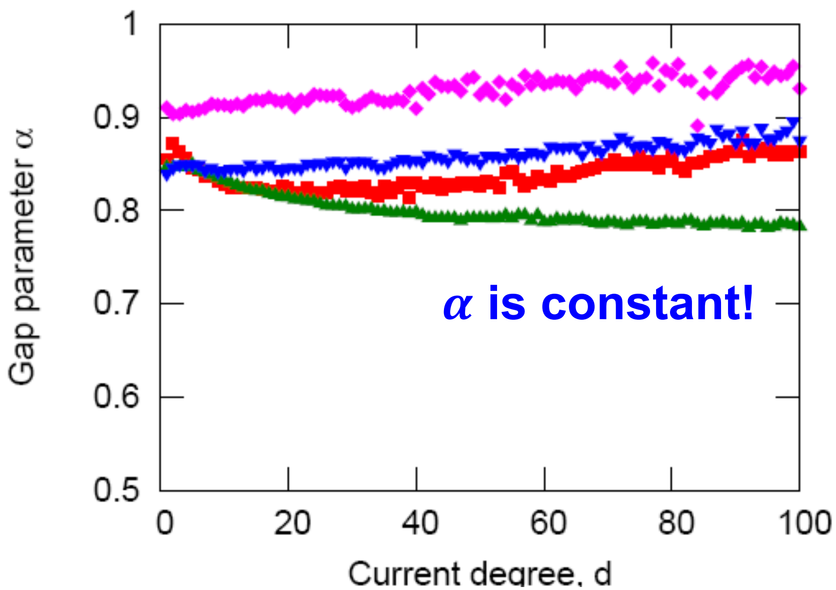
For every ***d*** we make a separate histogram

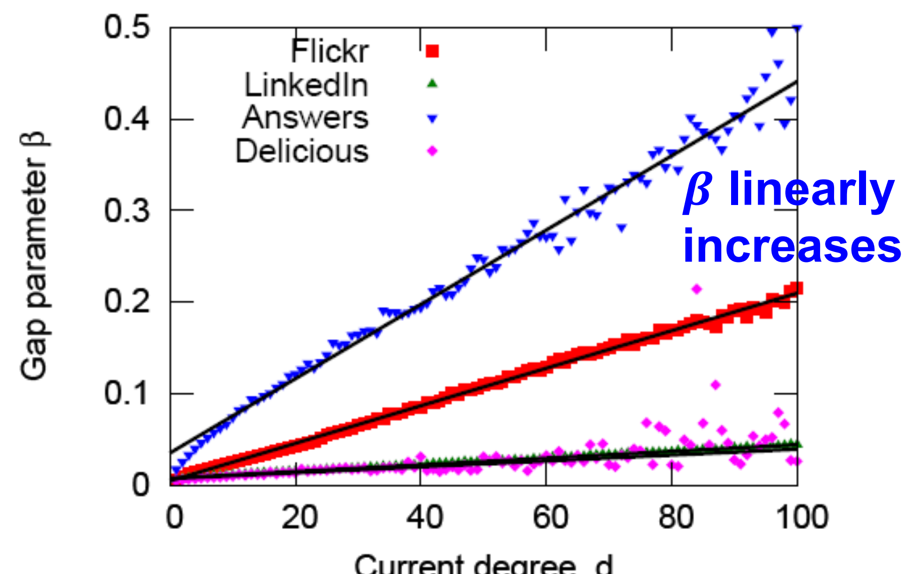
P2) How do α and β evolve with *d*?

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



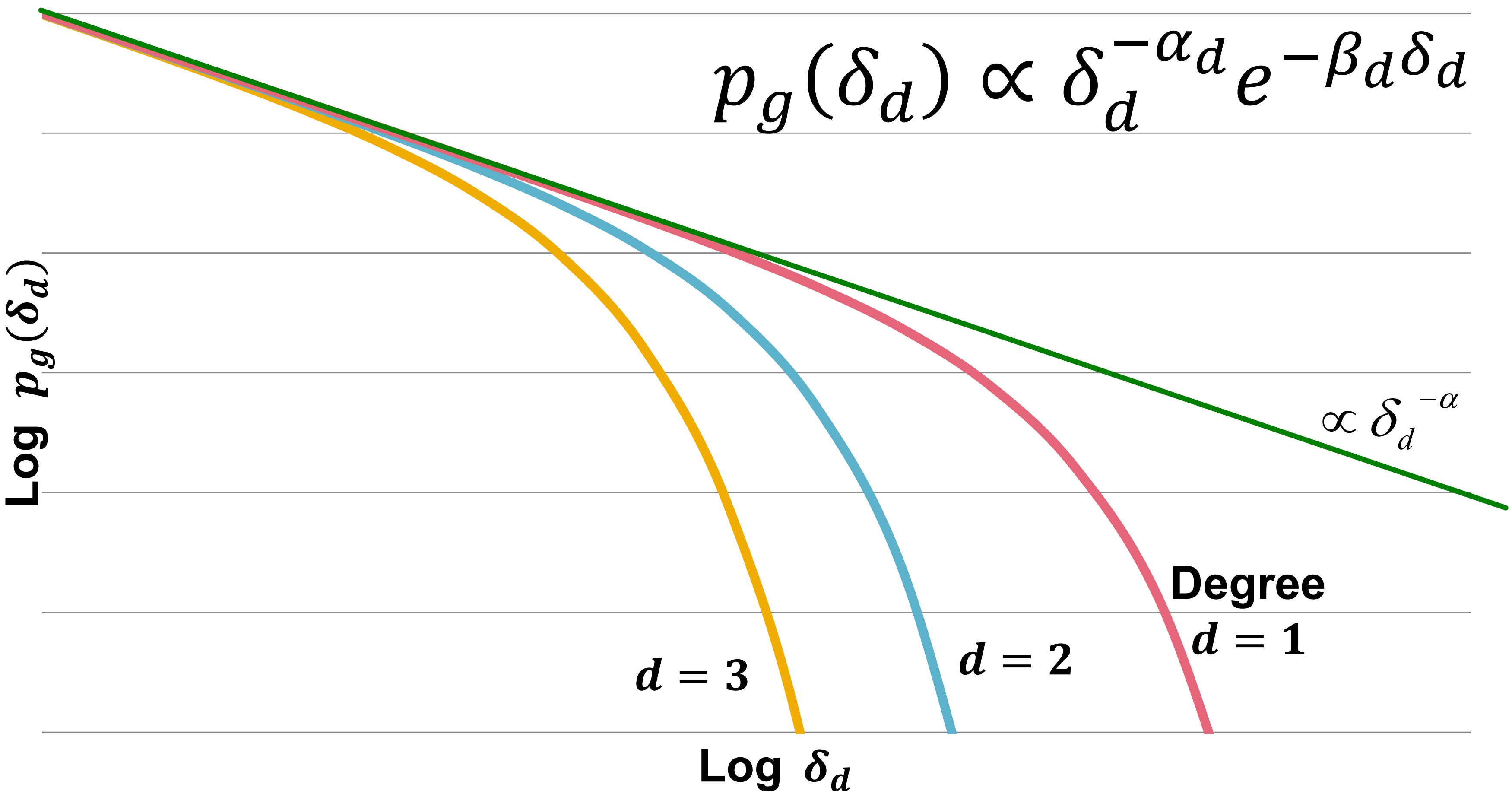
To each plot of δ_d fit: $p_g(\delta_d) \propto \delta_d^{-\alpha_d} e^{-\beta_d \delta_d}$





