Complex Graphs and Networks
Lecture 2: Generative models - preferential attachment schemes

Linyuan Lu

lu@math.sc.edu

University of South Carolina

BASICS2008 SUMMER SCHOOL
July 27 – August 2, 2008

Overview of talks

- Lecture 1: Overview and outlines
- Lecture 2: Generative models - preferential attachment schemes
- Lecture 3: Duplication models for biological networks
- Lecture 4: The rise of the giant component
- Lecture 5: The small world phenomenon: average distance and diameter
- Lecture 6: Spectrum of random graphs with given degrees

Preferential attachment scheme

The rich gets richer.

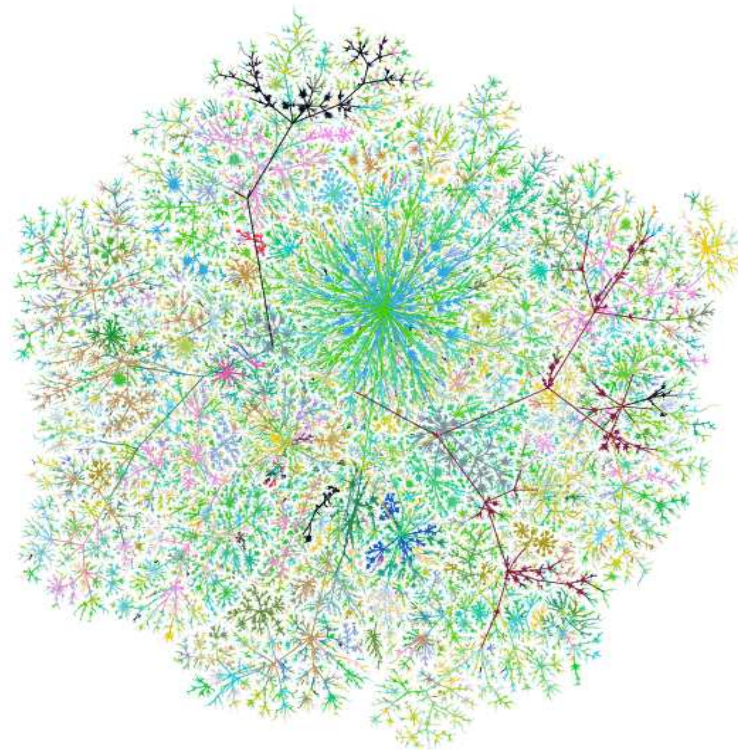
WWW Graphs

Call Graphs

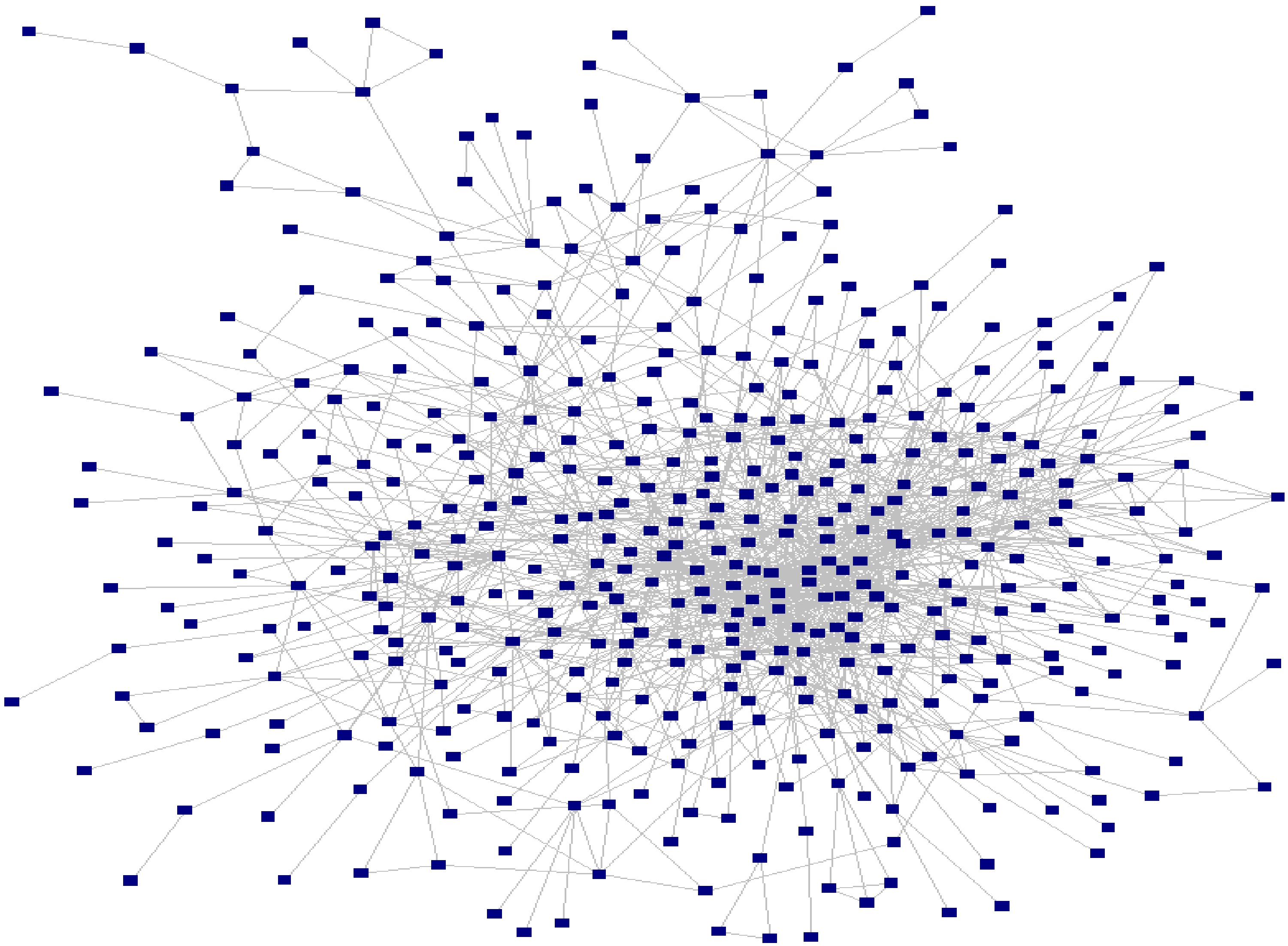
Collaboration Graphs

Costars Graph of Actors

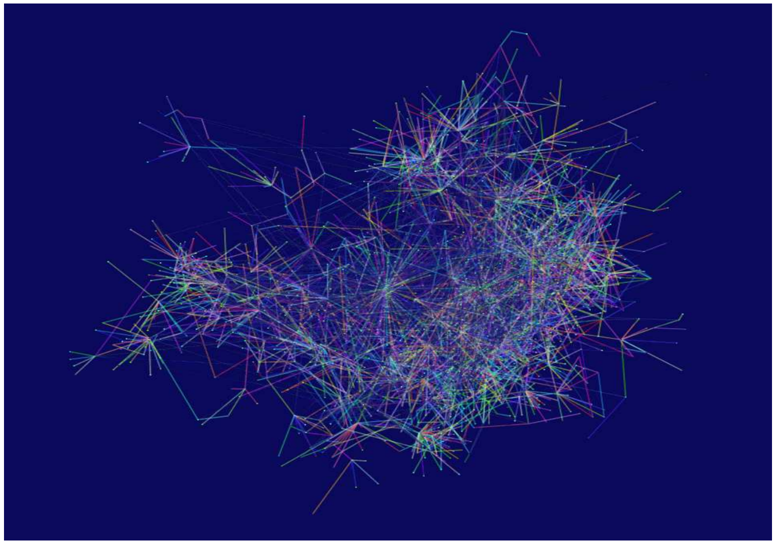
⋮



Erdős number 1 graph

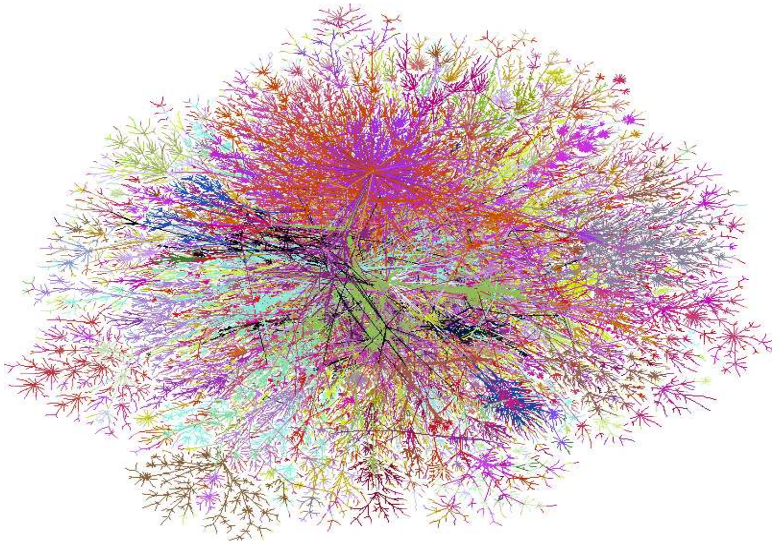


Power law graphs



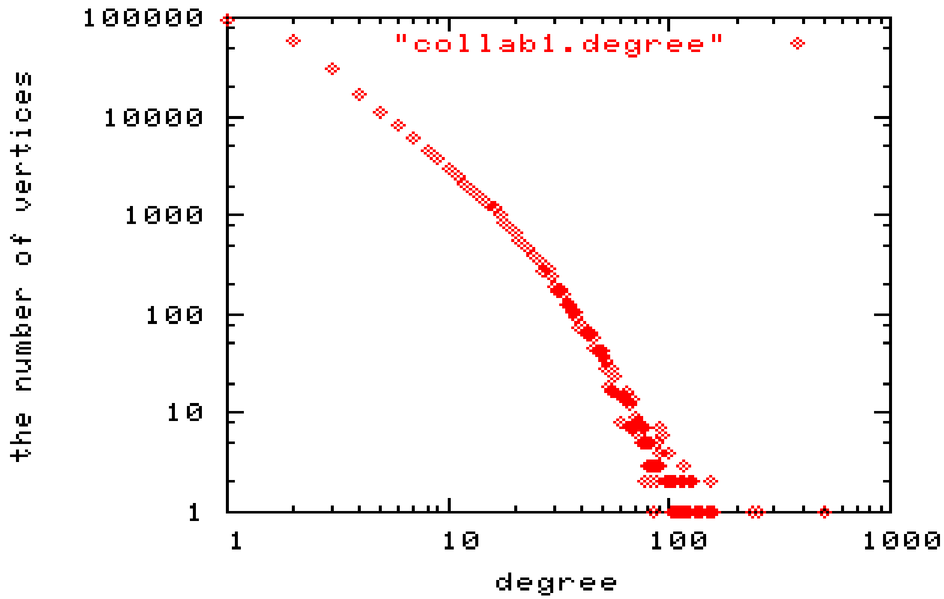
Left: Part of the collaboration graph (authors with Erdős number 2)

Right: An IP graph (by Bill Cheswick)



The power law

The number of vertices of degree k is approximately proportional to $k^{-\beta}$ for some positive β.



A power law graph is a graph whose degree sequence satisfies the power law.

History of power law

- **Lotka's Law (1926):** The distribution of authors in the index of Chemical Abstracts is power law.

History of power law

- **Lotka's Law (1926):** The distribution of authors in the index of Chemical Abstracts is power law.
- **Yule's Law (1942):** City population follows a power law.

History of power law

- **Lotka's Law (1926):** The distribution of authors in the index of Chemical Abstracts is power law.
- **Yule's Law (1942):** City population follows a power law.
- **Zipf's Law (1949):** The n-th most frequent word occurs at rate $\frac{1}{n}$.

History of power law

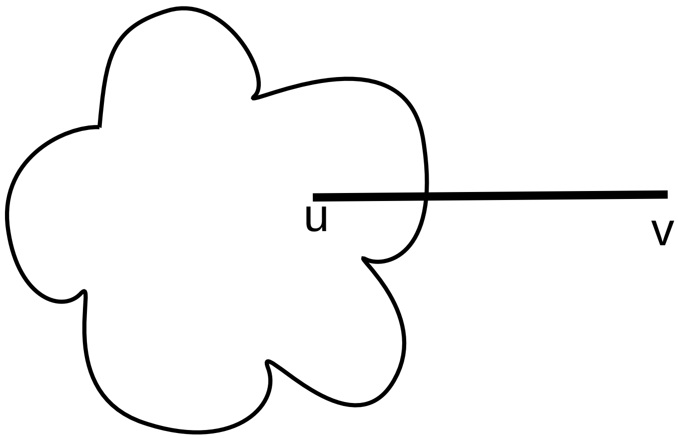
- **Lotka's Law (1926):** The distribution of authors in the index of Chemical Abstracts is power law.
- **Yule's Law (1942):** City population follows a power law.
- **Zipf's Law (1949):** The n-th most frequent word occurs at rate $\frac{1}{n}$.
- **Simon (1957):** Power law is common for various phenomena.

History of power law

