Preferential Attachment and Network Evolution

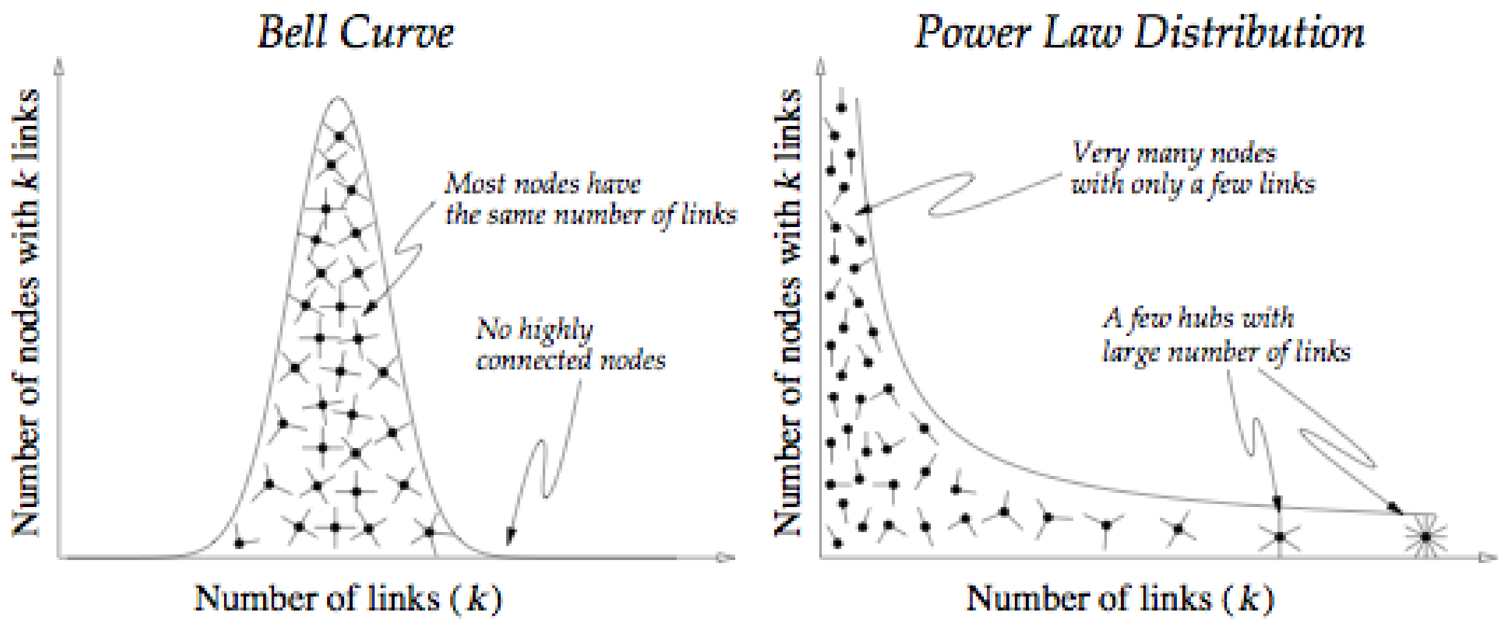
Thanks to Jure Leskovec, Stanford and Panayiotis Tsaparas, Univ. of Ioannina for slides

Agenda

- Preferential Attachment Model
- Microscopic Evolution of Social Networks
- Macroscopic Evolution of Social Networks
  - Forest Fire Model

Preferential Attachment Model

Exponential vs. Power-Law Tails

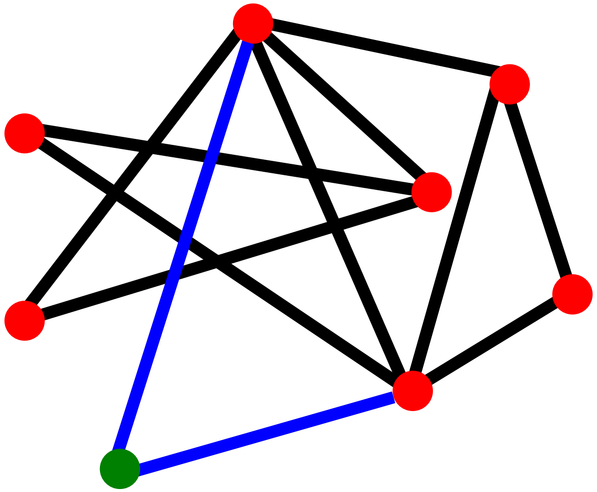


Model: G_{np} **?**

Model: Preferential attachment

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

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



Rich Get Richer

- **New nodes are more likely to link to nodes that already have high degree**
- **Herbert Simon's result:**
  - Power-laws arise from "**Rich get richer**" (**cumulative advantage**)
- **Examples**
  - **Citations** [de Solla Price '65]: New citations to a paper are proportional to the number it already has
    - **Herding:** If a lot of people cite a paper, then it must be good, and therefore I should cite it too
  - **Sociology:** Matthew effect
    - Eminent scientists often get more credit than a comparatively unknown researcher, even if their work is similar
    - http://en.wikipedia.org/wiki/Matthew_effect

Preferential attachment: Good news

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

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 sometime avoid linking to each other
- **Further questions:**
  - **What is a reasonable model for how people sample through network node and link to them?**
    - Short random walks

Generating Power-Law Values

- A simple trick to generate values that follow a power-law distribution:
  - Generate values r uniformly at random within the interval [0,1]
  - Transform the values using the equation
$$x = x_{min}(1 - r)^{-1/(\alpha-1)}$$
  - Generates values distributed according to power-law with exponent α

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

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

Small world

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.

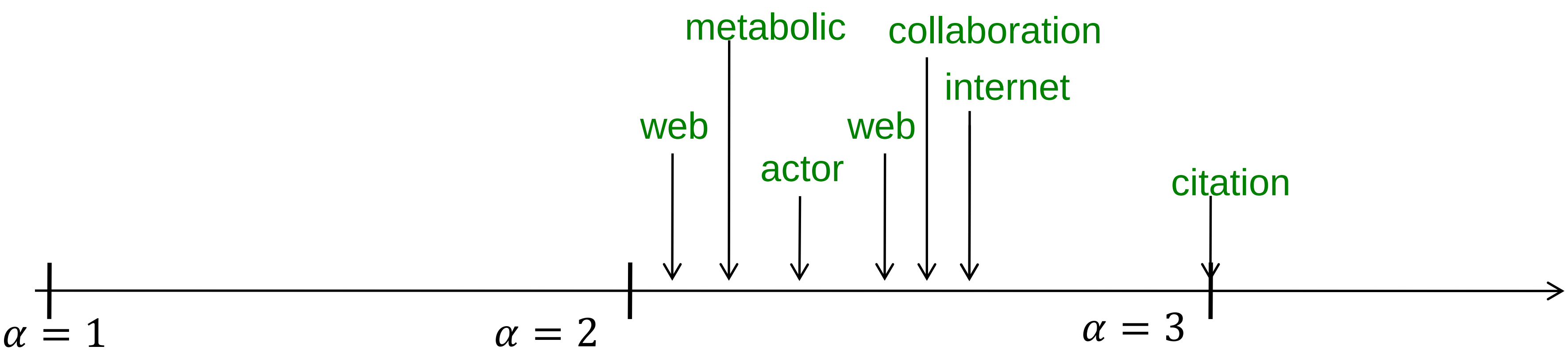
The average path length increases slower than logarithmically. In G_{np} all nodes have comparable degree, thus most paths will have comparable length. In a scale-free network vast majority of the path go through the few high degree hubs, reducing the distances between nodes.

