Complex Graphs and Networks

Lecture 5: The small world phenomenon: average distance and diameter

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

"Six degree separation"

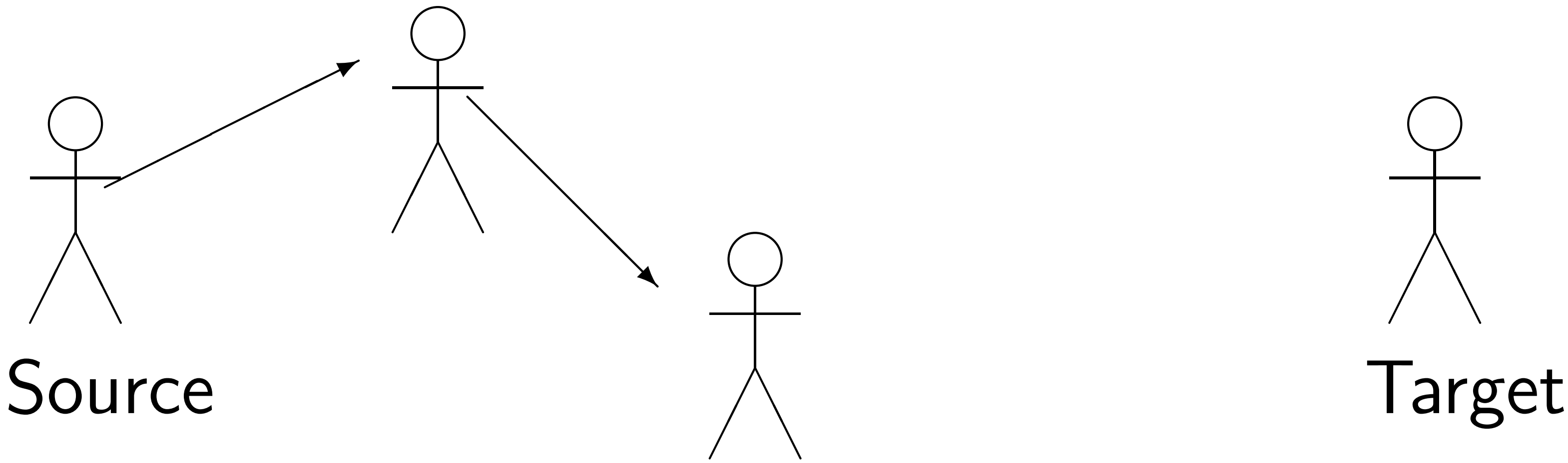
Experiments of Stanley Milgram (1967)

“Six degree separation”

Experiments of Stanley Milgram (1967)

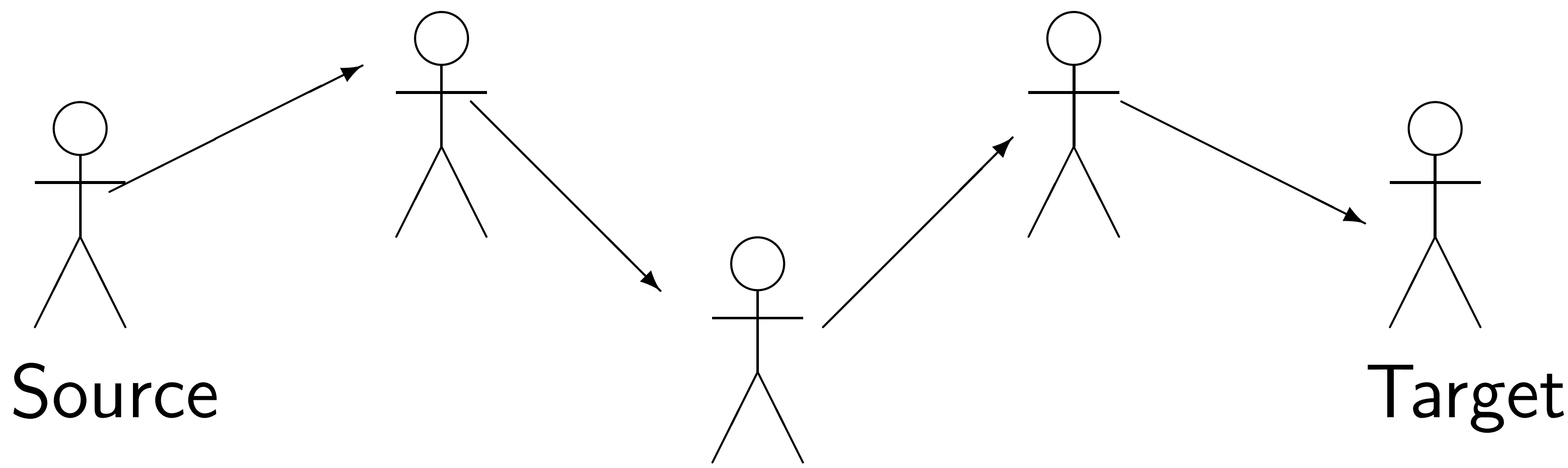
“Six degree separation”

Experiments of Stanley Milgram (1967)