- **Lotka's Law (1926):** The distribution of authors in the index of Chemical Abstracts is power law.
- **Yule's Law (1942):** City population follows a power law.
- **Zipf's Law (1949):** The n-th most frequent word occurs at rate $\frac{1}{n}$.
- **Simon (1957):** Power law is common for various phenomena.
- **Pareto, (1897):** Wealth distribution follows a power law.

Preferential attachment

$$G_0 \to G_1 \to G_2 \to \ldots \to G_t \to \ldots$$



Vertex-step: At time t, add a new vertex v to the existed network and attach v to a vertex u, which is selected with probability proportional to its current degree.

Barabási-Albert's model

m-**vertex-step:** At time t, add a new vertex v and m edges from v to the existed network using preferential attachment scheme.

Barabási-Albert's model

m-**vertex-step:** At time t, add a new vertex v and m edges from v to the existed network using preferential attachment scheme.

The number of vertices:

$$n_t = n_0 + t.$$

The number of edges:

$$e_t = e_0 + mt.$$

Heuristic analysis

Let $m_{k,t}$ be the number of vertices with degree k at time t.

$$\mathrm{E}(m_{k,t+1}) = (1 - m\frac{k}{2e_t})\mathrm{E}(m_{k,t}) + m\frac{k-1}{2e_t}\mathrm{E}(m_{k-1,t}).$$

Heuristic analysis

Let $m_{k,t}$ be the number of vertices with degree k at time t.

$$\mathrm{E}(m_{k,t+1}) = (1 - m\frac{k}{2e_t})\mathrm{E}(m_{k,t}) + m\frac{k-1}{2e_t}\mathrm{E}(m_{k-1,t}).$$

Write $\mathrm{E}(m_{k,t}) \approx M_k t$.

$$M_k(t+1) \approx (1 - m\frac{k}{2e_t})M_k t + m\frac{k-1}{2e_t}M_{k-1}t.$$

Heuristic analysis

Let $m_{k,t}$ be the number of vertices with degree k at time t.

$$\mathrm{E}(m_{k,t+1}) = (1 - m\frac{k}{2e_t})\mathrm{E}(m_{k,t}) + m\frac{k-1}{2e_t}\mathrm{E}(m_{k-1,t}).$$

Write $\mathrm{E}(m_{k,t}) \approx M_k t$.

$$M_k(t+1) \approx (1 - m\frac{k}{2e_t})M_k t + m\frac{k-1}{2e_t}M_{k-1}t.$$

Simplify it and let $t \to \infty$.

$$(1 + \frac{k}{2})M_k = \frac{k-1}{2}M_{k-1}.$$

Barabási-Albert's result

Write $M_k \approx ck^{-\beta}$. Then

$$\frac{M_k}{M_{k-1}} = (1 - \frac{1}{k})^{\beta} \approx 1 - \frac{\beta}{k}.$$

Thus,

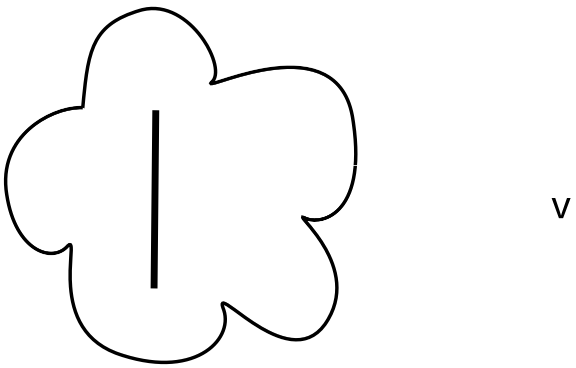
$$1 - \frac{\beta}{k} \approx \frac{(k-1)/2}{1 + k/2}.$$

It implies $\beta = 3$.