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

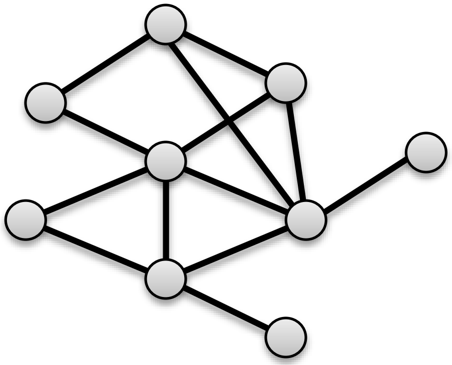
Microscopic Evolution of Social Networks

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]

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}

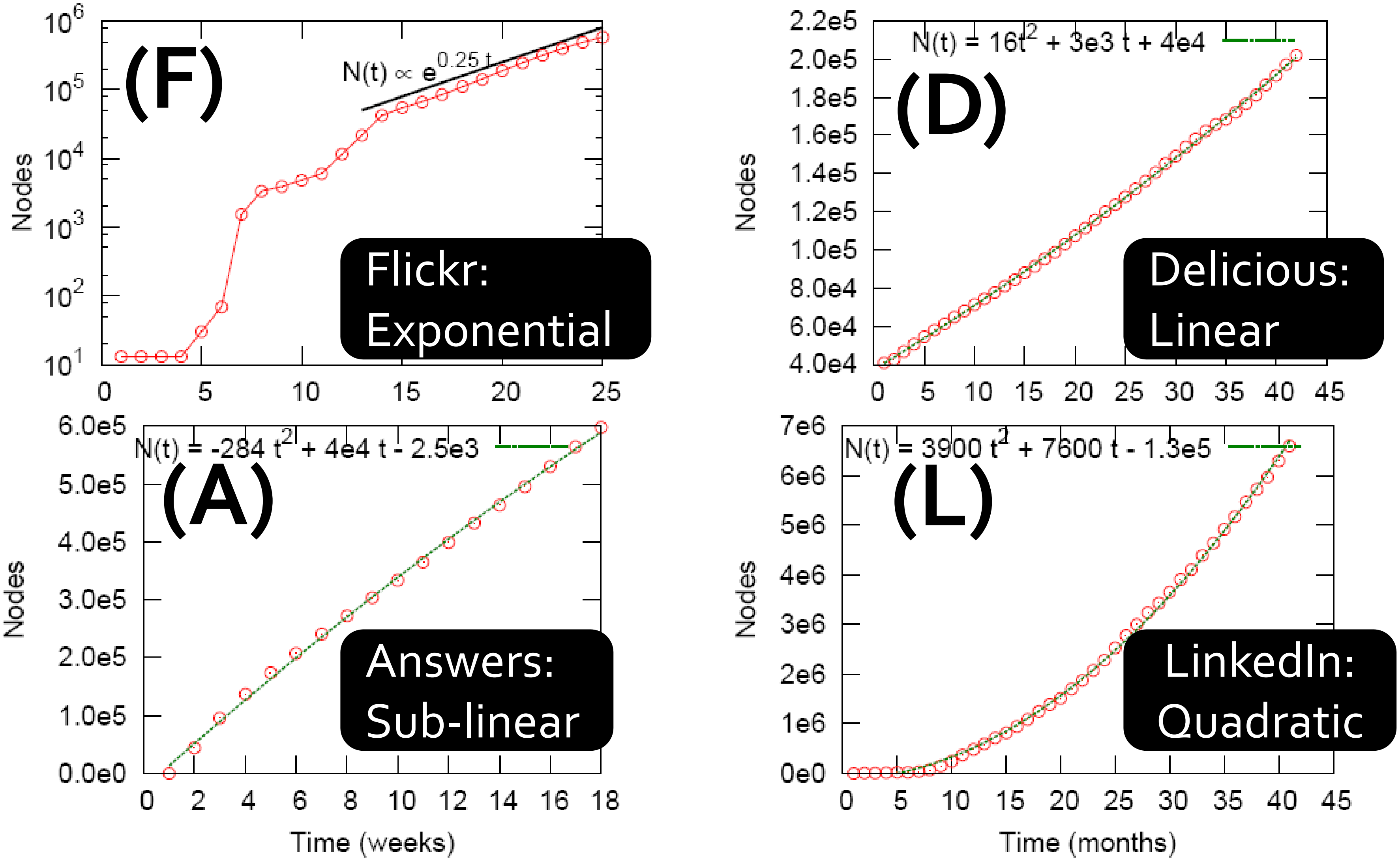


and so on for millions…

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

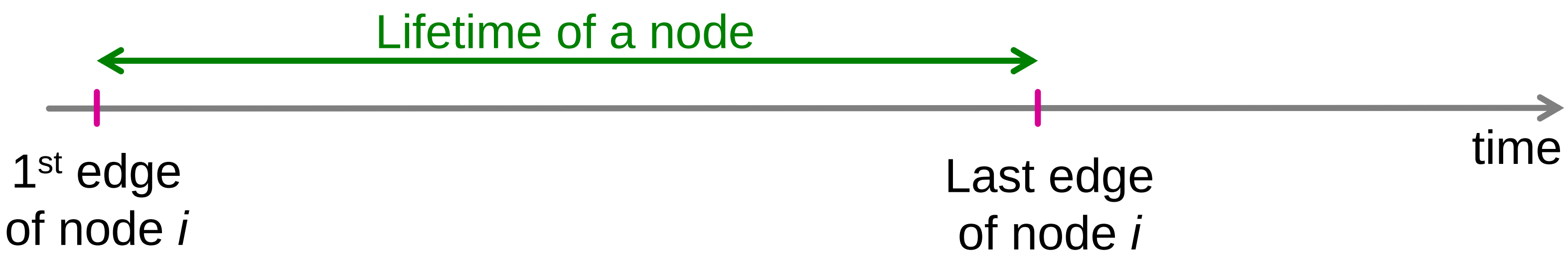
	Network	T	N	E
(F)	Flickr (03/2003–09/2005)	621	584,207	3,554,130
(D)	Delicious (05/2006–02/2007)	292	203,234	430,707
(A)	Answers (03/2007–06/2007)	121	598,314	1,834,217
(L)	LinkedIn (05/2003–10/2006)	1294	7,550,955	30,682,028

P1) When are New Nodes Arriving?

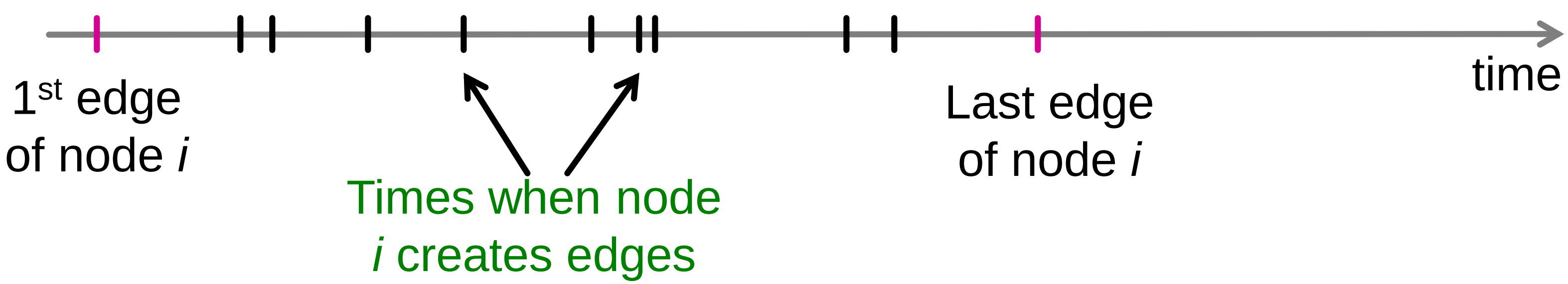


P2) When Do Nodes Create Edges?