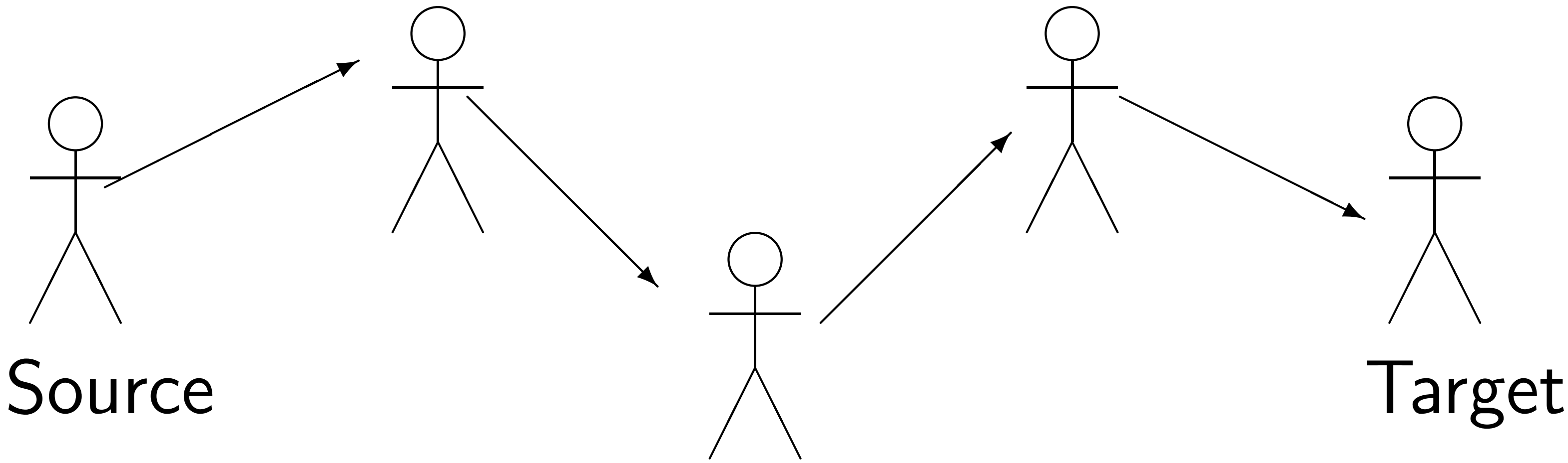
“Six degree separation”

Experiments of Stanley Milgram (1967)



“Six degree separation”

Experiments of Stanley Milgram (1967)



Milgram: “The average distance of the social graph is at most 6.”

Diameter and average distance

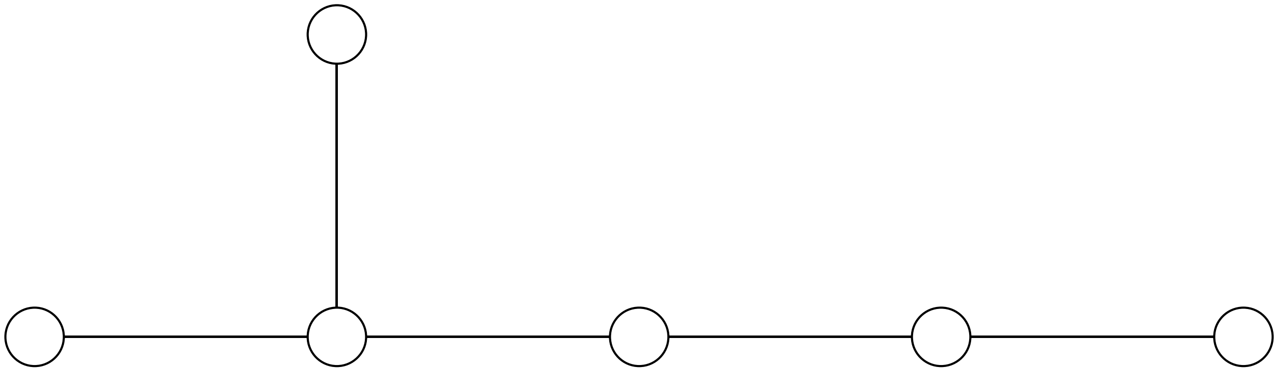
Diameter: the maximum distance $d(u, v)$, where u and v are in the same connected component.

Diameter and average distance

Diameter: the maximum distance $d(u, v)$, where u and v are in the same connected component.
Average distance: the average among all distance $d(u, v)$ for pairs of u and v in the same connected component.

Diameter and average distance

Diameter: the maximum distance $d(u, v)$, where u and v are in the same connected component.
Average distance: the average among all distance $d(u, v)$ for pairs of u and v in the same connected component.



Diameter is 4. Average distance is 2.13.

Experimental results

- The Hollywood graph: $n \approx 656,065$. The average Bacon number is 2.94. The maximum Bacon number is 9.

Experimental results

- The Hollywood graph: $n \approx 656,065$. The average Bacon number is 2.94. The maximum Bacon number is 9.

- The Collaboration graph: $n \approx 337,000$. The diameter is 27. The average distance is 7.73.

Experimental results

- The Hollywood graph: $n \approx 656,065$. The average Bacon number is 2.94. The maximum Bacon number is 9.
- The Collaboration graph: $n \approx 337,000$. The diameter is 27. The average distance is 7.73.
- The WWW subgraph: **Barabási (1999)** the diameter is 19.

Experimental results

- The Hollywood graph: $n \approx 656,065$. The average Bacon number is 2.94. The maximum Bacon number is 9.

- The Collaboration graph: $n \approx 337,000$. The diameter is 27. The average distance is 7.73.

- The WWW subgraph: **Barabási (1999)** the diameter is 19.

- The WWW subgraph: $n \approx 203,000,000$. **Kumar et al. (2001)** The diameter is about 500 (as a directed graph).

Experimental results

- The Hollywood graph: $n \approx 656,065$. The average Bacon number is 2.94. The maximum Bacon number is 9.
- The Collaboration graph: $n \approx 337,000$. The diameter is 27. The average distance is 7.73.
- The WWW subgraph: **Barabási (1999)** the diameter is 19.
- The WWW subgraph: $n \approx 203,000,000$. **Kumar et al. (2001)** The diameter is about 500 (as a directed graph).

Many real-world graphs have small diameters comparing to its sizes.

Disadvantage of experimental methods

- Case by case

Disadvantage of experimental methods

- Case by case
- Inadequate information

Disadvantage of experimental methods

- Case by case
- Inadequate information
- Dynamically changing

Disadvantage of experimental methods

- Case by case
- Inadequate information
- Dynamically changing
- Prohibitively large sizes

Questions

What is the magnitude of the diameter and the average distance with respect to the graph size?