A general generative model

At time t,

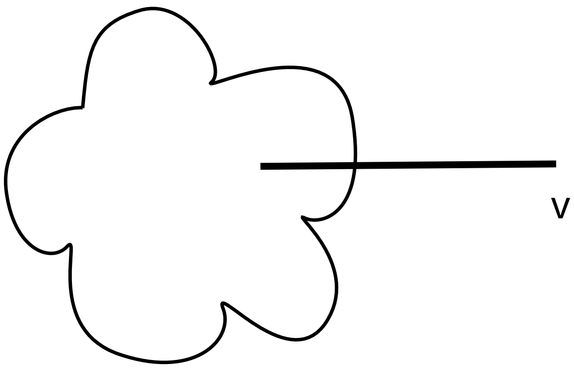
- add expected $\mu^{e,e}$ random random edges to existed network.
- add expected $\mu^{n,e}$ random edges between new vertex and existed network.
- add expected $\mu^{n,n}$ loops to the new vertex.



A general generative model

At time t,

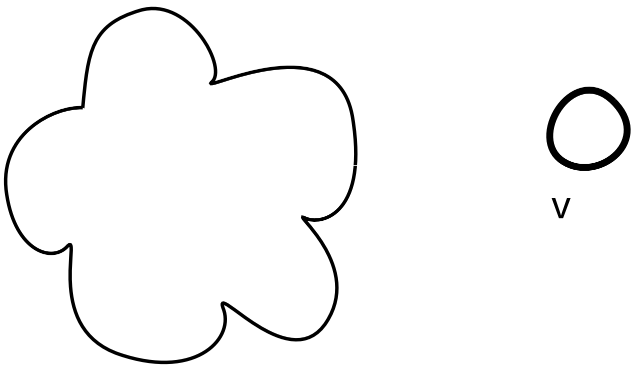
- add expected $\mu^{e,e}$ random random edges to existed network.
- add expected $\mu^{n,e}$ random edges between new vertex and existed network.
- add expected $\mu^{n,n}$ loops to the new vertex.



A general generative model

At time t,

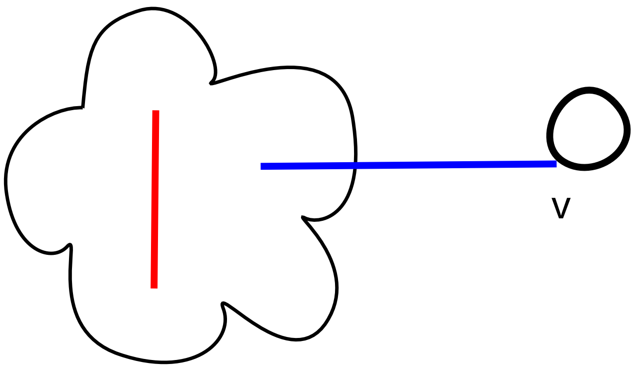
- add expected $\mu^{e,e}$ random random edges to existed network.
- add expected $\mu^{n,e}$ random edges between new vertex and existed network.
- add expected $\mu^{n,n}$ loops to the new vertex.



A general generative model

At time t,

- add expected $\mu^{e,e}$ random random edges to existed network.
- add expected $\mu^{n,e}$ random edges between new vertex and existed network.
- add expected $\mu^{n,n}$ loops to the new vertex.



This is the model D in the reference:

William Aiello, Fan Chung, and Linyuan Lu. Random evolution in massive graphs, *Handbook on Massive Data Sets*, (Eds. James Abello et al.), 97–122. The extended abstract is published in *Proceedings of the Forty-Second Annual Symposium on Foundations of Computer Science*, (2001), 510–519.

Aiello-Chung-Lu's result

Theorem (2001) *For model D, almost surely the degree sequence follows the power law distribution with the power* $2 + \frac{2\mu^{n,n}+\mu^{n,e}}{\mu^{n,e}+2\mu^{e,e}}$. *More precisely, we have*

$$Pr(|d_{i,t} - a_i' t| > 2M'\lambda\sqrt{t}) < e^{-\lambda^2/2},$$

where a_i' *satisfies*

$$a_i' = \frac{a'}{i^{2+\frac{2\mu^{n,n}+\mu^{n,e}}{\mu^{n,e}+2\mu^{e,e}}}}.$$

Here a', M' *are constants determined by distribution of* $(m^{e,e}, m^{n,e}, m^{n,n})$ *of this model, but independent of* i *and* t.

Two approaches

How to show the power law distribution?

Heuristic approach:
Assume $\mathrm{E}(m_{k,t}) \approx M_k t$.
Solve the recurrence.
Done!

Rigorous approach:
Solve the recurrence honestly.
Concentration Properties
...

Two approaches

How to show the power law distribution?

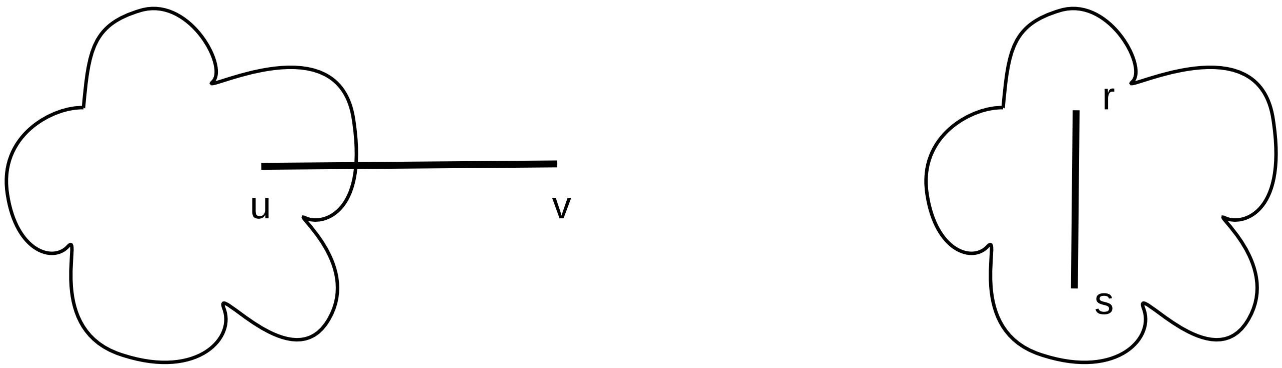
Heuristic approach:	Rigorous approach:
Assume $\mathrm{E}(m_{k,t}) \approx M_k t$.	Solve the recurrence honestly.
Solve the recurrence.	Concentration Properties
Done!	...

In the rest of talk, we will show how the rigorous proof can be done through a simple model $G(p)$.

A generative model $G(p)$.

$$G_0 \to G_1 \to G_2 \to \ldots \to G_t \to \ldots$$

Operations



Vertex step

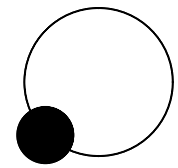
Edge step

Vertex u, r, s are randomly selected with probability proportional to their current degrees.

A generative model $G(p)$

Parameter p: $0 < p < 1$.

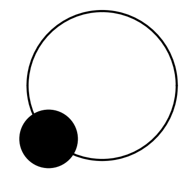
Initial graph G_0



A generative model $G(p)$

Parameter p: $0 < p < 1$.

Initial graph G_0



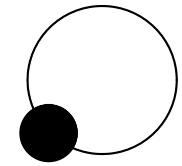
At time t, G_t is formed by modifying from G_{t-1} as follows

- With probability p, take a vertex-step.
- With probability $1 - p$, take a edge-step.

A generative model $G(p)$

Parameter p: $0 < p < 1$.