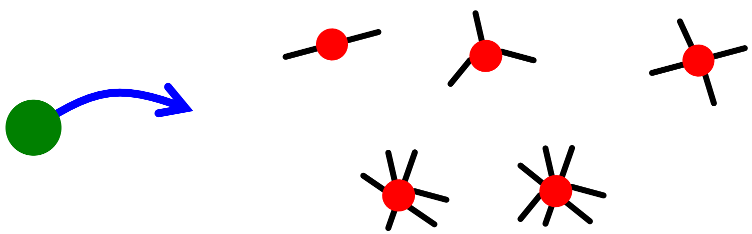
P2) Evolution of Edge Gaps

- α const., β linear in d. What does this mean?
- Gaps get smaller with d!

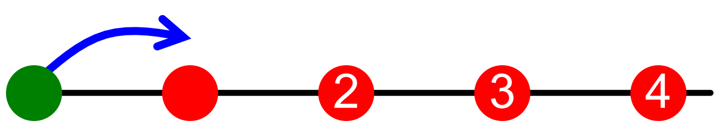


P3) How to Select Destination?

- **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*?**
    - **Does preferential attachment really hold?**

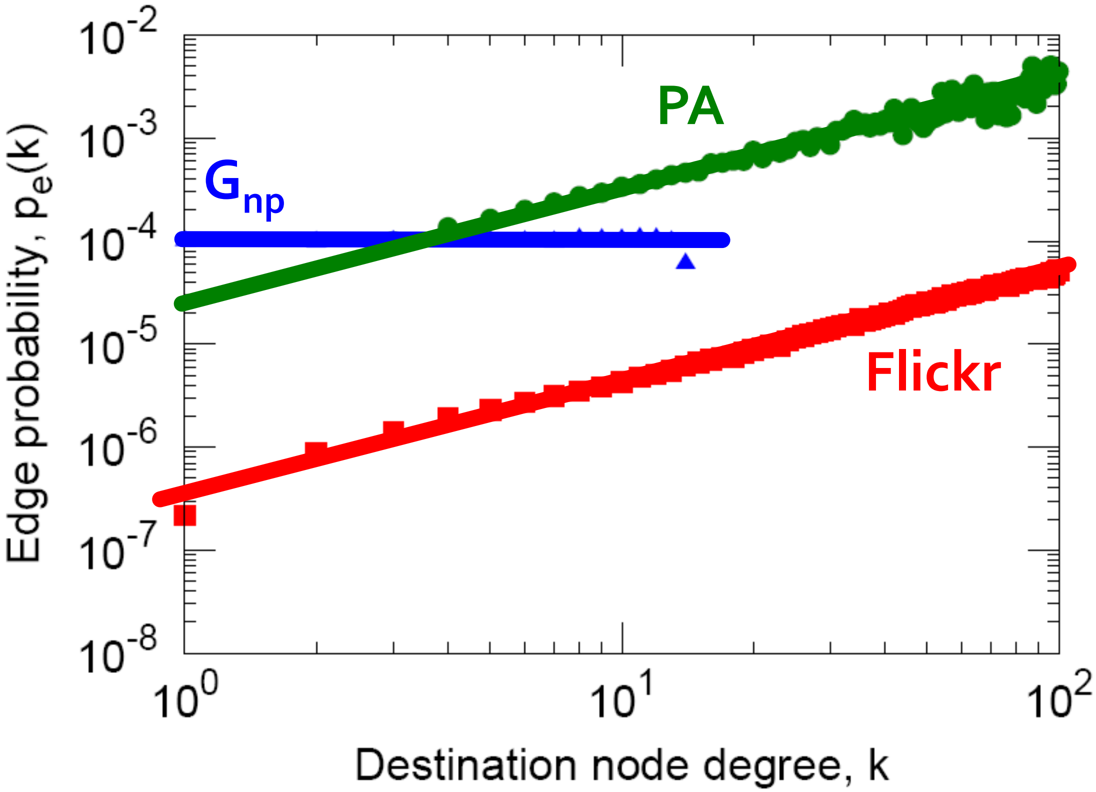


  - **How many hops away is the target *j*?**
    - **Are edges attaching locally?**