Questions

What is the magnitude of the diameter and the average distance with respect to the graph size?

How to characterize these graphs?

Modelling graphs

We will use random graphs to model real-world graphs because

- Data sets are too large and dynamic for exact analysis.
- Most real-world graphs have a random or statistical nature.

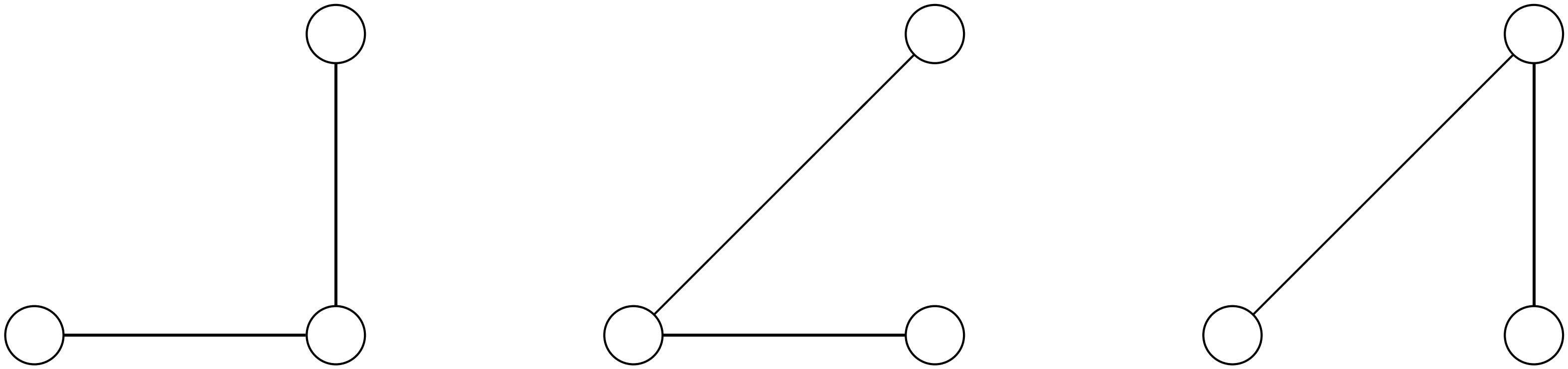
Random graphs

A random graph is a set of graphs together with a probability distribution on that set.

Random graphs

A random graph is a set of graphs together with a probability distribution on that set.

Example: A random graph on 3 vertices and 2 edges with the uniform distribution on it.

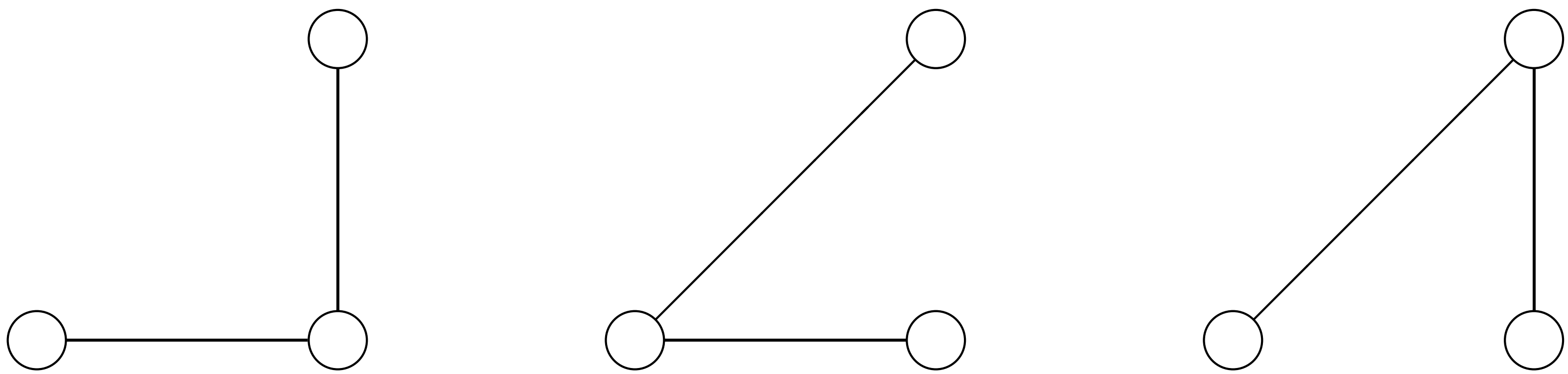


Probability $\frac{1}{3}$ Probability $\frac{1}{3}$ Probability $\frac{1}{3}$

Random graphs

A random graph is a set of graphs together with a probability distribution on that set.

Example: A random graph on 3 vertices and 2 edges with the uniform distribution on it.



Probability $\frac{1}{3}$ Probability $\frac{1}{3}$ Probability $\frac{1}{3}$

A random graph G *almost surely* satisfies a property P, if

$$Pr(G \text{ satisfies } P) = 1 - o_n(1).$$

Erdős-Rényi model $G(n, p)$

- n nodes

Erdős-Rényi model $G(n, p)$

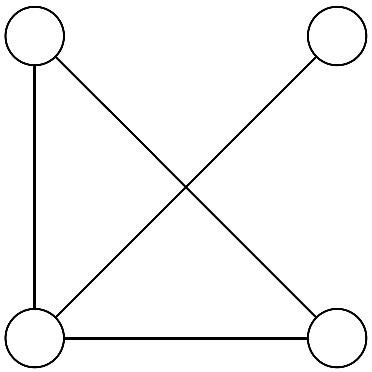
- n nodes
- For each pair of vertices, create an edge independently with probability p.

Erdős-Rényi model $G(n, p)$

- n nodes
- For each pair of vertices, create an edge independently with probability p.
- The graph with e edges has the probability $p^e(1-p)^{\binom{n}{2}-e}$.