Initial graph G_0



At time t, G_t is formed by modifying from G_{t-1} as follows

- With probability p, take a vertex-step.
- With probability $1 - p$, take a edge-step.

The number of edges:

$$e_t = 1 + t.$$

The number of vertices n_t

$$n_t = 1 + \sum_{i=1}^{t} s_i$$

where

$$\begin{aligned} Pr(s_j = 1) &= p, \\ Pr(s_j = 0) &= 1 - p. \end{aligned}$$

The number of vertices n_t

$$n_t = 1 + \sum_{i=1}^{t} s_i$$

where

$$\begin{aligned} Pr(s_j = 1) &= p, \\ Pr(s_j = 0) &= 1 - p. \end{aligned}$$

The expected value is

$$\mathrm{E}(n_t) = 1 + tp.$$

The number of vertices n_t

$$n_t = 1 + \sum_{i=1}^{t} s_i$$

where

$$\begin{aligned} Pr(s_j = 1) &= p, \\ Pr(s_j = 0) &= 1 - p. \end{aligned}$$

The expected value is

$$\mathrm{E}(n_t) = 1 + tp.$$

Concentration property?

Chernoff inequality (1981)

Let $X_1, \ldots, X_n$ *be independent random variables with*

$$Pr(X_i = 1) = p_i, \qquad Pr(X_i = 0) = 1 - p_i.$$

We consider the sum $X = \sum_{i=1}^n X_i$*, with expectation* $\mathrm{E}(X) = \sum_{i=1}^n p_i$*. Then we have*

$$\begin{aligned} \textit{(Lower tail)} \qquad Pr(X \leq E(X) - \lambda) \;&\leq\; e^{-\lambda^2/2\mathrm{E}(X)}, \\ \textit{(Upper tail)} \qquad Pr(X \geq E(X) + \lambda) \;&\leq\; e^{-\frac{\lambda^2}{2(\mathrm{E}(X)+\lambda/3)}}. \end{aligned}$$

Chernoff inequality (1981)

Let $X_1, \ldots, X_n$ *be independent random variables with*

$$Pr(X_i = 1) = p_i, \qquad Pr(X_i = 0) = 1 - p_i.$$

We consider the sum $X = \sum_{i=1}^n X_i$, *with expectation* $\mathrm{E}(X) = \sum_{i=1}^n p_i$. *Then we have*

$$\begin{array}{lrcl} \textit{(Lower tail)} & Pr(X \leq E(X) - \lambda) & \leq & e^{-\lambda^2/2\mathrm{E}(X)}, \\ \textit{(Upper tail)} & Pr(X \geq E(X) + \lambda) & \leq & e^{-\frac{\lambda^2}{2(\mathrm{E}(X)+\lambda/3)}}. \end{array}$$

We can show n_t is exponentially concentrated around $\mathrm{E}(n_t)$.

Notations

$m_{k,t}$: the number of vertices of degree k at time t.
Initially

$$m_{2,0} = 1 \text{ and } m_{0,t} = 0.$$

$\mathcal{F}_t$: the σ-algebra associated with the probability space at time t.

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_t.$$

Conditional probability identity:

$$\mathrm{E}(\mathrm{E}(X \mid \mathcal{F}_t)) = \mathrm{E}(X).$$

Recurrence formula for $m_{1,t}$

Case	Probability	Contribution
A new vertex	p	$+1$
$d_u^{t-1} = 1 \rightarrow d_u^t = 2$	$\frac{2-p}{2t}$	-1

Recurrence formula for $m_{1,t}$

Case	Probability	Contribution
A new vertex	p	$+1$
$d_u^{t-1} = 1 \to d_u^t = 2$	$\frac{2-p}{2t}$	-1

For $t > 0$ and $k = 1$, we have

$$E(m_{1,t}|\mathcal{F}_{t-1}) = m_{1,t-1}(1 - \frac{(2-p)}{2t}) + p.$$

Recurrence formula for $m_{1,t}$

Case	Probability	Contribution
A new vertex	p	$+1$
$d_u^{t-1} = 1 \to d_u^t = 2$	$\frac{2-p}{2t}$	-1

For $t > 0$ and $k = 1$, we have

$$E(m_{1,t}|\mathcal{F}_{t-1}) = m_{1,t-1}(1 - \frac{(2-p)}{2t}) + p.$$

Thus,

$$\mathrm{E}(m_{1,t}) = \mathrm{E}(m_{1,t-1})(1 - \frac{(2-p)}{2t}) + p.$$

Recurrence formula for $m_{k,t}$

Case	Probability	Contribution
$d_u^{t-1} = k-1 \rightarrow d_u^t = k$	$(2-p)\frac{k-1}{2t}$	$+1$
$d_u^{t-1} = k \rightarrow d_u^t = k+1$	$(2-p)\frac{k}{2t}$	-1

Recurrence formula for $m_{k,t}$

Case	Probability	Contribution
$d_u^{t-1} = k-1 \rightarrow d_u^t = k$	$(2-p)\frac{k-1}{2t}$	$+1$
$d_u^{t-1} = k \rightarrow d_u^t = k+1$	$(2-p)\frac{k}{2t}$	-1

For $t > 0$ and $k > 1$, we have

$$\begin{aligned} \mathrm{E}(m_{k,t} \mid \mathcal{F}_{t-1}) &= m_{k,t-1}\left(1 - \frac{(2-p)2k}{2t}\right) \\ &\quad + m_{k-1,t-1}\left(\frac{(2-p)(k-1)}{2t}\right). \end{aligned}$$

Recurrence formula for $m_{k,t}$

Case	Probability	Contribution
$d_u^{t-1} = k-1 \to d_u^t = k$	$(2-p)\frac{k-1}{2t}$	$+1$
$d_u^{t-1} = k \to d_u^t = k+1$	$(2-p)\frac{k}{2t}$	-1

For $t > 0$ and $k > 1$, we have

$$\begin{aligned} \mathrm{E}(m_{k,t}|\mathcal{F}_{t-1}) &= m_{k,t-1}(1 - \frac{(2-p)2k}{2t}) \\ &\quad + m_{k-1,t-1}(\frac{(2-p)(k-1)}{2t}). \end{aligned}$$

Thus,
$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1 - \frac{(2-p)2k}{2t}) + \mathrm{E}(m_{k-1,t-1})(\frac{(2-p)(k-1)}{2t})$.

A useful lemma for rigorous proofs

Lemma: *Suppose* $\{a_t\}$ *satisfies* $a_{t+1} = (1 - \frac{b_t}{t+t_1})a_t + c_t$ *for* $t \geq t_0$. $\lim_{t\to\infty} b_t = b > 0$ *and* $\lim_{t\to\infty} c_t = c$. *Then*

$$\lim_{t\to\infty} \frac{a_t}{t} \text{ exists and } \lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

A useful lemma for rigorous proofs

Lemma: *Suppose* $\{a_t\}$ *satisfies* $a_{t+1} = (1 - \frac{b_t}{t+t_1})a_t + c_t$ *for* $t \geq t_0$. $\lim_{t\to\infty} b_t = b > 0$ *and* $\lim_{t\to\infty} c_t = c$. *Then*

$$\lim_{t\to\infty} \frac{a_t}{t} \text{ exists and } \lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

Proof: Define $s_t = |\frac{a_t}{t} - \frac{c}{1+b}|$. Then

$$s_{t+1} \leq s_t|1 - \frac{1+b_t}{t+1}| + |\frac{(1+b)c_t - (1+b_t)c}{(1+b)(1+t)}|.$$

$|s_{t+1} - \epsilon| \leq |s_t - \epsilon|(1-\epsilon)$ for large t.