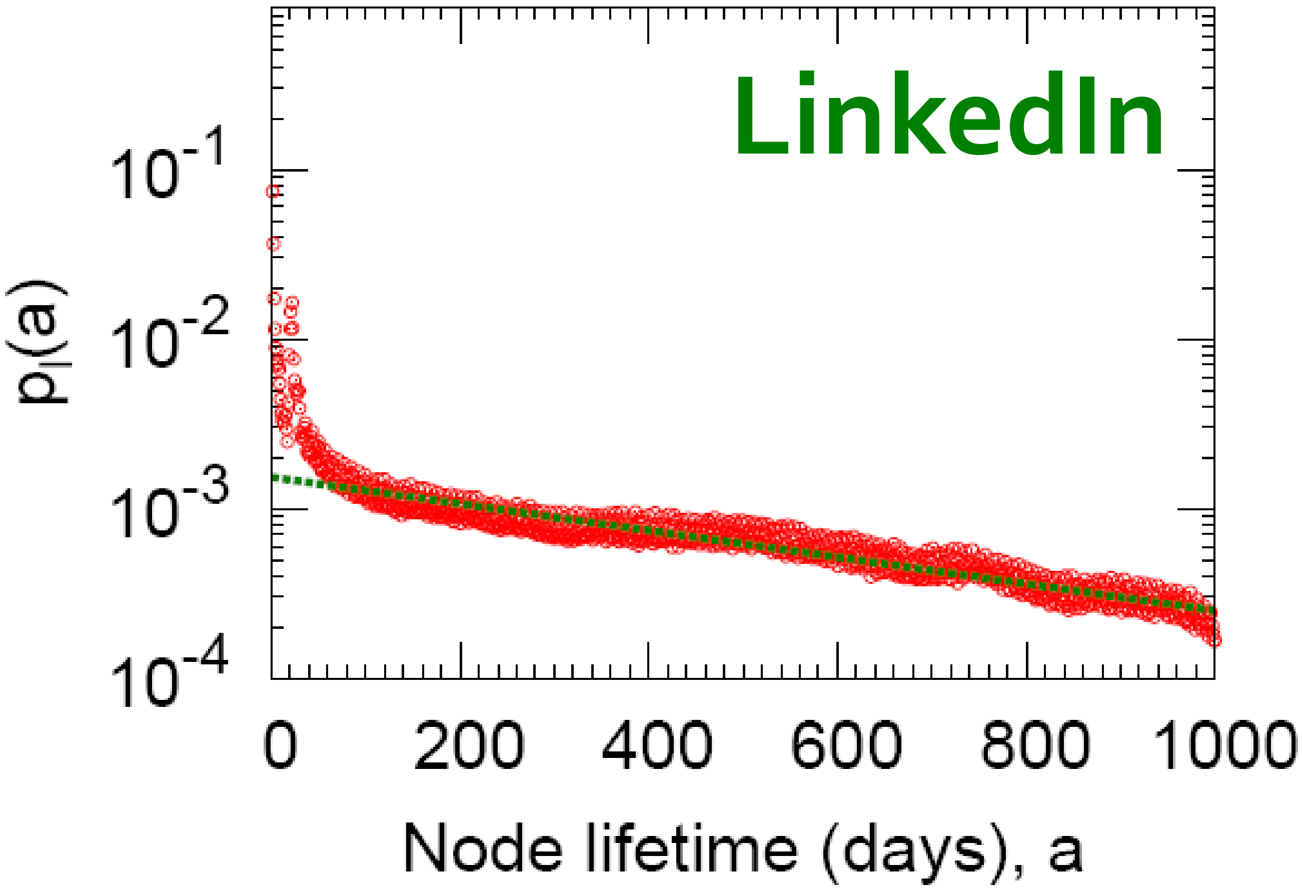
- **How long do nodes live?**
  - Node life-time is the time between the 1st and the last edge of a node



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



P2) What is Node Lifetime?



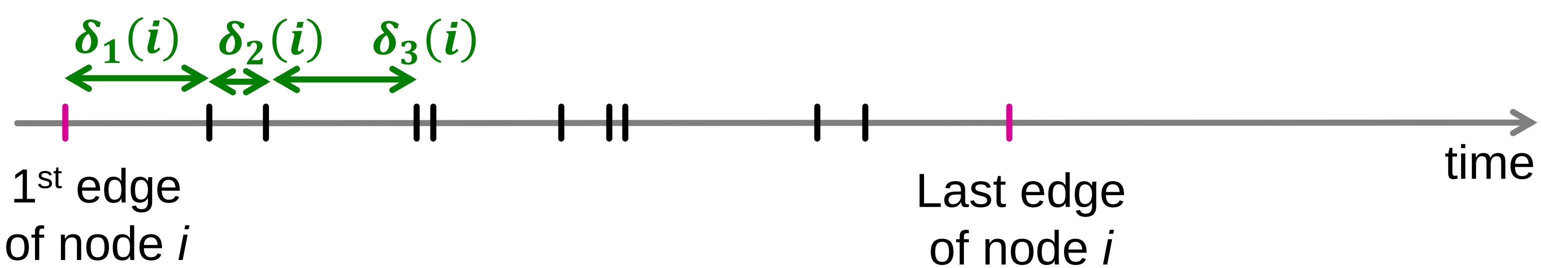
- **Lifetime *a***: Time between node's first and last edge

Node lifetime is **exponentially distributed:**

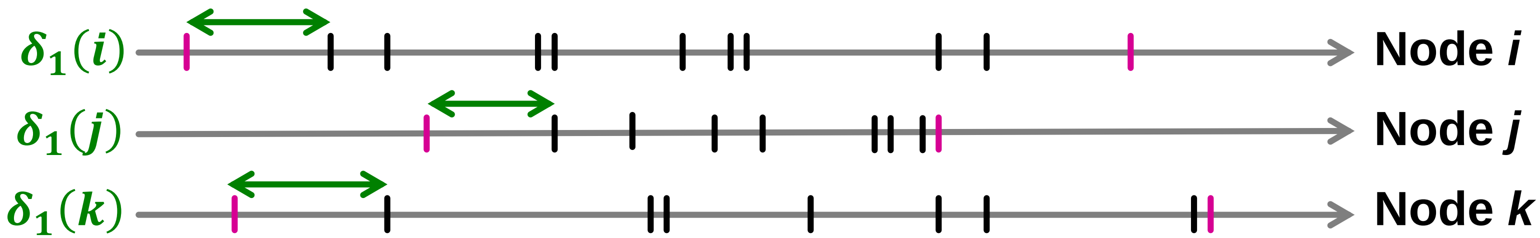
$$p_l(a) = \lambda e^{-\lambda a}$$

P2) When do Nodes Create Edges?

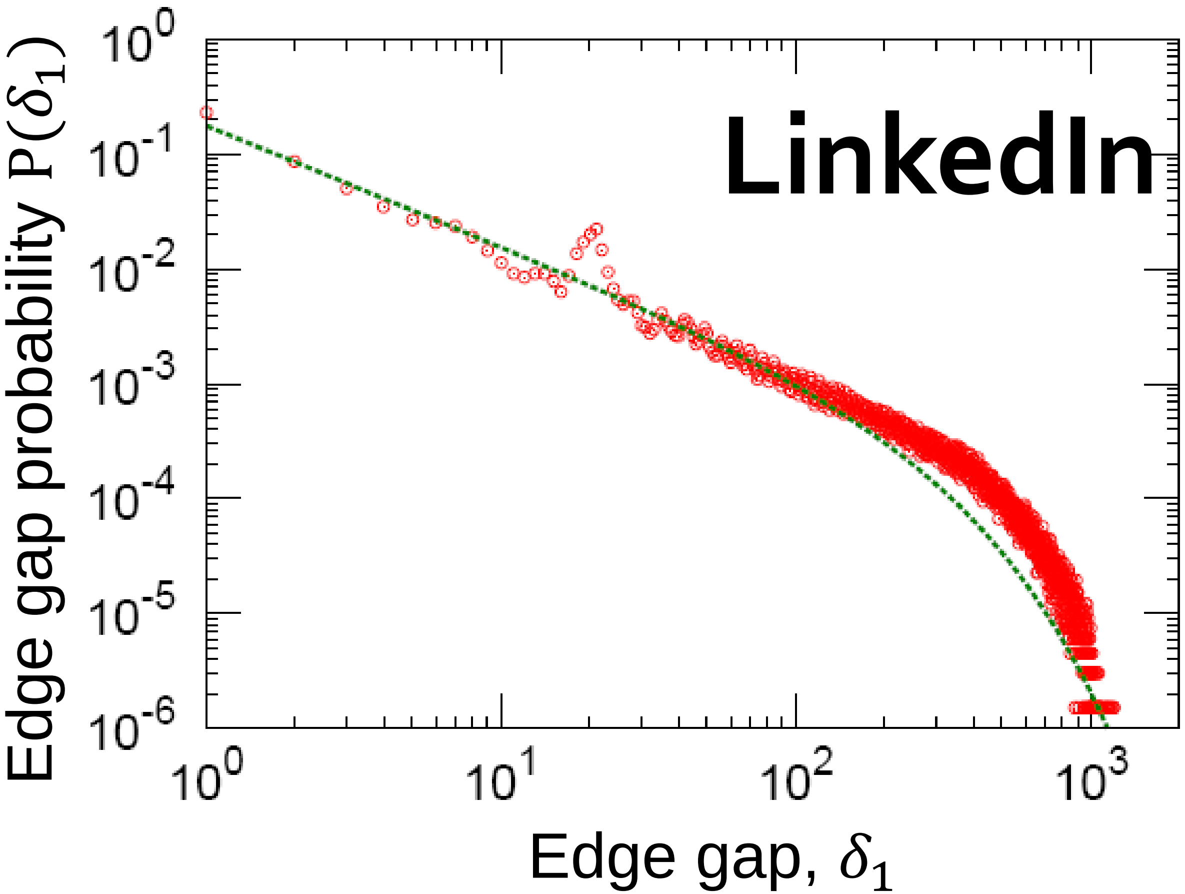
- **How do nodes "wake up" to create edges?**
  - **Edge gap $\delta_d(i)$**: time between d^{th} and $d+1^{st}$ edge of node i:
    - Let $t_d(i)$ be the creation time of d-th edge of node i
    - $\delta_d(i) = t_{d+1}(i) - t_d(i)$