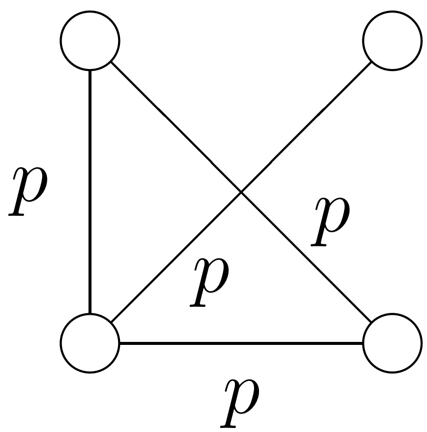
Erdős-Rényi model $G(n,p)$

- n nodes
- For each pair of vertices, create an edge independently with probability p.
- The graph with e edges has the probability $p^e(1-p)^{\binom{n}{2}-e}$.



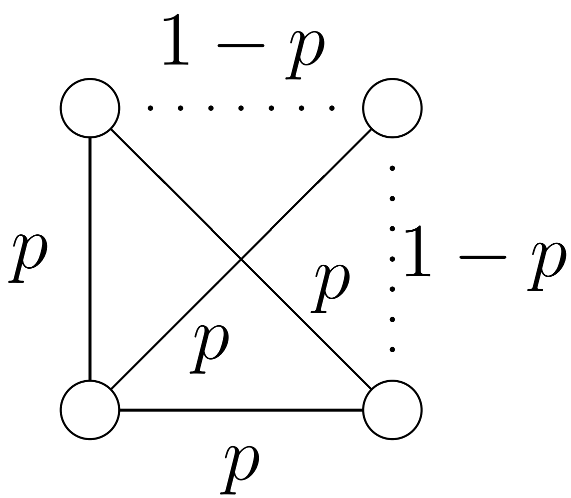
Erdős-Rényi model $G(n,p)$

- n nodes
- For each pair of vertices, create an edge independently with probability p.
- The graph with e edges has the probability $p^e(1-p)^{\binom{n}{2}-e}$.



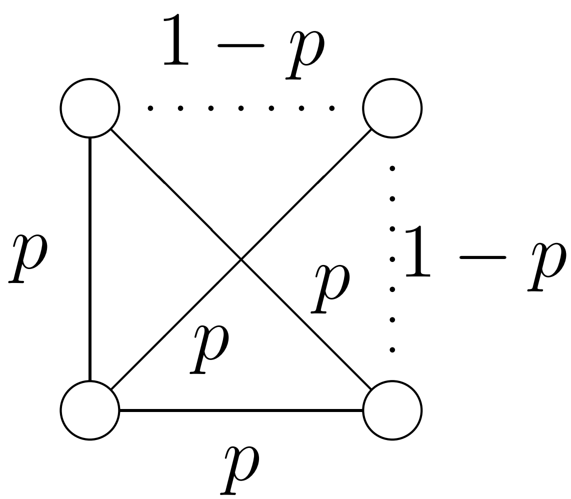
Erdős-Rényi model $G(n,p)$

- n nodes
- For each pair of vertices, create an edge independently with probability p.
- The graph with e edges has the probability $p^e(1-p)^{\binom{n}{2}-e}$.



Erdős-Rényi model $G(n, p)$

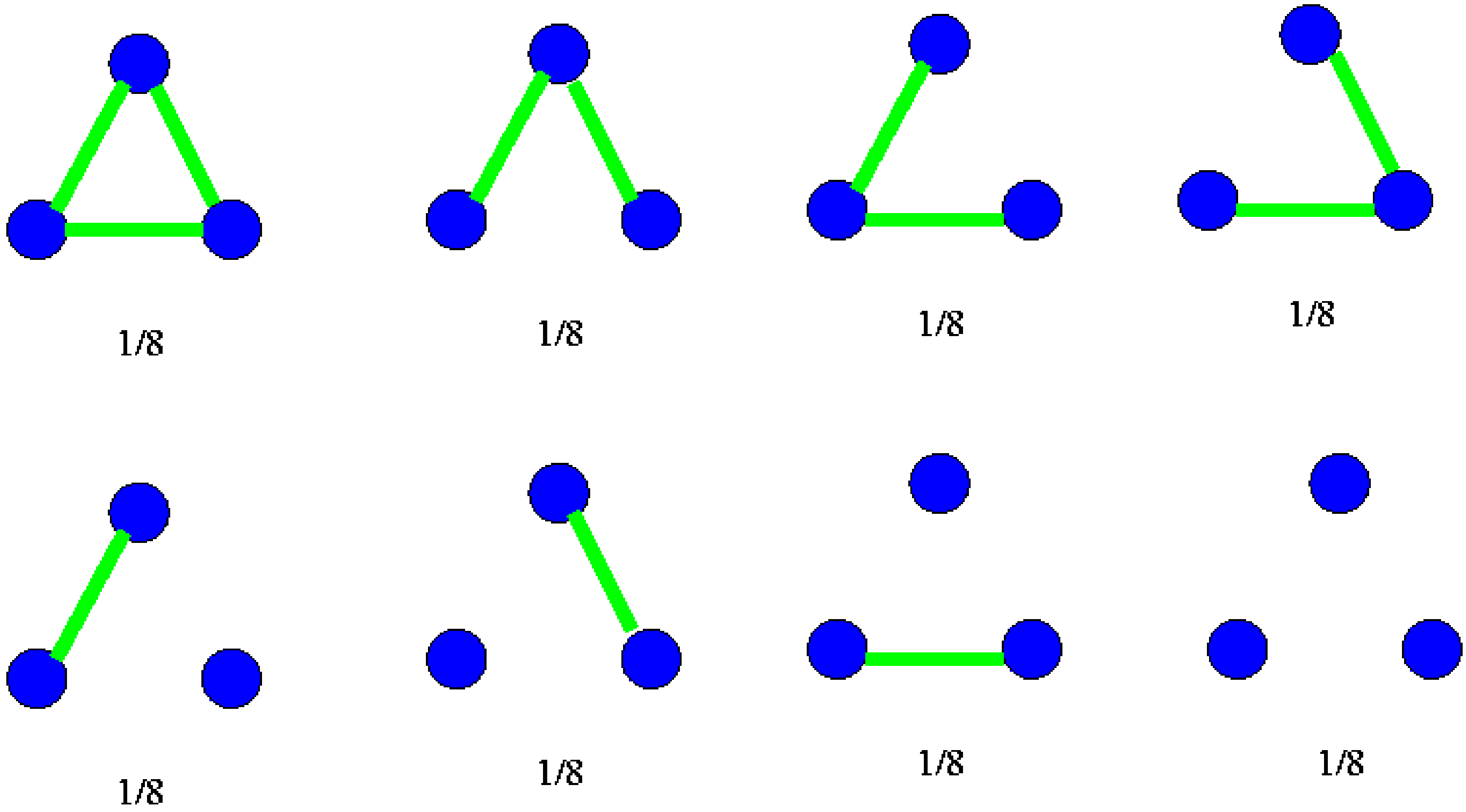
- n nodes
- For each pair of vertices, create an edge independently with probability p.
- The graph with e edges has the probability $p^e(1-p)^{\binom{n}{2}-e}$.