Solve the case $k = 1$

$$\mathrm{E}(m_{1,t}) = \mathrm{E}(m_{1,t-1})(1 - \frac{(2-p)}{2t}) + p.$$

Solve the case $k = 1$

$$\mathrm{E}(m_{1,t}) = \mathrm{E}(m_{1,t-1})(1 - \frac{(2-p)}{2t}) + p.$$

We apply the lemma with

$$\begin{aligned} a_t &= \mathrm{E}(m_{1,t}), \\ b_t &= b = (2-p)/2, \\ c_t &= c = p. \end{aligned}$$

We have $\lim_{t\to\infty} E(m_{1,t})/t$ exists and

$$M_1 = \lim_{t\to\infty} \frac{E(m_{1,t})}{t} = \frac{2p}{4-p}.$$

Solve the case $k > 1$

$$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1 - \tfrac{(2-p)2k}{2t}) + \mathrm{E}(m_{k-1,t-1})(\tfrac{(2-p)(k-1)}{2t}).$$

Solve the case $k > 1$

$$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1 - \frac{(2-p)2k}{2t}) + \mathrm{E}(m_{k-1,t-1})(\frac{(2-p)(k-1)}{2t}).$$

We apply the lemma with $a_t = \mathrm{E}(m_{k,t})$,

$$\begin{aligned} b_t &= b = k(2-p)/2, \\ c_t &= E(m_{k-1,t-1})(2-p)(k-1)/(2t). \end{aligned}$$

Solve the case $k > 1$

$$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1 - \frac{(2-p)2k}{2t}) + \mathrm{E}(m_{k-1,t-1})(\frac{(2-p)(k-1)}{2t}).$$

We apply the lemma with $a_t = \mathrm{E}(m_{k,t})$,

$$\begin{aligned} b_t &= b = k(2-p)/2, \\ c_t &= E(m_{k-1,t-1})(2-p)(k-1)/(2t). \end{aligned}$$

$\lim_{t\to\infty} c_t = M_{k-1}(2-k)(k-1)\frac{1}{2}$ (inductive hypothesis)

Solve the case $k > 1$

$\mathrm{E}(m_{k,t}) = \mathrm{E}(m_{k,t-1})(1 - \frac{(2-p)2k}{2t}) + \mathrm{E}(m_{k-1,t-1})(\frac{(2-p)(k-1)}{2t})$.

We apply the lemma with $a_t = \mathrm{E}(m_{k,t})$,

$$\begin{aligned} b_t &= b = k(2-p)/2, \\ c_t &= E(m_{k-1,t-1})(2-p)(k-1)/(2t). \end{aligned}$$

$\lim_{t\to\infty} c_t = M_{k-1}(2-k)(k-1)\frac{1}{2}$ (inductive hypothesis)

The limit $\lim_{t\to\infty} E(m_{k,t})/t$ exists and is equal to

$$M_k = M_{k-1}\frac{(2-p)(k-1)}{2+k(2-p)} = M_{k-1}\frac{k-1}{k+\frac{2}{2-p}}.$$

The function $\Gamma(s)$

Recall $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$.

$$\Gamma(s) = s\Gamma(s-1).$$

The function $\Gamma(s)$

Recall $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$.

$$\Gamma(s) = s\Gamma(s-1).$$

Stirling formula

$$\Gamma(x) = (1 + O(\frac{1}{x}))\frac{\sqrt{2\pi}}{\sqrt{x}}(\frac{x}{e})^x.$$

The function $\Gamma(s)$

Recall $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$.

$$\Gamma(s) = s\Gamma(s-1).$$

Stirling formula

$$\Gamma(x) = (1 + O(\frac{1}{x}))\frac{\sqrt{2\pi}}{\sqrt{x}}(\frac{x}{e})^x.$$

$$\begin{aligned} \frac{\Gamma(x)}{\Gamma(x+p)} &= (1 + O(\frac{1}{x}))\frac{\sqrt{x+p}}{\sqrt{x}} \frac{(\frac{x}{e})^x}{(\frac{x+p}{e})^{x+p}} \\ &= (1 + O(\frac{1}{x}))x^p. \end{aligned}$$

Power Law

We have

$$M_k = \frac{2p}{4-p}\prod_{j=2}^{k}\frac{j-1}{j+\frac{2}{2-p}}$$

Power Law

We have

$$
\begin{aligned}
M_k &= \frac{2p}{4-p}\prod_{j=2}^{k}\frac{j-1}{j+\frac{2}{2-p}} \\
&= \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(2+\frac{2}{2-p})}{\Gamma(k+2+\frac{p}{2-p})}
\end{aligned}
$$

Power Law

We have

$$
\begin{aligned}
M_k &= \frac{2p}{4-p}\prod_{j=2}^{k}\frac{j-1}{j+\frac{2}{2-p}} \\
&= \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(2+\frac{2}{2-p})}{\Gamma(k+2+\frac{p}{2-p})} \\
&\approx \frac{2p}{4-p}\Gamma(2+\frac{2}{2-p})k^{-(2+\frac{p}{2-p})}.
\end{aligned}
$$

Power Law

We have

$$
\begin{aligned}
M_k &= \frac{2p}{4-p}\prod_{j=2}^{k}\frac{j-1}{j+\frac{2}{2-p}} \\
&= \frac{2p}{4-p}\frac{\Gamma(k)\Gamma(2+\frac{2}{2-p})}{\Gamma(k+2+\frac{p}{2-p})} \\
&\approx \frac{2p}{4-p}\Gamma(2+\frac{2}{2-p})k^{-(2+\frac{p}{2-p})}.
\end{aligned}
$$

$\{M_k\}$ follows a power law distribution with $\beta = 2 + \frac{p}{2-p}$.

Are we done?

Are we done?

No.

$$\text{“}\{\mathrm{E}(m_{k,t})\}_k \text{ power law”} \quad \not\Longrightarrow \quad \{m_{k,t}\}_k \text{ power law”}$$

Are we done?

No.

$$\text{“}\{\mathrm{E}(m_{k,t})\}_k \text{ power law”} \quad \not\Longrightarrow \quad \{m_{k,t}\}_k \text{ power law”}$$

We need prove $m_{k,t}$ concentrates on $\mathrm{E}(m_{k,t})$.

Our result

Chung, Lu *For the preferential attachment model* $G(p)$*, almost surely the number of vertices with degree* k *at time* t *is*

$$M_k t + O(4\sqrt{k^3 t \ln(t)}).$$

Our result

Chung, Lu *For the preferential attachment model $G(p)$, almost surely the number of vertices with degree k at time t is*

$$M_k t + O(4\sqrt{k^3 t \ln(t)}).$$

In other words, almost surely the graphs generated by $G(p)$ have the power law degree distribution with the exponent $\beta = 2 + \frac{p}{2-p}$.

A claim

Claim: For $k \geq 1$, $c > 0$, with probability at least $1 - 2(t+1)^{k-1}e^{-c^2}$, we have

$$|m_{k,t} - M_k(t+1)| \leq 4kc\sqrt{t}.$$

A claim

Claim: For $k \geq 1$, $c > 0$, with probability at least $1 - 2(t+1)^{k-1}e^{-c^2}$, we have

$$|m_{k,t} - M_k(t+1)| \leq 4kc\sqrt{t}.$$

Choose $c = \sqrt{k \ln t}$. Note that

$$2(t+1)^{k-1}e^{-c^2} = 2(t+1)^{k-1}t^{-k} = o(1).$$

A claim

Claim: For $k \geq 1$, $c > 0$, with probability at least $1 - 2(t+1)^{k-1}e^{-c^2}$, we have

$$|m_{k,t} - M_k(t+1)| \leq 4kc\sqrt{t}.$$

Choose $c = \sqrt{k \ln t}$. Note that

$$2(t+1)^{k-1}e^{-c^2} = 2(t+1)^{k-1}t^{-k} = o(1).$$

From the Claim, with probability $1 - o(1)$, we have

$$|m_{k,t} - M_k(t+1)| \leq 4\sqrt{k^3 t \ln t},$$

as desired.