Edge Attachment Degree Bias

- **Are edges more likely to connect to higher degree nodes? YES!**



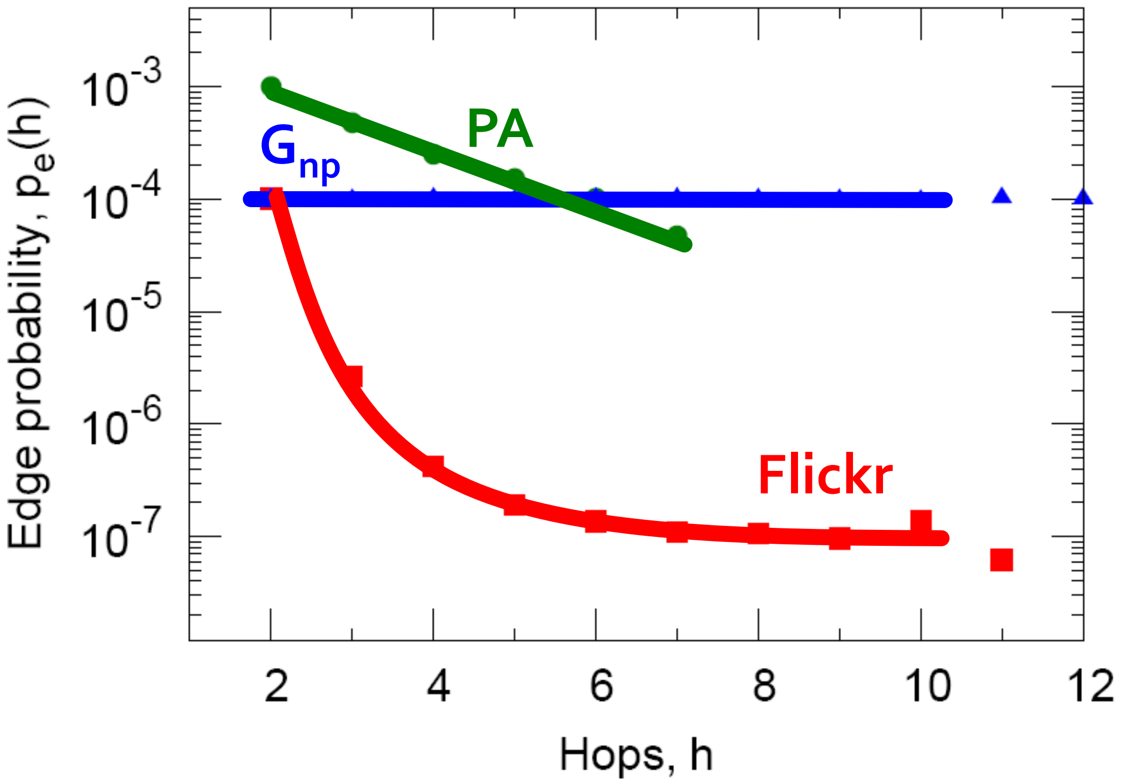
$$p_e(k) \propto k^{\tau}$$

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*?**



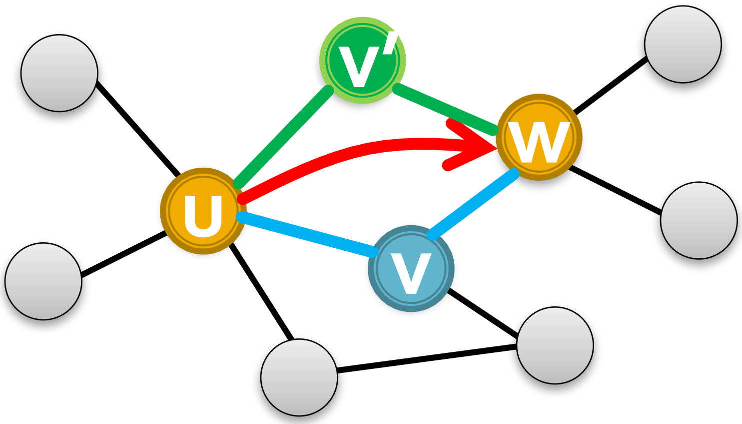
Fraction of triad closing edges

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**
- **2 step walk model:**
  - **u is about to create an edge**
  1. ***u*** choses neighbor ***v***
  2. ***v*** choses neighbor ***w*** and **u** connects to ***w***



- One can use different strategies for choosing ***v*** and ***w***: **Random-Random works well. Why?**
  - More common friends (more paths) helps
  - High-degree nodes are more likely to be hit

Triad Closing Strategies

Extra!

- **Improvement in log-likelihood over baseline:**
  - **Baseline:** Pick a random node 2 hops away

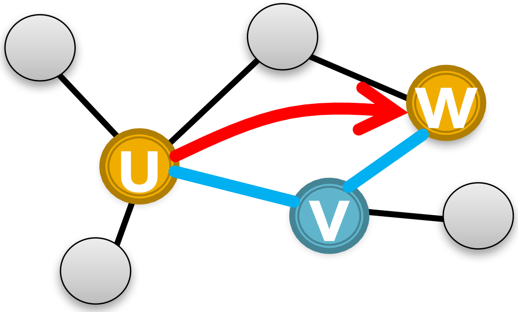
Strategy to select *v* (1st node)

Select *w* (2nd node)

FLICKR	random	$\text{deg}^{0.2}$	com	$\text{last}^{-0.4}$	$\text{comlast}^{-0.4}$
random	13.6	13.9	14.3	16.1	15.7
$\text{deg}^{0.1}$	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**



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

- **Comments:**
  - The proof is based on a combination of exponentials
  - Interesting as temporal behavior predicts a structural network property

Evolving the Networks

- **Given the model one can take an existing network and 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

degree exponent

Macroscopic Evolution of Networks

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?