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



P2) When do Nodes Create Edges?



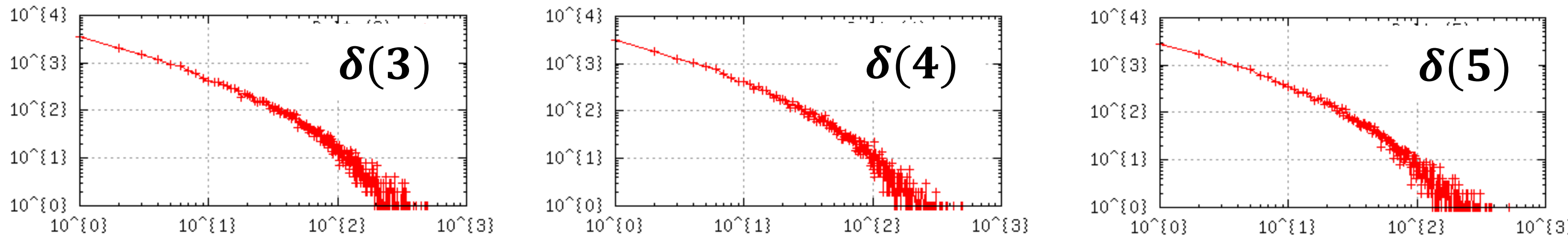
- **Edge gap δ_d**: inter-arrival time between d^{th} and $d+1^{st}$ edge **is distributed by a power-law with exponential cut-off**

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

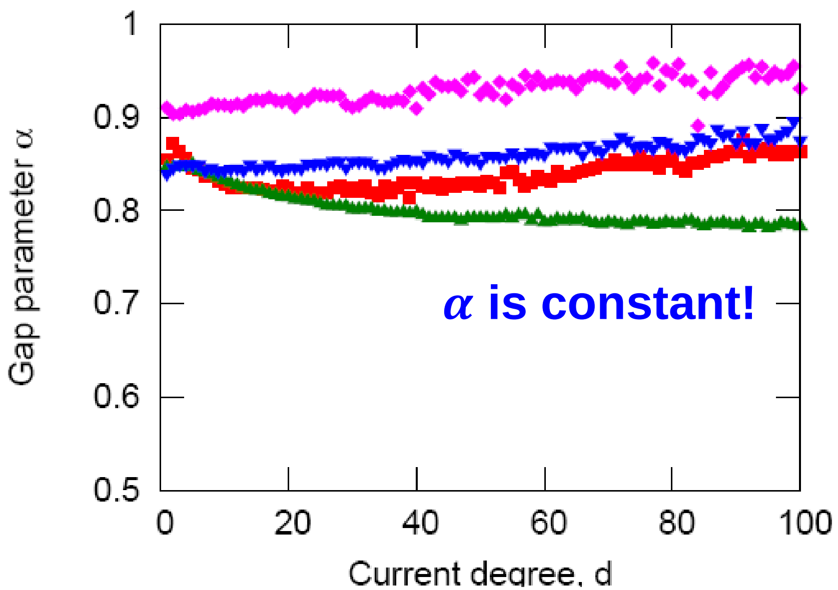
For every ***d*** we make a separate histogram

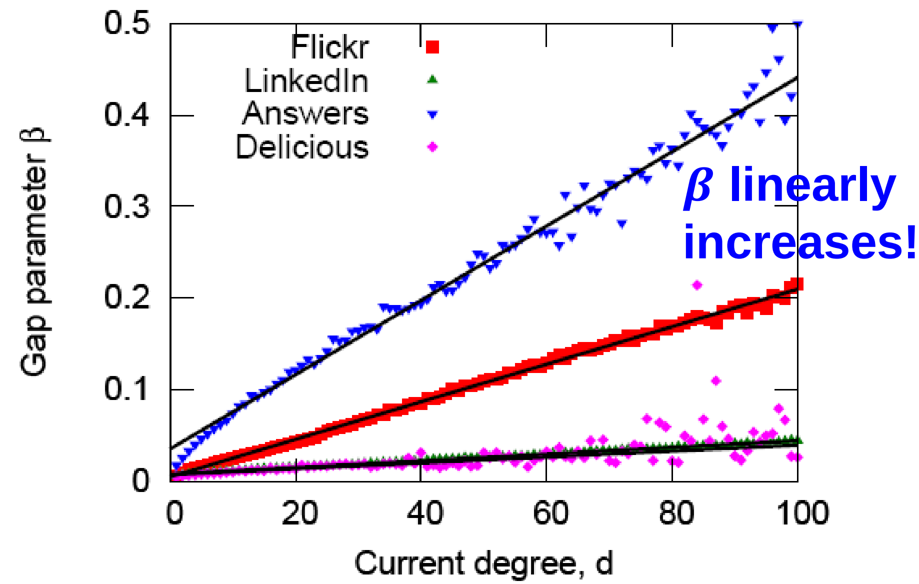
P2) How do α and β evolve with *d*?