Martingale inequality

A martingale is a sequence of random variables $X_0, X_1, \ldots$ so that

$$\mathrm{E}(X_{n+1} \mid X_0, X_1, \ldots, X_n) = X_n.$$

Martingale inequality

A martingale is a sequence of random variables $X_0, X_1, \ldots$ so that

$$E(X_{n+1} \mid X_0, X_1, \ldots, X_n) = X_n.$$

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, the martingale X is said to be c-Lipschitz if

$$|X_i - X_{i-1}| \leq c_i \text{ for } i = 1, 2, \ldots, n.$$

Martingale inequality

A martingale is a sequence of random variables $X_0, X_1, \ldots$ so that

$$E(X_{n+1} \mid X_0, X_1, \ldots, X_n) = X_n.$$

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, the martingale X is said to be c-Lipschitz if

$$|X_i - X_{i-1}| \leq c_i \text{ for } i = 1, 2, \ldots, n.$$

Azuma's martingale inequality:
If a martingale X is c-Lipschitz, then

$$\Pr(|X - E(X)| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}.$$

Proof of the claim for $k = 1$

Rewrite recursive formula as

$$\mathrm{E}(m_{1,t} - M_1(t+1)|\mathcal{F}_{t-1}) = (m_{1,t-1} - M_1 t)(1 - \frac{2-p}{2t}).$$

Proof of the claim for $k = 1$

Rewrite recursive formula as

$$\mathrm{E}(m_{1,t} - M_1(t+1) | \mathcal{F}_{t-1}) = (m_{1,t-1} - M_1 t)(1 - \frac{2-p}{2t}).$$

Let $X_{1,t} = \frac{m_{1,t} - M_1(t+1)}{\prod_{j=1}^{t}(1 - \frac{2-p}{2j})}$. $1 = X_{1,0}, X_{1,1}, \cdots, X_{1,t}$ forms a martingale.

Proof of the claim for $k = 1$

Rewrite recursive formula as

$$\mathrm{E}(m_{1,t} - M_1(t+1)|\mathcal{F}_{t-1}) = (m_{1,t-1} - M_1 t)(1 - \frac{2-p}{2t}).$$

Let $X_{1,t} = \frac{m_{1,t} - M_1(t+1)}{\prod_{j=1}^{t}(1 - \frac{2-p}{2j})}$. $1 = X_{1,0}, X_{1,1}, \cdots, X_{1,t}$ forms a martingale.
We can show

$$|X_i - X_{i-1}| \leq \frac{4}{\prod_{j=1}^{t}(1 - \frac{2-p}{2j})}.$$

continue...

Let $c_i = \frac{4}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}$. We have

$$\begin{aligned} \sum_{i=1}^{t} c_i^2 &= \sum_{i=1}^{t} \frac{16}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})^2} \\ &= 16\sum_{i=1}^{t}(\Gamma(\frac{p}{2})^2) + O(\frac{1}{i}))i^{2-p} \\ &\approx \frac{16\Gamma^2(\frac{p}{2})}{3-p} t^{3-p} \\ &< 8\Gamma^2(\frac{p}{2})t^{3-p}. \end{aligned}$$

Apply martingale inequality

Choose $\lambda = c\sqrt{2\sum_{i=1}^{t} c_i^2}$. We have

$$\Pr(|X_{1,t} - \mathrm{E}(X_{1,t})| \geq \lambda) \leq e^{-c^2}.$$

Apply martingale inequality

Choose $\lambda = c\sqrt{2\sum_{i=1}^{t} c_i^2}$. We have

$$\Pr(|X_{1,t} - \mathrm{E}(X_{1,t})| \geq \lambda) \leq e^{-c^2}.$$

With probability at least $1 - e^{-c^2}$,

$$\begin{aligned} |m_{1,t} - M_1(t+1)| &= |X_{1,t} - \mathrm{E}(X_{1,t})| \prod_{i=1}^{t}(1 - \frac{2-p}{2j}) \\ &\leq \lambda \prod_{i=1}^{t}(1 - \frac{2-p}{2j})) \\ &\approx 4c\sqrt{t}. \qquad \square \end{aligned}$$

Induction on k

By the induction hypothesis, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$|m_{k-1,t-1} - M_{k-1}t| \leq 4(k-1)c\sqrt{t-1}.$$

Induction on k

By the induction hypothesis, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$|m_{k-1,t-1} - M_{k-1}t| \leq 4(k-1)c\sqrt{t-1}.$$

For k, we define

$$X_{k,t} = \frac{m_{k,t} - M_k(t+1) - 4(k-1)c\sqrt{t}}{\prod_{j=1}^{t}(1 - \frac{(2-p)k}{2j})}.$$

Induction on k

By the induction hypothesis, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$|m_{k-1,t-1} - M_{k-1}t| \leq 4(k-1)c\sqrt{t-1}.$$

For k, we define

$$X_{k,t} = \frac{m_{k,t} - M_k(t+1) - 4(k-1)c\sqrt{t}}{\prod_{j=1}^{t}(1 - \frac{(2-p)k}{2j})}.$$

Therefore, $0 = X_{k,0}, X_{k,1}, \cdots, X_{k,t}$ forms a *submartingale* with fail probability at most $2t^{k-2}e^{-c^2}$.

Submartingale

For a filter $\mathbf{F}$:

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

$X_0, X_1, \ldots, X_n$ is called a *submartingale* if

- X_i is $\mathcal{F}_i$-measurable,
- $E(X_i \mid \mathcal{F}_{i-1}) \leq X_{i-1}$, for $1 \leq i \leq n$.

Submartingale inequality

Suppose that a submartingale X *associated with a filter* $\mathbf{F}$*, satisfies*

$$\mathrm{Var}(X_i \mid \mathcal{F}_{i-1}) \leq \sigma_i^2$$

$$X_i - E(X_i \mid \mathcal{F}_{i-1}) \leq C$$

for $1 \leq i \leq n$ *with exceptional set* B_i*. Then*

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n \sigma_i^2 + C\lambda/3)}} + \sum_{i=1}^n \Pr(B_i).$$

put together

Applying submartingale inequality, $\sigma_i^2 = \frac{4}{\prod_{j=1}^{i}(1-\frac{(2-p)k}{2j})^2}$, and $C = \frac{4}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}$.

put together

Applying submartingale inequality, $\sigma_i^2 = \frac{4}{\prod_{j=1}^{i}(1-\frac{(2-p)k}{2j})^2}$, and $C = \frac{4}{\prod_{j=1}^{t}(1-\frac{2-p}{2j})}$. We have

$$\begin{aligned}
\Pr(X_{k,t} \geq \mathrm{E}(X_{k,t}) + \lambda) &\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2 + M\lambda/3)}} + \sum_{i=1}^{n} \Pr(B_i) \\
&\leq e^{-c^2} + \sum_{i=1}^{n} (i+1)^{k-2} e^{-c^2} \\
&\leq (t+1)^{k-1} e^{-c^2}.
\end{aligned}$$

Here we choose λ properly so that $e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2 + M\lambda/3)}} \leq e^{-c^2}$.

Continue...

Choose $\lambda = \frac{2c\sqrt{t}}{\prod_{j=1}^{t}(1-\frac{(2-p)k}{2j})} \approx 2\Gamma(1-\frac{(2-p)k}{2})ct^{1/2+k(2-p)/2}$.
With probability at least $1-(t+1)^{k-1}e^{-c^2}$, we have

$$m_{k,t} - M_k(t+1) \leq 2kc\sqrt{t}.$$

Similar for the other direction. Thus, With probability at least $1-2(t+1)^{k-1}e^{-c^2}$, we have

$$|m_{k,t} - M_k(t+1)| \leq 2kc\sqrt{t}.$$

The proof of the claim is finished. □

A preferential model $G(p, m, G_0)$

Parameters: $0 < p < 1$. $m \geq 1$, initial graph G_0. At time t, G_t is formed by modifying from G_{t-1} as follows

- With probability p, take a m-vertex-step.
- With probability $1 - p$, take a m-edge-step.

A preferential model $G(p, m, G_0)$

Parameters: $0 < p < 1$. $m \geq 1$, initial graph G_0. At time t, G_t is formed by modifying from G_{t-1} as follows

- With probability p, take a m-vertex-step.
- With probability $1 - p$, take a m-edge-step.

Add m edges at each step.