Network Evolution

- $\boldsymbol{N(t)}$ … nodes at time $\boldsymbol{t}$
- $\boldsymbol{E(t)}$ … edges at time $\boldsymbol{t}$
- Suppose that

  $$\boldsymbol{N(t+1) = 2 \cdot N(t)}$$

- **Q:** what is:

  $\boldsymbol{E(t+1) = ?}$ **Is it $\boldsymbol{2 \cdot E(t)?}$**

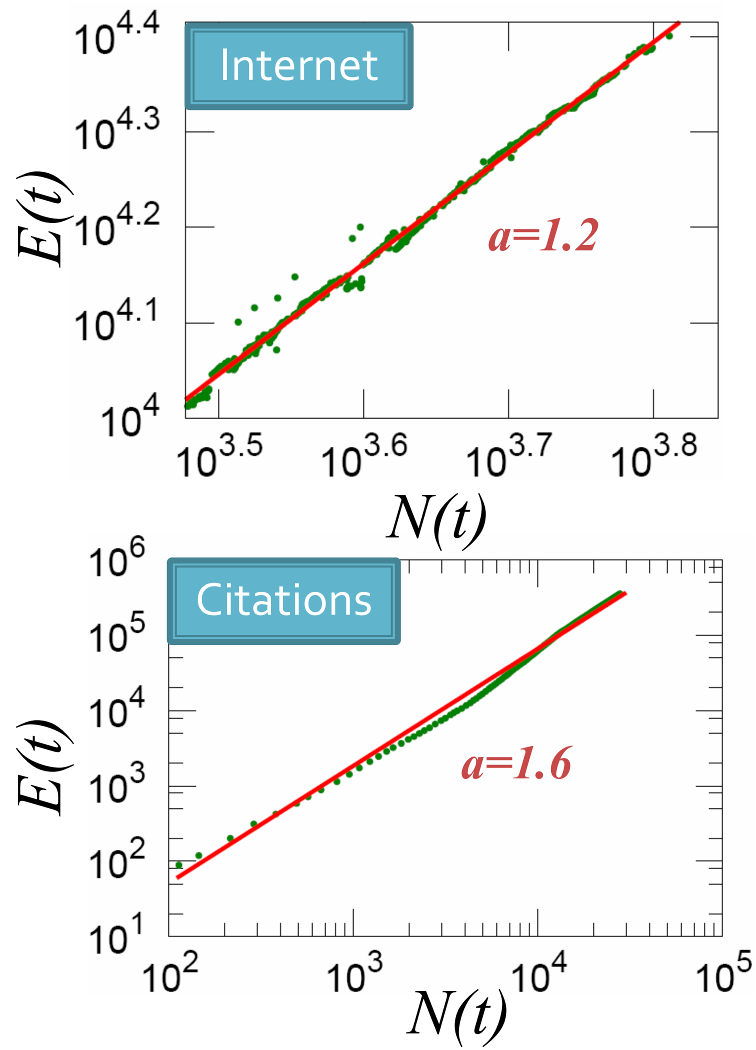
- **A: More than doubled!**
  - But obeying the **Densification Power Law**

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 ≤ **a** ≤ 2)



Densification Power Law

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

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

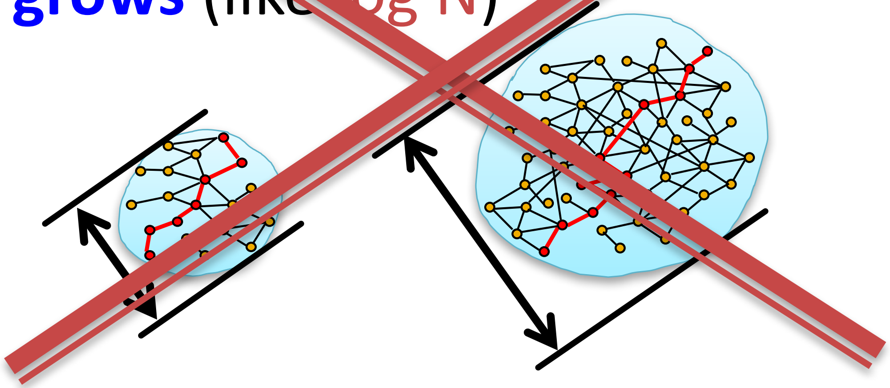
or equivalently $\dfrac{\log(E(t))}{\log(N(t))} = const$

a … densification exponent: 1 ≤ **a** ≤ 2:

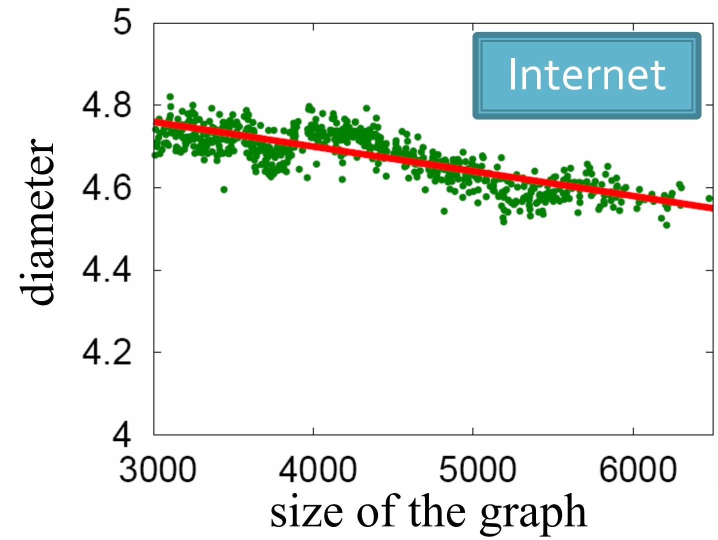
- **a=1: linear growth** – constant out-degree (traditionally assumed)
- **a=2: quadratic growth** – fully connected graph