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



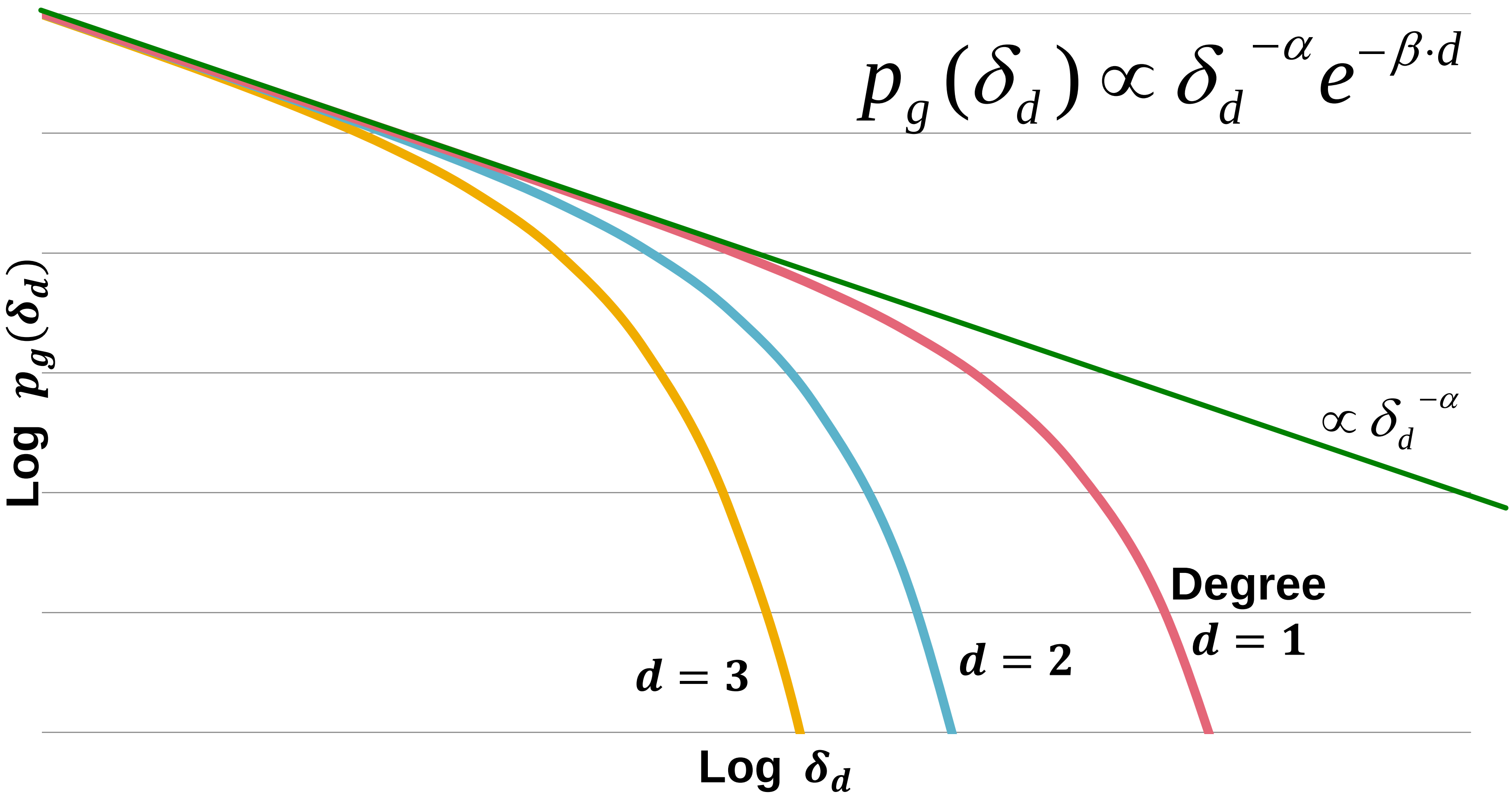
To each plot of δ_d fit: $p_g(\delta_d) \propto \delta_d^{-\alpha_d} e^{-\beta_d}$





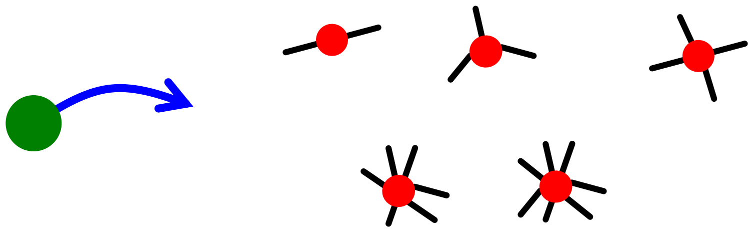
P2) Evolution of Edge Gaps

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

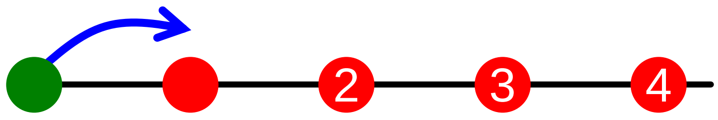


P3) How to Select Destination?

- 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?

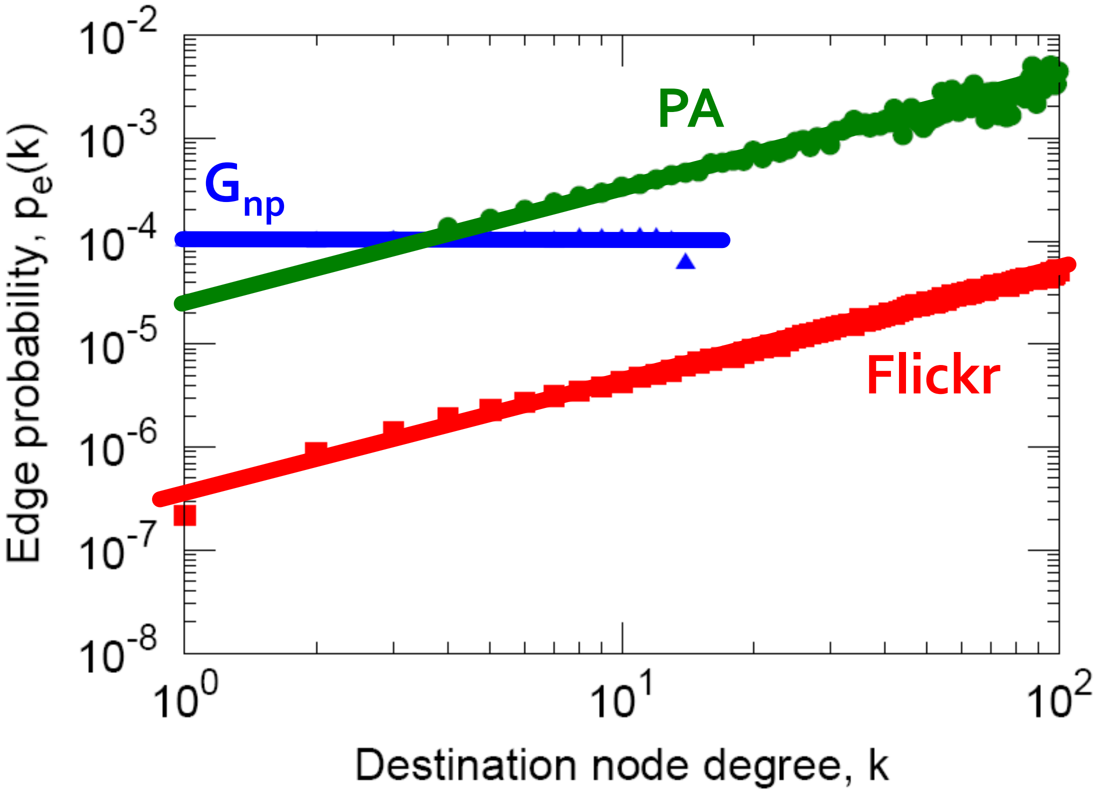


  - How many hops away is the target j?
    - Are edges attaching locally?