Result on $G(p, m, G_0)$

Chung, Lu *For $G(p, m, G_0)$, almost surely the number of vertices with degree k at time t is*

$$mM_k t + m_{k,0} + O(4m\sqrt{(k+m-1)^3 t \ln(t)}).$$

Here $M_m = \frac{2p}{4-p}$ and
$M_k = \frac{2p}{4-p}\frac{\Gamma(k-m)\Gamma(1+\frac{2}{2-p})}{\Gamma(k-m+1+\frac{2}{2-p})} = O(k^{-(2+\frac{p}{2-p})})$, *for $k \geq m+1$.*

Result on $G(p, m, G_0)$

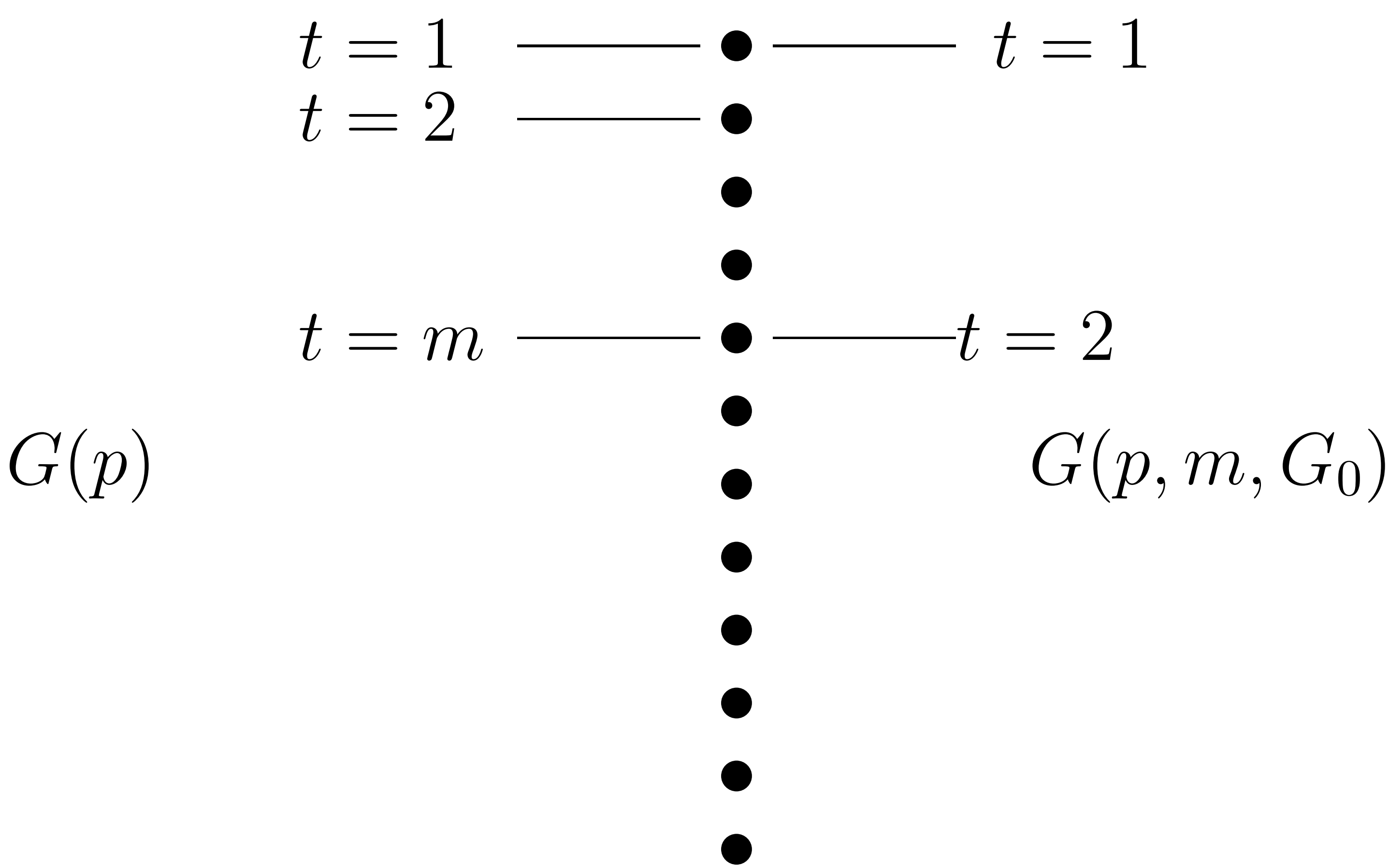
Chung, Lu *For $G(p, m, G_0)$, almost surely the number of vertices with degree k at time t is*

$$mM_k t + m_{k,0} + O(4m\sqrt{(k+m-1)^3 t \ln(t)}).$$

Here $M_m = \frac{2p}{4-p}$ and
$M_k = \frac{2p}{4-p}\frac{\Gamma(k-m)\Gamma(1+\frac{2}{2-p})}{\Gamma(k-m+1+\frac{2}{2-p})} = O(k^{-(2+\frac{p}{2-p})})$, *for $k \geq m+1$.*

In other words, almost surely the graphs generated by $G(p)$ have the power law degree distribution with the exponent $\beta = 2 + \frac{p}{2-p}$.

Scale-free networks



Same power law exponent, different edge density.

Directed graphs

- The WWW graph as a directed graph:

  Kumar et al (1999) and independently Albert and Barabasi (1999) reported that a power law of exponent 2.1 for in-degree distribution and a power law of exponent 2.7 for out-degree distribution.

Directed graphs

- The WWW graph as a directed graph:

  Kumar et al (1999) and independently Albert and Barabasi (1999) reported that a power law of exponent 2.1 for in-degree distribution and a power law of exponent 2.7 for out-degree distribution.

- The call graph also has different power law distributions for in-degrees and out-degrees.

Directed graphs

- The WWW graph as a directed graph:

  Kumar et al (1999) and independently Albert and Barabasi (1999) reported that a power law of exponent 2.1 for in-degree distribution and a power law of exponent 2.7 for out-degree distribution.

- The call graph also has different power law distributions for in-degrees and out-degrees.

How to model directed graphs using preferential attachment scheme?

Model A

- At time 1, add a node with in-weight 1 and out-weight 1.
- A time $t+1$:
  - With probability $1-\alpha$, add a node with in-weight 1 and out-weight 1.
  - With probability α, add an edge uv. Here the origin u is chosen with probability proportional to the current out-weight $w_{u,t}^{out} \stackrel{def}{=} 1 + \delta_{u,t}^{out}$ and the destination v is chosen with probability proportional to the current in-weight $w_{v,t}^{in} \stackrel{def}{=} 1 + \delta_{v,t}^{in}$.

Result on Model A

Aiello, Chung, Lu (2001) For model A, the distribution of in-degree and out-degree sequences follow the power law distribution with power $1 + \frac{1}{\alpha}$. The joint distribution of in-degree and out-degree sequence follows the power law distribution with power $2 + \frac{1}{\alpha}$. More precisely, we have

$$Pr(|d_{i,j,t}^{joint} - a_{i,j}t| > \lambda\sqrt{t} + 2) < e^{-\lambda^2/8},$$

$$Pr(|d_{i,t}^{in} - b_i t| > \lambda\sqrt{t} + 2) < e^{-\lambda^2/2},$$

$$Pr(|d_{j,t}^{out} - c_j t| > \lambda\sqrt{t} + 2) < e^{-\lambda^2/2}.$$

Continue...

where $a_{i,j}, b_i, c_j$ are constants satisfying

$$a_{i,j} = \frac{(1-\alpha)(i+j-2)!\alpha^{i+j-2}}{\prod_{l=2}^{i+j}(1+l\alpha)} = \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+2)}{(i+j)^{\frac{1}{\alpha}+2}} + o_{i+j}(1)$$

$$b_i = \frac{(1-\alpha)!\alpha^{i-1}}{\prod_{l=1}^{i}(1+l\alpha)} = \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+1)}{i^{\frac{1}{\alpha}+1}} + o_i(1)$$

$$c_j = \frac{(1-\alpha)(j-1)!\alpha^{j-1}}{\prod_{l=1}^{j}(1+l\alpha)} = \frac{(\frac{1}{\alpha}-1)\Gamma(\frac{1}{\alpha}+1)}{j^{\frac{1}{\alpha}+1}} + o_j(1)$$

Model B

Two parameters: γ^{in} and γ^{out}.

- At time 1, add a node with in-weight γ^{in} and out-weight γ^{out}.
- A time $t+1$:
  - With probability $1-\alpha$, add a node with in-weight 1 and out-weight 1.
  - With probability α, add an edge uv. Here the origin u is chosen with probability proportional to the current out-weight $w_{u,t}^{out} \stackrel{def}{=} \gamma^{out} + \delta_{u,t}^{out}$ and the destination v is chosen with probability proportional to the current in-weight $w_{v,t}^{in} \stackrel{def}{=} \gamma^{in} + \delta_{v,t}^{in}$.