The probability of this graph is

$$p^4(1-p)^2.$$

A example: $G(3, \frac{1}{2})$



The birth of random graph theory

Paul Erdős and A. Rényi, On the evolution of random graphs
Magyar Tud. Akad. Mat. Kut. Int. Kozl. **5** (1960) 17-61.

The birth of random graph theory

ON THE EVOLUTION OF RANDOM GRAPHS

by

P. ERDÖS and A. RÉNYI
Institute of Mathematics
Hungarian Academy of Sciences, Hungary

1. Definition of a random graph

Let $E_{n,N}$ denote the set of all graphs having n given labelled vertices $V_1, V_2, \cdots,$ V_n and N edges. The graphs considered are supposed to be not oriented, without parallel edges and without slings (such graphs are sometimes called linear graphs). Thus a graph belonging to the set $E_{n,N}$ is obtained by choosing N out of the possible $\binom{n}{2}$ edges between the points $V_1, V_2, \cdots, V_n$, and therefore the number of elements of $E_{n,N}$ is equal to $\binom{\binom{n}{2}}{N}$. A random graph $\Gamma_{n,N}$ can be defined as an element of $E_{n,N}$ chosen at random, so that each of the elements of $E_{n,N}$ have the same probability to be chosen, namely $1/\binom{\binom{n}{2}}{N}$. There is however an other slightly

Evolution of $G(n, p)$

p	
0	the diameter.
$\frac{c}{n}$	$\Theta(\frac{\log n}{\log np})$
$\omega(\frac{1}{n})$	$(1 + o(1))\frac{\log n}{\log np}$
$\omega(\frac{\log n}{n})$	$\lfloor \frac{\log n}{\log np} \rfloor$ or $\lceil \frac{\log n}{\log np} \rceil$
$\vdots$	$\vdots$
$n^{-3/4}$	
	3 or 4
$n^{-2/3}$	
	2 or 3
$n^{-\frac{1}{2}}$	
	2
1	1

Diameter of $G(n,p)$

Bollobás (1985): (denser graph)

$$diam(G(n,p)) = \lfloor \frac{\log n}{\log np} \rfloor \text{ or } \lceil \frac{\log n}{\log np} \rceil \text{ if } np \gg \log n.$$

Diameter of $G(n,p)$

Bollobás (1985): (denser graph)

$$diam(G(n,p)) = \lfloor \frac{\log n}{\log np} \rfloor \text{ or } \lceil \frac{\log n}{\log np} \rceil \text{ if } np \gg \log n.$$

Chung Lu, (2000) (Sparser graph)

$$diam(G(n,p)) = \begin{cases} (1+o(1))\frac{\log n}{\log np} & \text{if } np \to \infty \\ \Theta(\frac{\log n}{\log np}) & \text{if } \infty > np > 1. \end{cases}$$

Model $G(w_1, w_2, \ldots, w_n)$

Random graph model with given expected degree sequence

- n nodes with weights $w_1, w_2, \ldots, w_n$.

Model $G(w_1, w_2, \ldots, w_n)$

Random graph model with given expected degree sequence

- n nodes with weights $w_1, w_2, \ldots, w_n$.
- For each pair (i, j), create an edge independently with probability $p_{ij} = w_i w_j \rho$, where $\rho = \frac{1}{\sum_{i=1}^{n} w_i}$.

Model $G(w_1, w_2, \ldots, w_n)$

Random graph model with given expected degree sequence

- n nodes with weights $w_1, w_2, \ldots, w_n$.
- For each pair (i, j), create an edge independently with probability $p_{ij} = w_i w_j \rho$, where $\rho = \frac{1}{\sum_{i=1}^n w_i}$.
- The graph H has probability

$$\prod_{ij \in E(H)} p_{ij} \prod_{ij \notin E(H)} (1 - p_{ij}).$$

Model $G(w_1, w_2, \ldots, w_n)$

Random graph model with given expected degree sequence

- n nodes with weights $w_1, w_2, \ldots, w_n$.
- For each pair (i, j), create an edge independently with probability $p_{ij} = w_i w_j \rho$, where $\rho = \frac{1}{\sum_{i=1}^n w_i}$.
- The graph H has probability

$$\prod_{ij \in E(H)} p_{ij} \prod_{ij \notin E(H)} (1 - p_{ij}).$$

- The expected degree of vertex i is w_i.