Edge Attachment Degree Bias

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

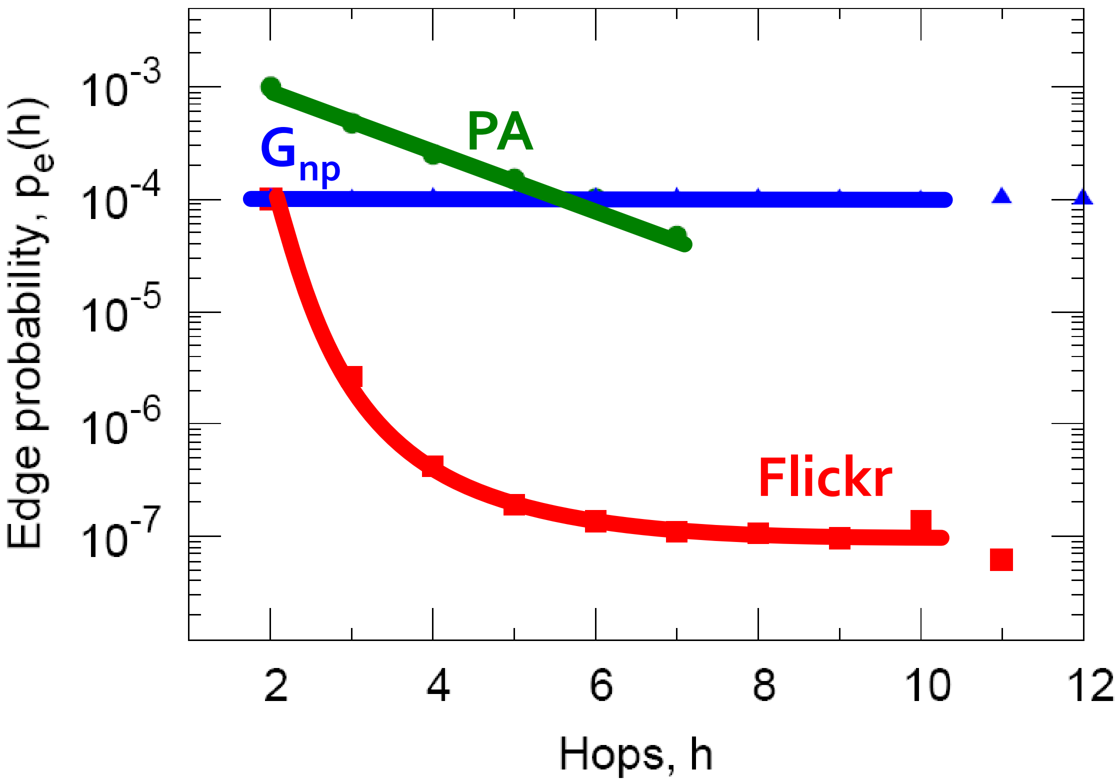


$$p_e(k) \propto k^{\tau}$$

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

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%

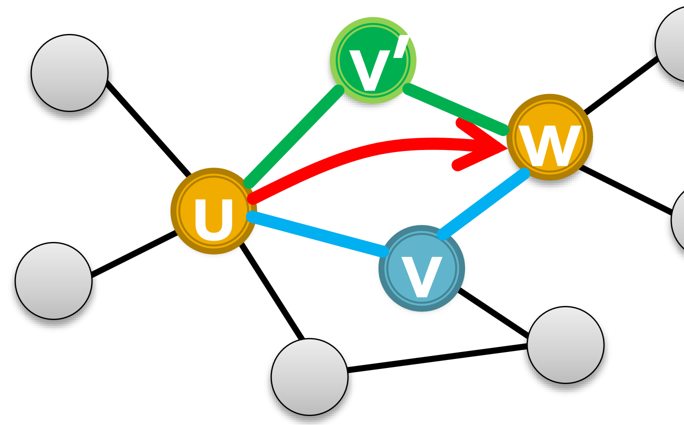
Real edges are local!
Most of them close **triangles**!

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*** chooses neighbor ***v***

  2. ***v*** chooses 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

[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

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

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

  $$N(t+1) = 2 \cdot N(t)$$

- **Q:** what is:

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

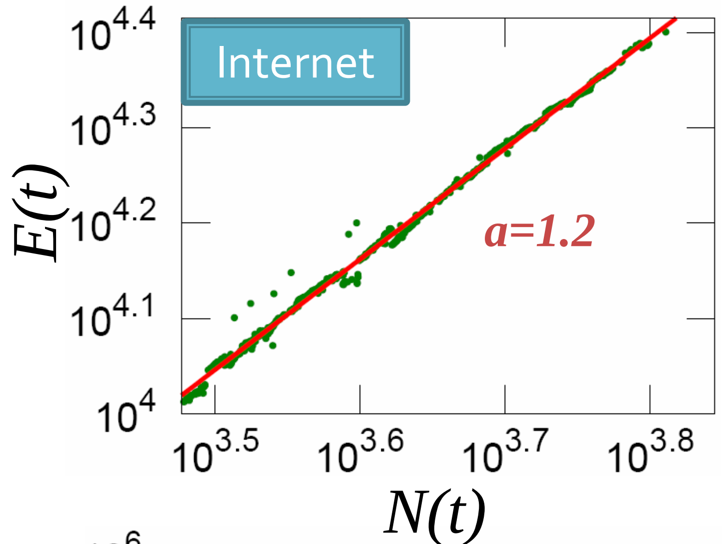
- **A: More than doubled!**
  - But obeying the **Densification Power Law**

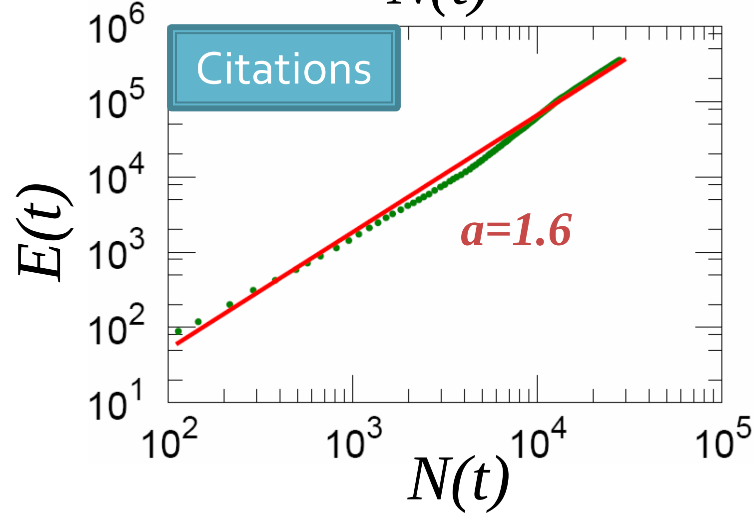
Q1) Network Evolution

- **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**

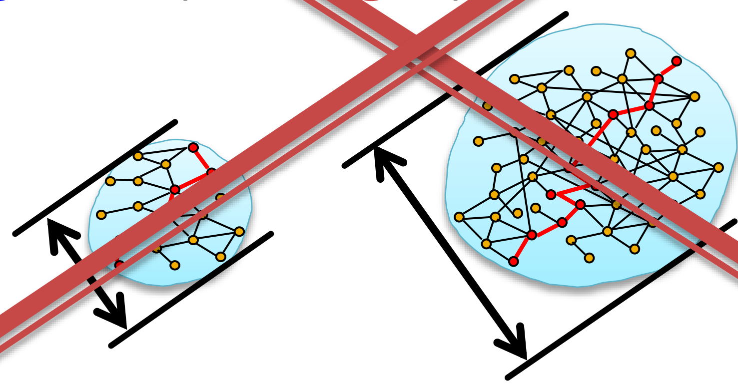
$$\boxed{E(t) \propto N(t)^a}$$ or equivalently $$\frac{\log(E(t))}{\log(N(t))} = const$$

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

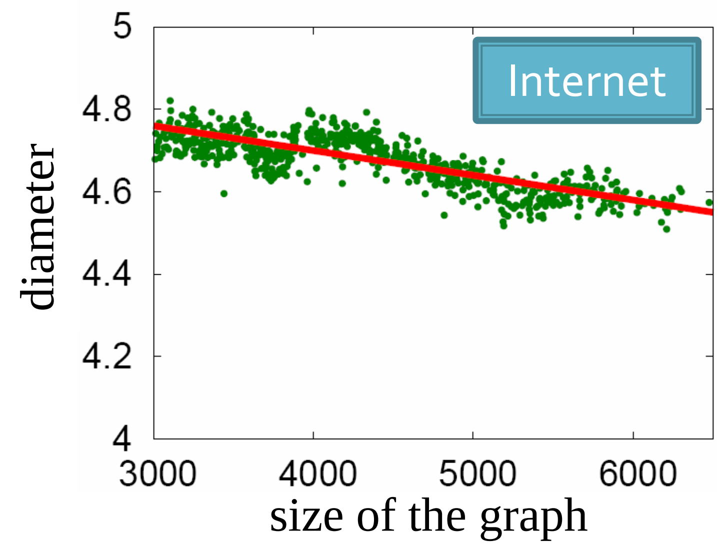
- **a=1: linear growth** – constant out-degree (traditionally assumed)
- **a=2: quadratic growth** – fully connected graph