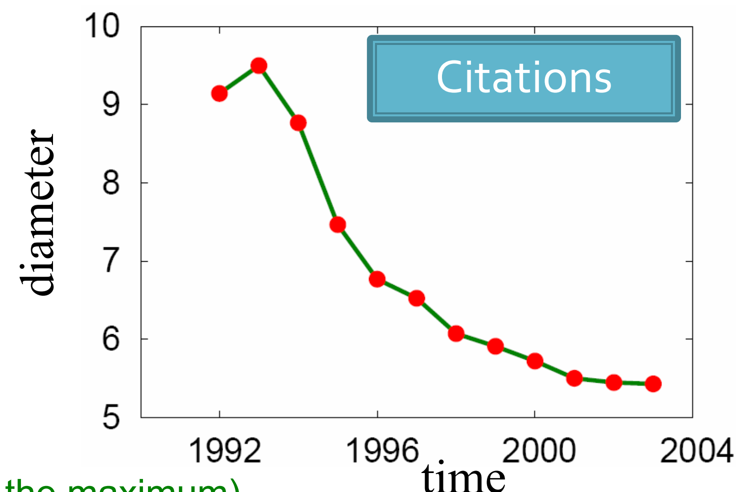
Q1) Network Evolution

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



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

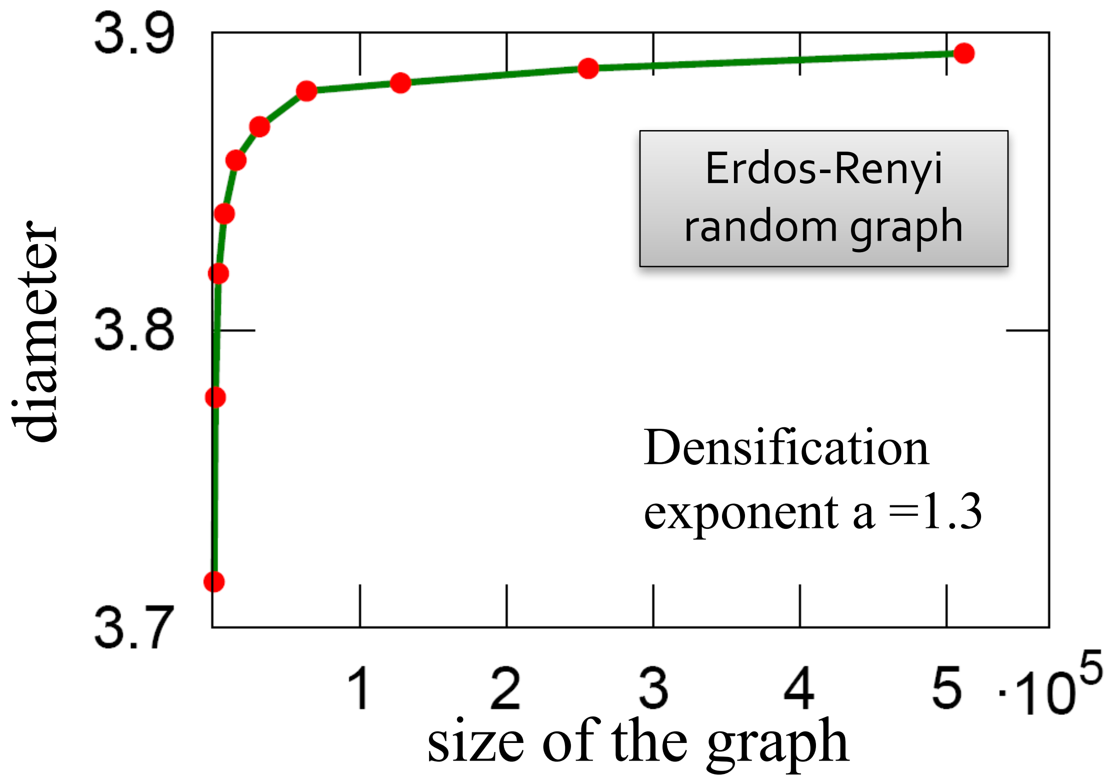




How do we compute diameter in practice?
-- Long paths: Take 90th-percentile or average path length (not the maximum)
-- Disconnected components: Take only largest component or average only over connected pairs of nodes

Diameter of a Densifying G_{np}

Is shrinking diameter just a consequence of densification?



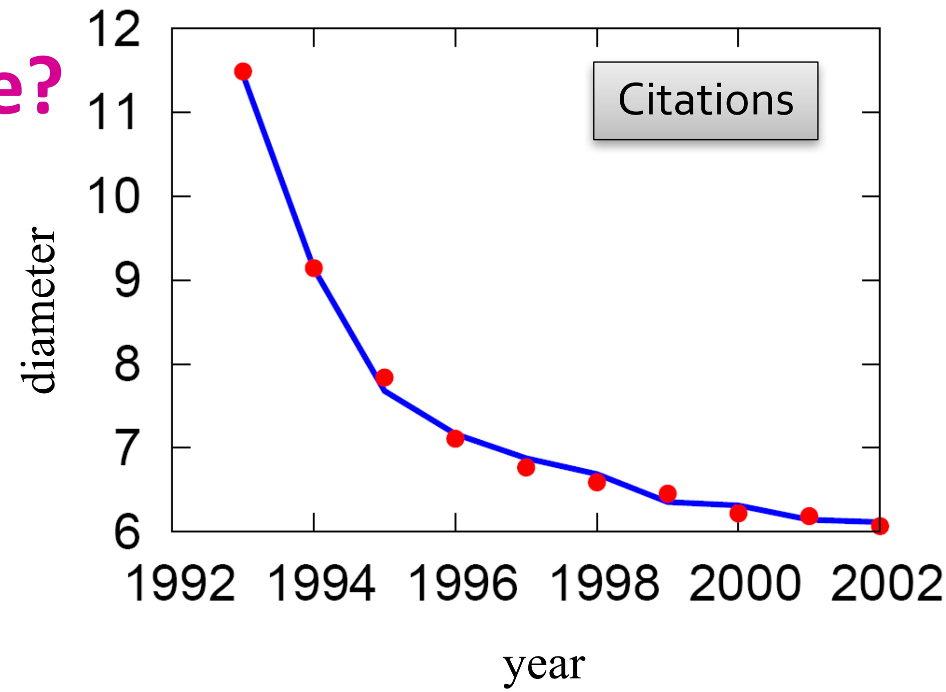
Densifying random graph has increasing diameter
⇒ **There is more to shrinking diameter than just densification!**

Diameter of a Rewired Network

Is it the degree sequence?