Model $G(w_1, w_2, \ldots, w_n)$

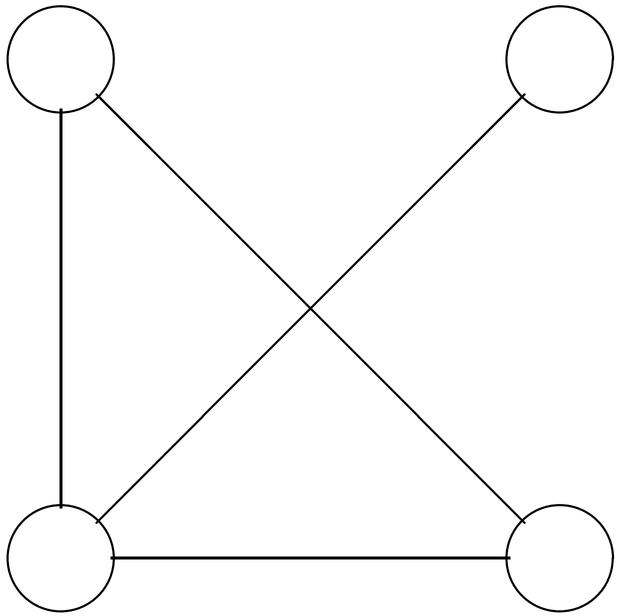
Random graph model with given expected degree sequence

- n nodes with weights $w_1, w_2, \ldots, w_n$.
- For each pair (i, j), create an edge independently with probability $p_{ij} = w_i w_j \rho$, where $\rho = \frac{1}{\sum_{i=1}^n w_i}$.
- The graph H has probability

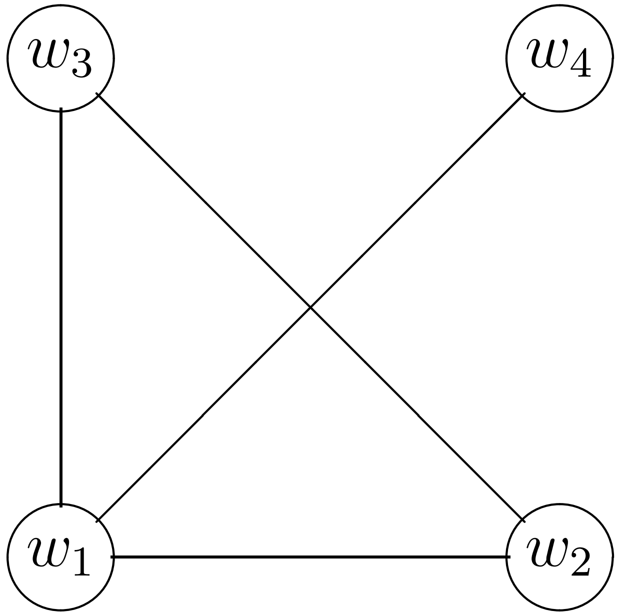
$$\prod_{ij \in E(H)} p_{ij} \prod_{ij \notin E(H)} (1 - p_{ij}).$$

- The expected degree of vertex i is w_i.