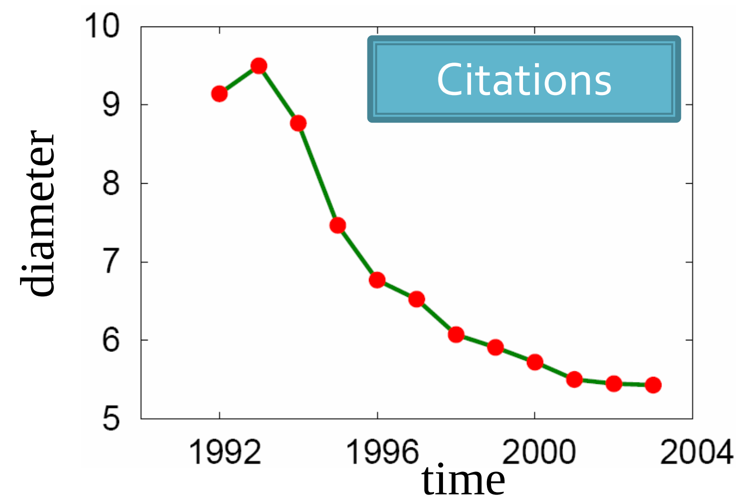
Q1) Network Evolution

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



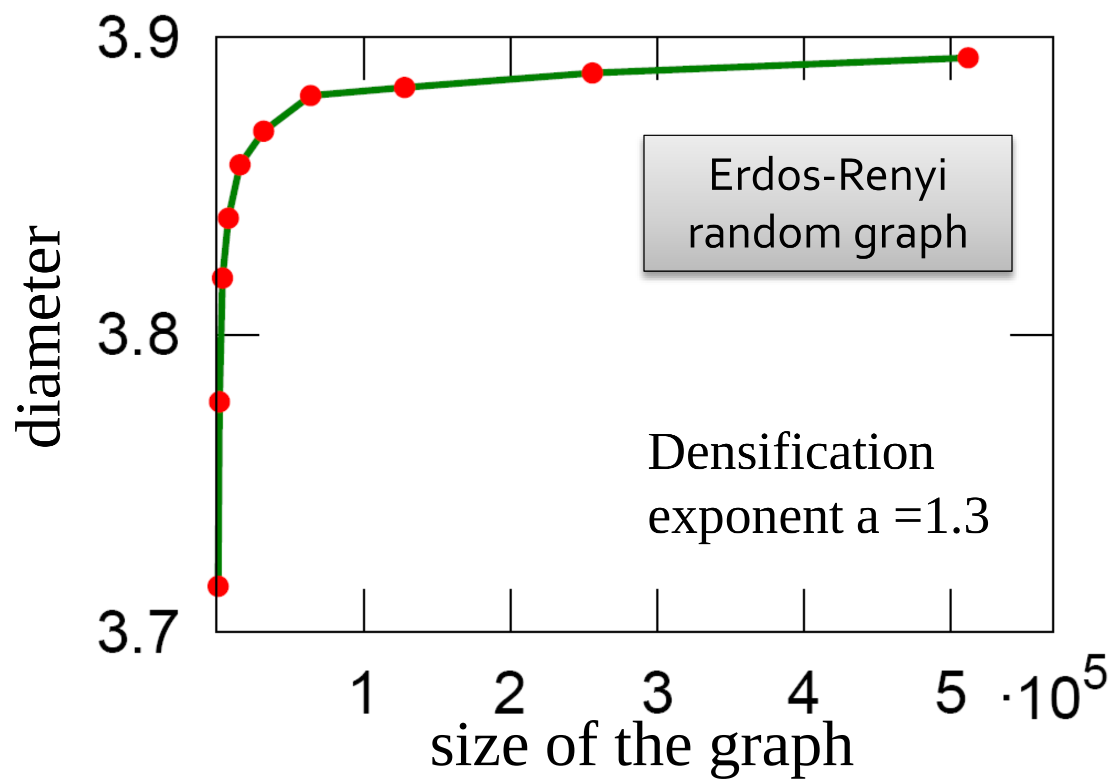
- **Diameter shrinks over time**
  - as the network grows the distances between the nodes slowly **decrease**





Diameter of a Densifying G_{np}

Is shrinking diameter just a consequence of densification?



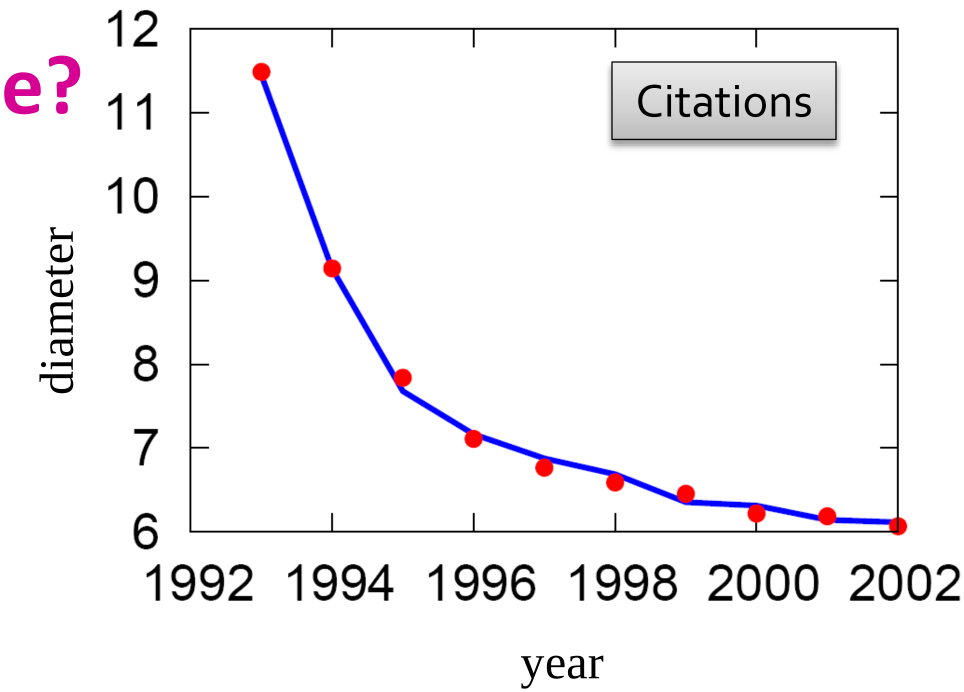
Densifying random graph has increasing diameter
$\Rightarrow$ **There is more to shrinking diameter than just densification!**

Diameter of a Rewired Network

Is it the degree sequence?

Compare diameter of a:

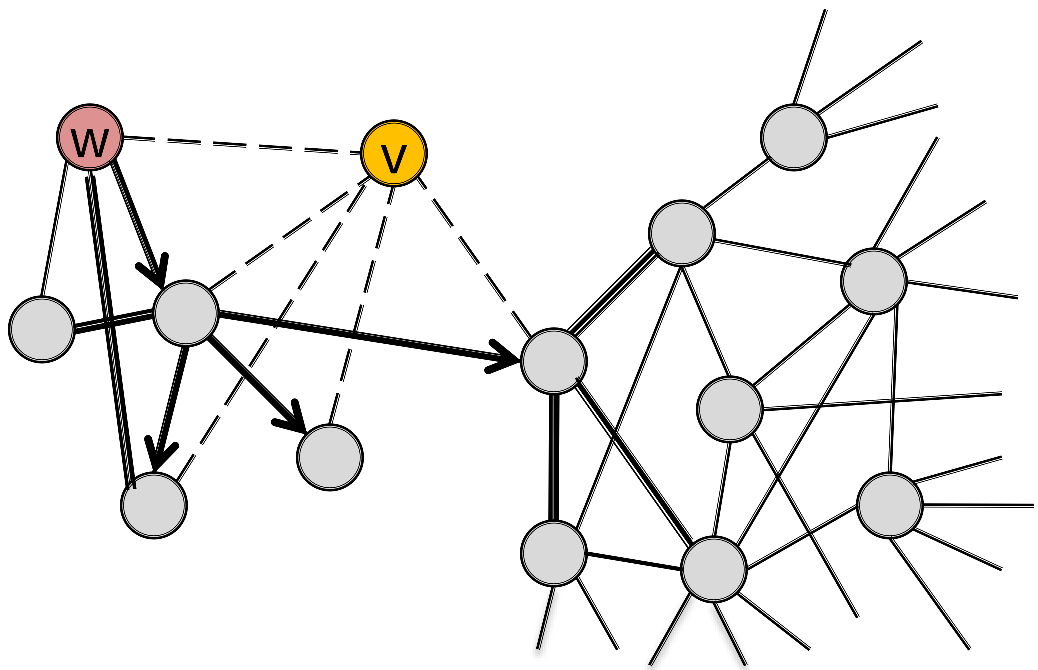
- Real network (**red**)
- Random network with the same degree distribution (**blue**)