Compare diameter of a:

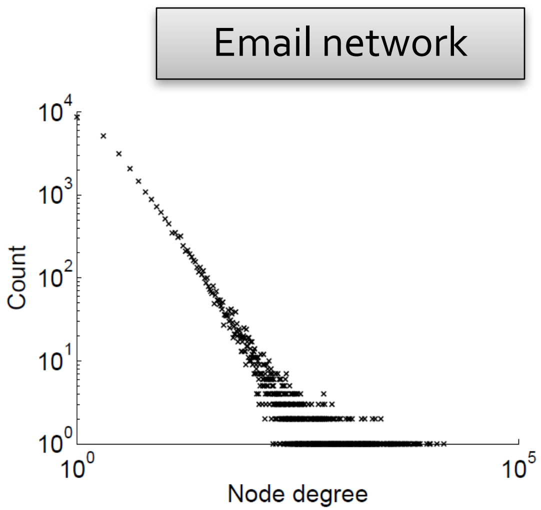
- Real network (**red**)
- Random network with the same degree distribution (**blue**)

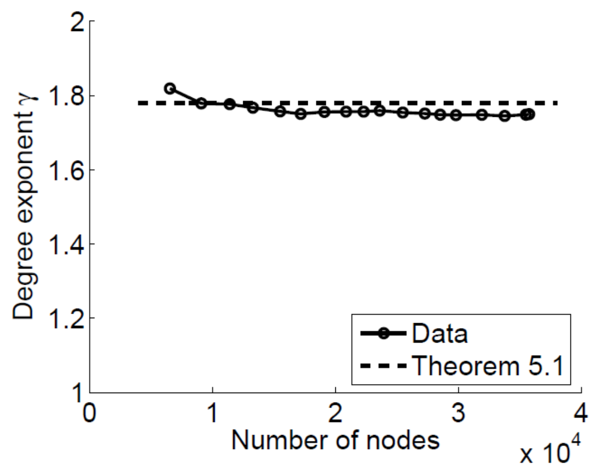


Densification + degree sequence gives shrinking diameter

Connecting Degrees & Densification

- **How does degree distribution evolve to allow for densification?**
- **Option 1) Degree exponent γ_t is constant:**
  - **Fact 1:** If $\gamma_t = \gamma \in [1, 2]$, then: $a = 2/\gamma$





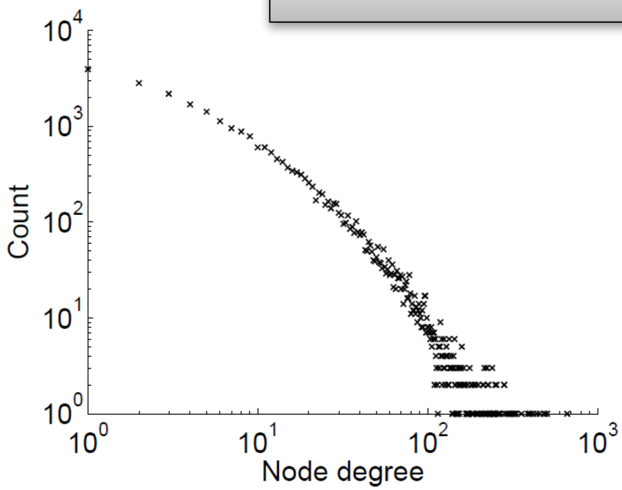
A consequence of what we learned in the Power law lecture:
- Power-laws with exponents <2 have infinite expectations.
- So, by maintaining constant degree exponent α the average degree grows.

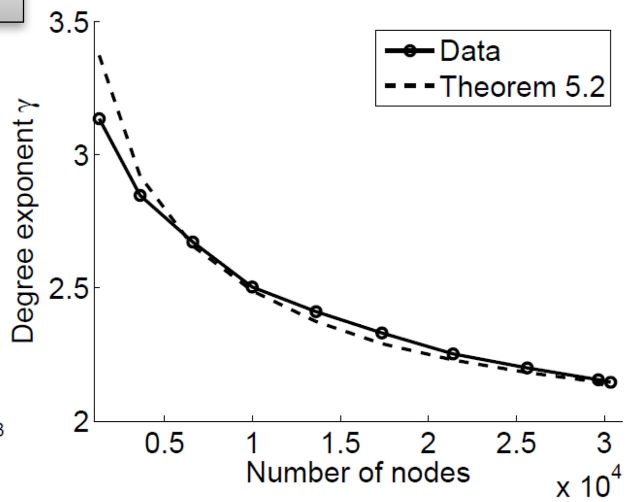
Connecting Degrees & Densification

- **How does degree distribution evolve to allow for densification?**
- **Option 2) γ_t evolves with graph size n:**
  - **Fact 2:** If $\gamma_t = \frac{4n_t^{x-1}-1}{2n_t^{x-1}-1}$, then: $a = x$