Result on Model B

Aiello, Chung, Lu (2001) For model B, the distribution of in-degree sequence follows the power law distribution with power $2 + \frac{\gamma^{in}}{\Delta}$, and the distribution of out-degree sequence follows the power law distribution with power $2 + \frac{\gamma^{out}}{\Delta}$. Here $\Delta = \frac{\alpha}{1-\alpha}$ is the asymptotic edge density. More precisely, we have

$$Pr(|d_{i,t}^{in} - b_i' t| > 2\lambda\sqrt{t}) < e^{-\lambda^2/2},$$

$$Pr(|d_{j,t}^{out} - c_j' t| > 2\lambda\sqrt{t}) < e^{-\lambda^2/2}.$$

continue...

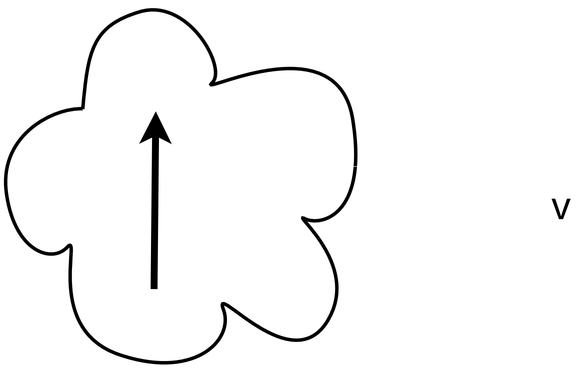
where b'_i, c'_j are constants satisfying

$$
\begin{aligned}
b'_i &= (1-\alpha)\left(\frac{1}{\gamma^{in}}+\frac{1}{\Delta}\right)\prod_{l=1}^{i+1}\frac{l-2+\gamma^{in}}{l+\frac{\gamma^{in}}{\alpha}} \\
&= (1-\alpha)\left(\frac{1}{\gamma^{in}}+\frac{1}{\Delta}\right)\frac{\Gamma(\frac{\gamma^{in}}{\alpha}+1)}{\Gamma(\gamma^{in}-1)}\frac{1}{i^{\frac{\gamma^{in}}{\Delta}+2}}+o_i(1) \\
c'_j &= (1-\alpha)\left(\frac{1}{\gamma^{out}}+\frac{1}{\Delta}\right)\prod_{l=1}^{j+1}\frac{l-2+\gamma^{out}}{l+\frac{\gamma^{out}}{\alpha}} \\
&= (1-\alpha)\left(\frac{1}{\gamma^{out}}+\frac{1}{\Delta}\right)\frac{\Gamma(\frac{\gamma^{out}}{\alpha}+1)}{\Gamma(\gamma^{out}-1)}\frac{1}{j^{\frac{\gamma^{out}}{\Delta}+2}}+o_j(1)
\end{aligned}
$$

Model C

At time t,

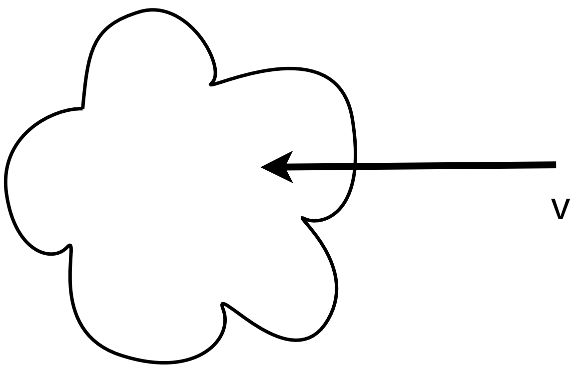
- add expected $\mu^{e,e}$ random random directed edges to existed network.
- add expected $\mu^{n,e}$ random edges from new vertex to existed network.
- add expected $\mu^{e,n}$ random edges from existed network to new vertex.
- add expected $\mu^{n,n}$ loops to the new vertex.



Model C

At time t,

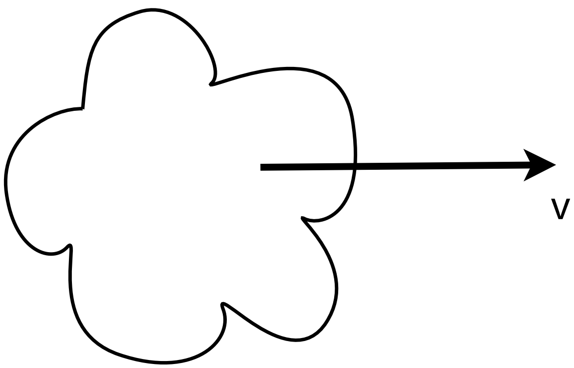
- add expected $\mu^{e,e}$ random random directed edges to existed network.
- add expected $\mu^{n,e}$ random edges from new vertex to existed network.
- add expected $\mu^{e,n}$ random edges from existed network to new vertex.
- add expected $\mu^{n,n}$ loops to the new vertex.



Model C

At time t,

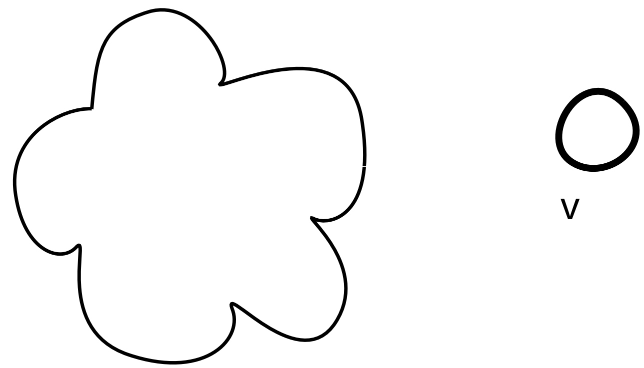
- add expected $\mu^{e,e}$ random random directed edges to existed network.
- add expected $\mu^{n,e}$ random edges from new vertex to existed network.
- add expected $\mu^{e,n}$ random edges from existed network to new vertex.
- add expected $\mu^{n,n}$ loops to the new vertex.



Model C

At time t,

- add expected $\mu^{e,e}$ random random directed edges to existed network.
- add expected $\mu^{n,e}$ random edges from new vertex to existed network.
- add expected $\mu^{e,n}$ random edges from existed network to new vertex.
- add expected $\mu^{n,n}$ loops to the new vertex.



Result on Model C

Aiello, Chung, Lu (2001) For model C, almost surely the out-degree sequence follows the power law distribution with the power $2 + \frac{\mu^{n,n}+\mu^{n,e}}{\mu^{e,n}+\mu^{e,e}}$. Almost surely the in-degree sequence follows the power law distribution with the power $2 + \frac{\mu^{n,n}+\mu^{e,n}}{\mu^{n,e}+\mu^{e,e}}$. More precisely, we have

$$Pr(|d_{i,t}^{in} - b_i''t| > 2M\lambda\sqrt{t}) < e^{-\lambda^2/2},$$

$$Pr(|d_{j,t}^{out} - c_j''t| > 2M\lambda\sqrt{t}) < e^{-\lambda^2/2}.$$

Continue...

where b_i'', c_j'' satisfy

$$b_i'' = \frac{b''}{i^{2+\frac{\mu^{n,n}+\mu^{e,n}}{\mu^{n,e}+\mu^{e,e}}}} + o_i(1),$$

$$c_j'' = \frac{c''}{j^{2+\frac{\mu^{n,n}+\mu^{e,n}}{\mu^{n,e}+\mu^{e,e}}}} + o_j(1).$$

Here b'', c'', M are constants determined by the joint distribution of $m^{e,e}$, $m^{n,e}$, $m^{e,n}$, $m^{n,n}$ of this model, but independent of i and t.

Overview of talks

- Lecture 1: Overview and outlines
- Lecture 2: Generative models - preferential attachment schemes
- Lecture 3: Duplication models for biological networks
- Lecture 4: The rise of the giant component
- Lecture 5: The small world phenomenon: average distance and diameter
- Lecture 6: Spectrum of random graphs with given degrees