Densification + degree sequence gives shrinking diameter

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

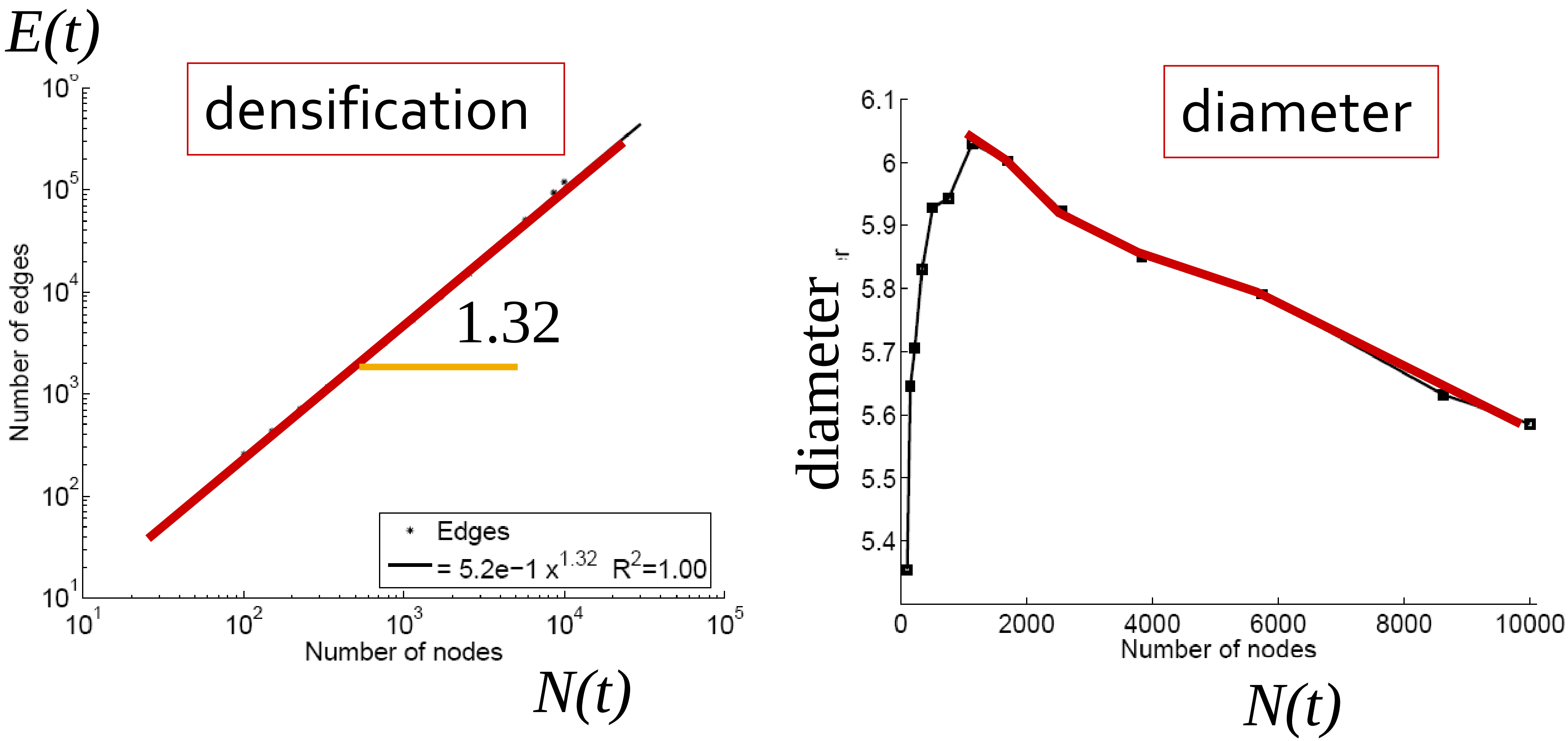
- **The Forest Fire model has 2 parameters:**
  - ***p*** … forward burning probability
  - ***r*** … backward burning probability
- **The model: Directed Graph**
  - Each turn a new node *v* arrives
  - Uniformly at random chooses an "ambassador" *w*
  - Flip 2 geometric coins (based on *p* and *r*)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

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

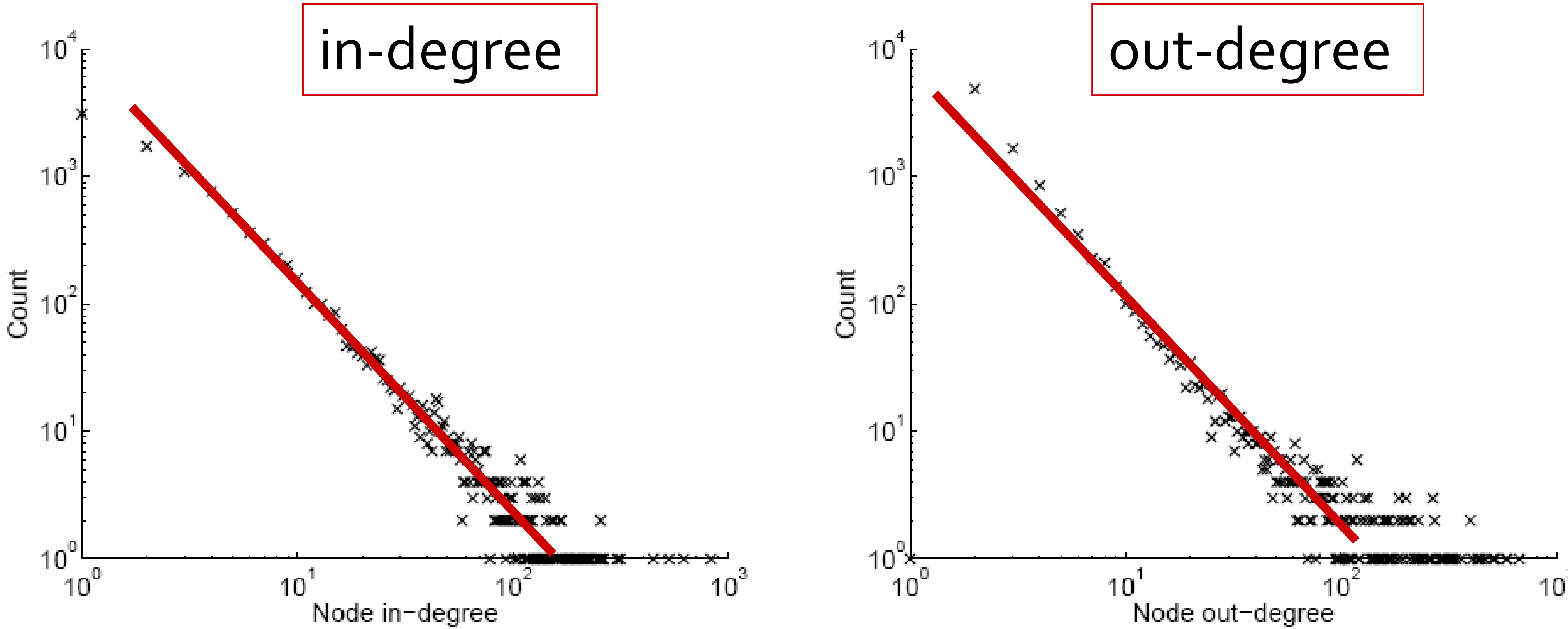
Forest Fire Model

- **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**