Notice: $\gamma_t \to 2$ as $n_t \to \infty$

Citation network





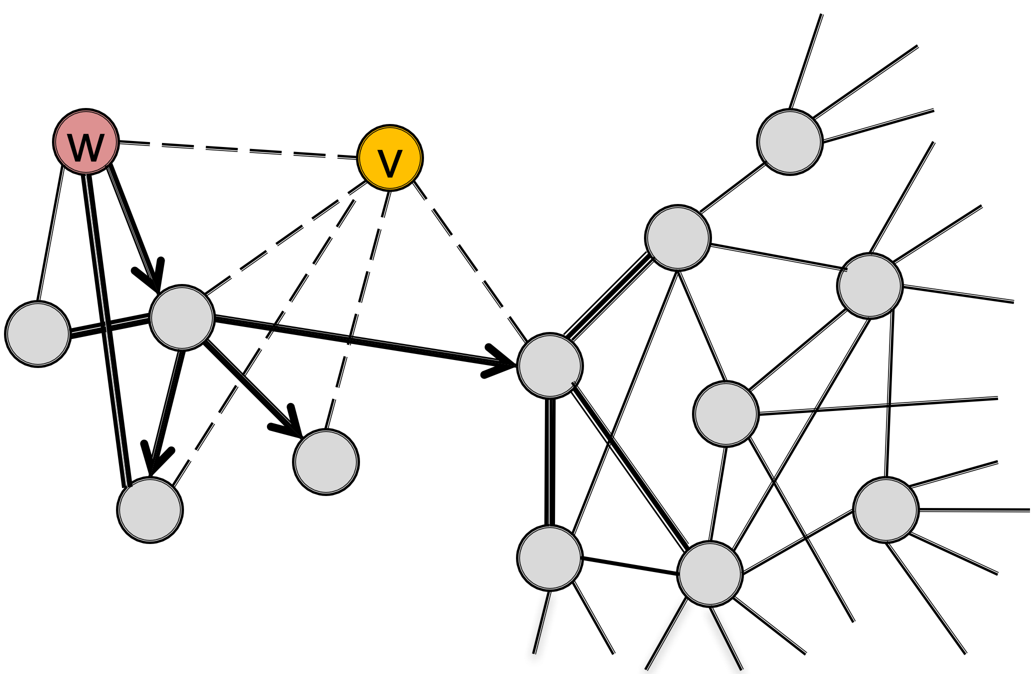
Remember, the expected degree in a power law is:

$$E[X] = \frac{\gamma_t - 1}{\gamma_t - 2} x_m$$

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

Forest Fire Model

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



Forest Fire Model

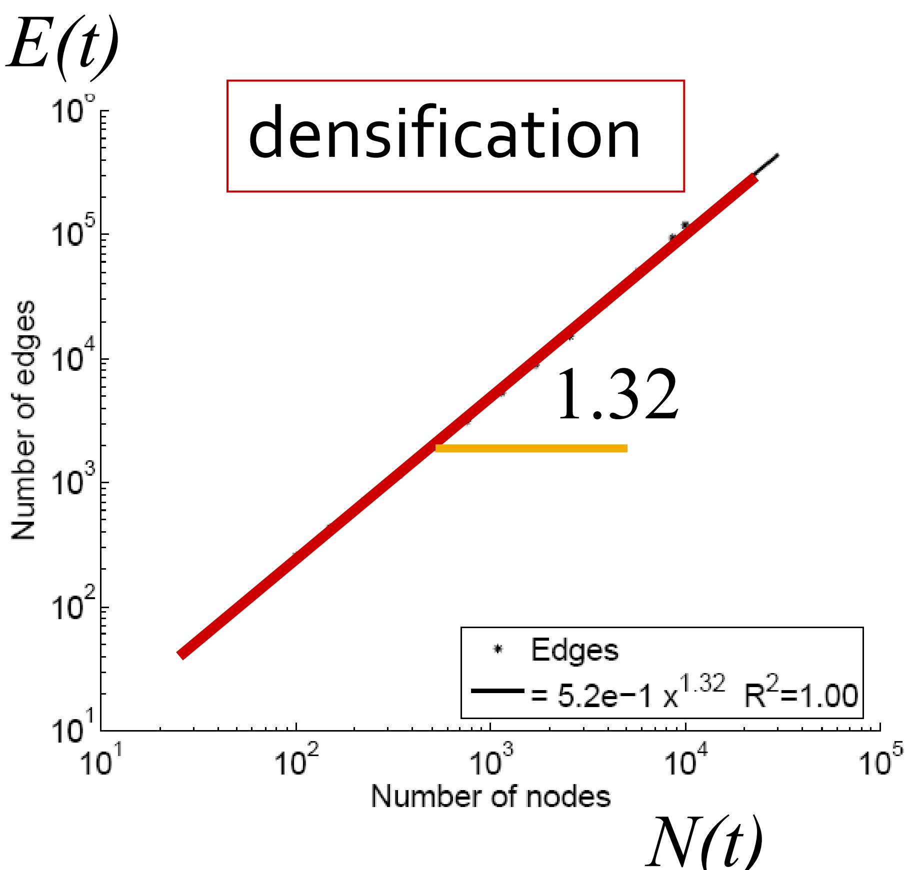
- **The Forest Fire model has 2 parameters:**
  - ***p*** … forward burning probability
  - ***r*** … backward burning probability
- **The model: Directed Graph**
  - Each turn a new node *v* arrives
  - Uniformly at random chooses an "ambassador" *w*
  - Flip 2 geometric coins (based on *p* and *r*) 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

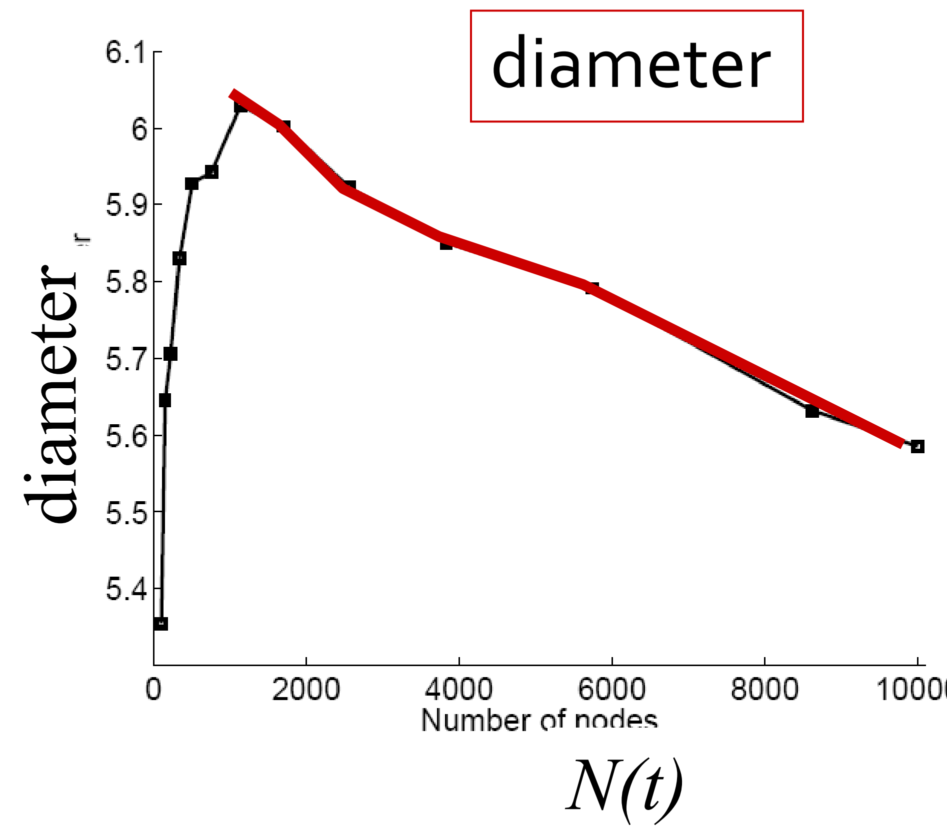
Geometric distribution:

$$\Pr(X = k) = (1 - p)^{k-1} p$$

Forest Fire Model

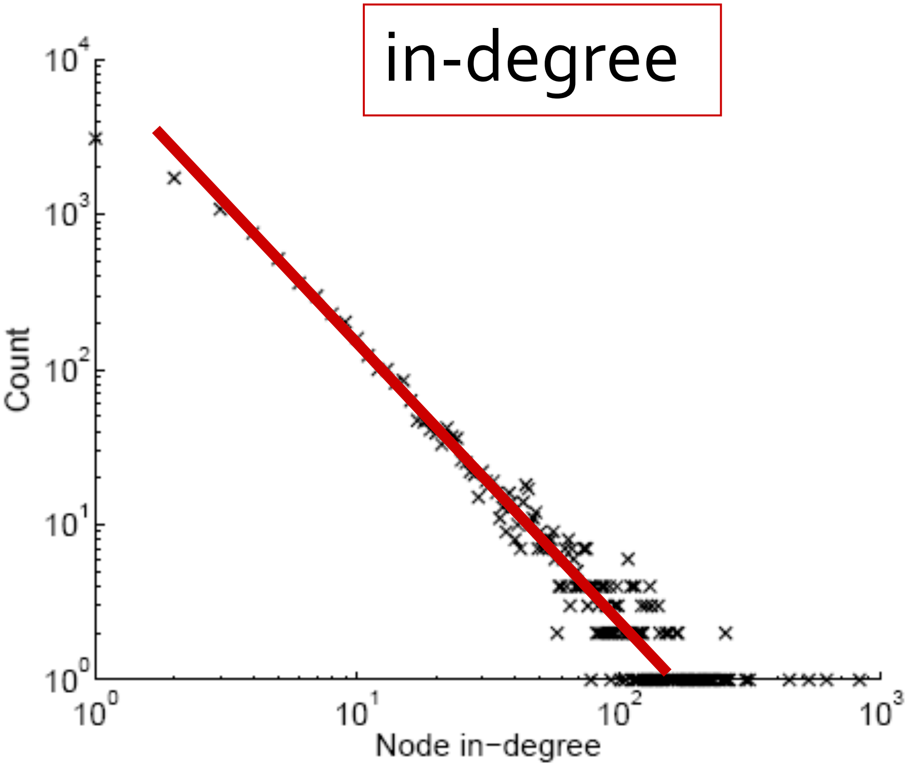
- **Forest Fire generates graphs that densify and have shrinking diameter**

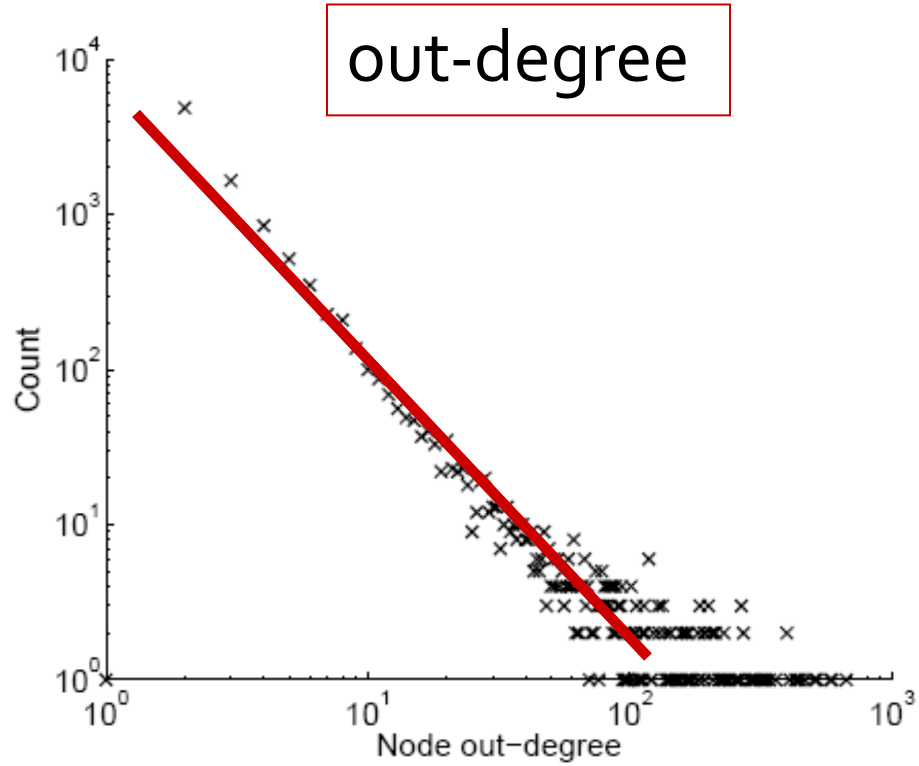




Forest Fire Model

- **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
- **The "sweet spot" is very narrow**