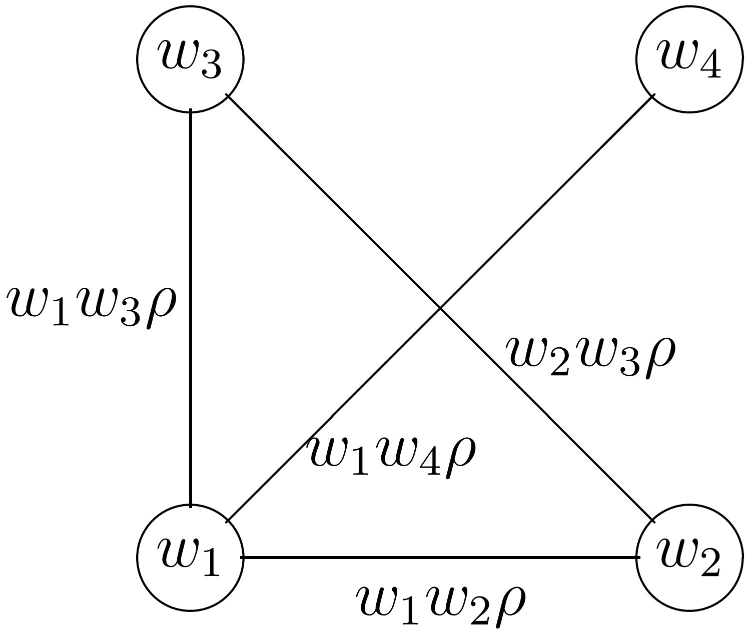
An example: $G(w_1, w_2, w_3, w_4)$



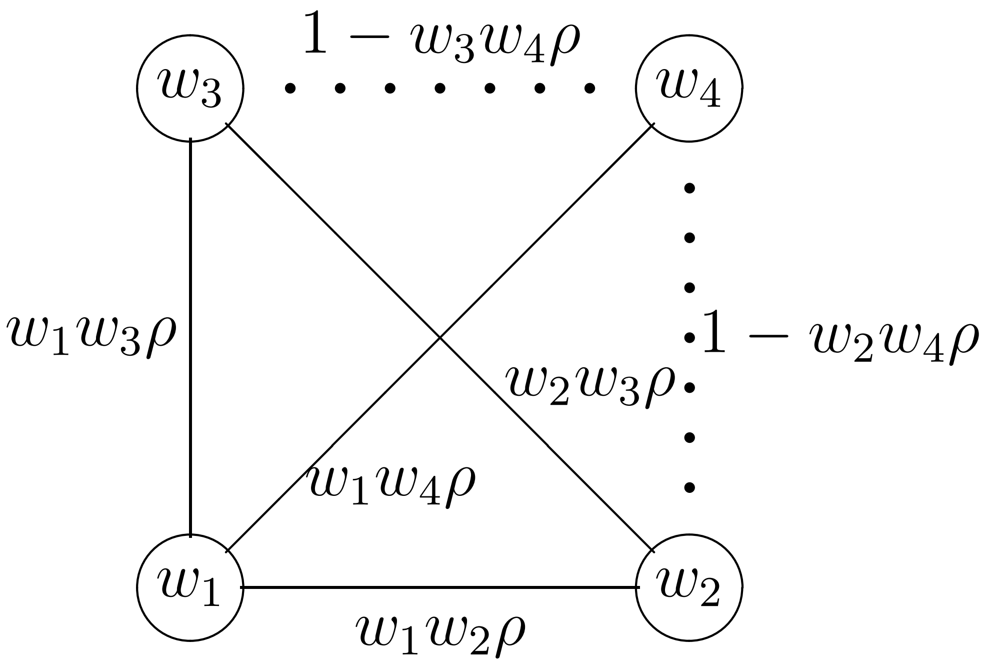
An example: $G(w_1, w_2, w_3, w_4)$



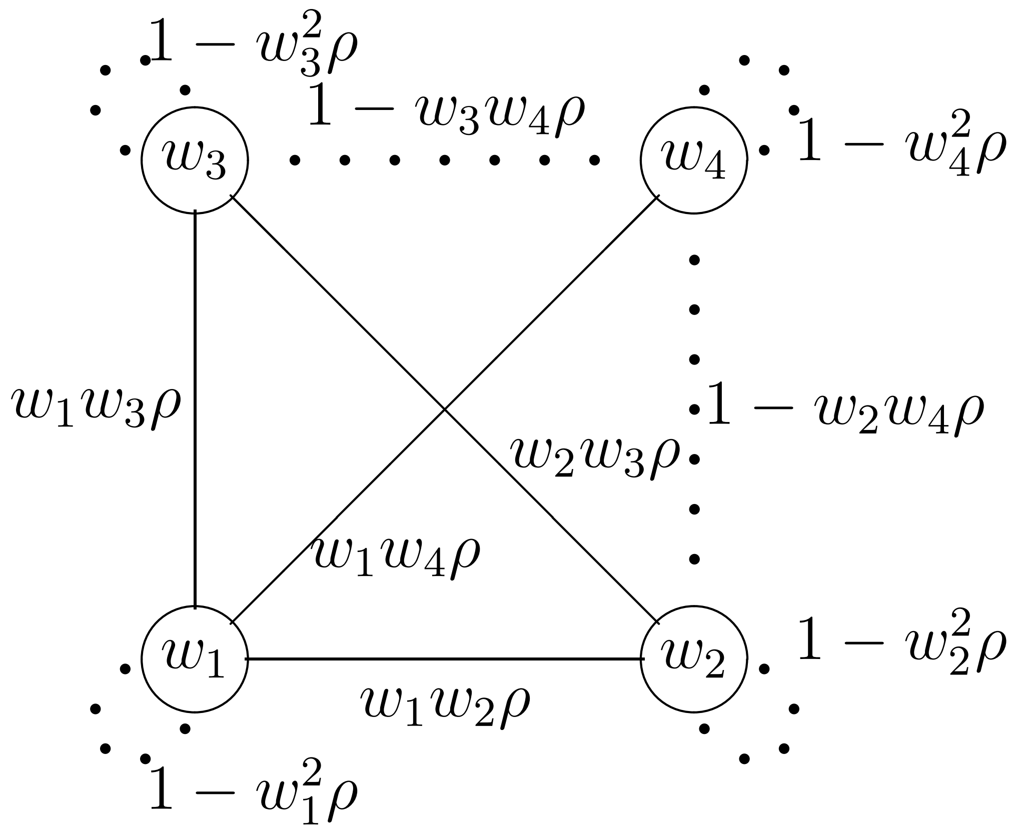
An example: $G(w_1, w_2, w_3, w_4)$



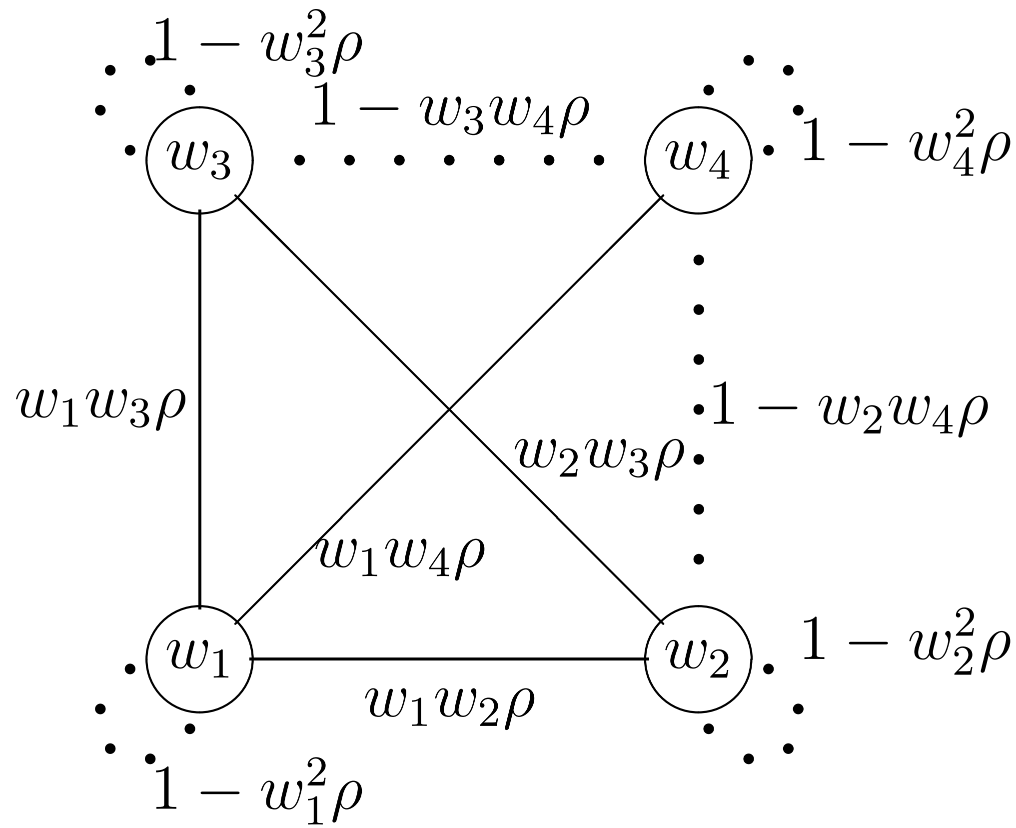
An example: $G(w_1, w_2, w_3, w_4)$



An example: $G(w_1, w_2, w_3, w_4)$



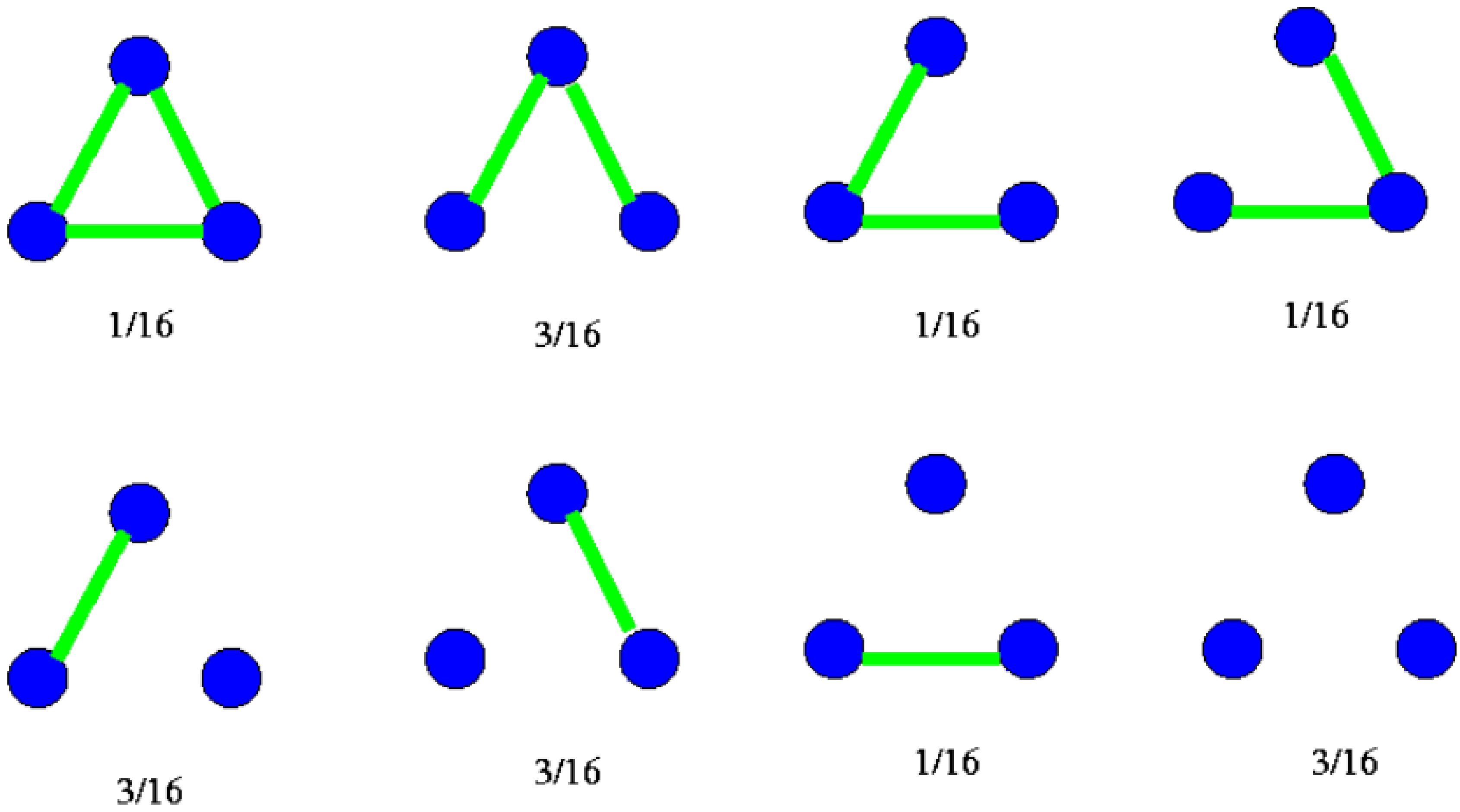
An example: $G(w_1, w_2, w_3, w_4)$



The probability of the graph is

$$w_1^3 w_2^2 w_3^2 w_4 \rho^4 (1 - w_2 w_4 \rho) \times (1 - w_3 w_4 \rho) \prod_{i=1}^{4} (1 - w_i^2 \rho).$$

A example: $G(1,2,1)$



Loops are omitted here.

Notations

For $G = G(w_1, \ldots, w_n)$, let

- $d = \frac{1}{n}\sum_{i=1}^n w_i$
- $\tilde{d} = \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i}$.
- The volume of S: $\mathrm{Vol}(S) = \sum_{i\in S} w_i$.
- The k-th volume of S: $\mathrm{Vol}_k(S) = \sum_{i\in S} w_i^k$.

Notations

For $G = G(w_1, \ldots, w_n)$, let

- $d = \frac{1}{n}\sum_{i=1}^n w_i$
- $\tilde{d} = \frac{\sum_{i=1}^n w_i^2}{\sum_{i=1}^n w_i}$.
- The volume of S: $\mathrm{Vol}(S) = \sum_{i \in S} w_i$.
- The k-th volume of S: $\mathrm{Vol}_k(S) = \sum_{i \in S} w_i^k$.

We have

$$\tilde{d} \geq d$$

"$=$" holds if and only if $w_1 = \cdots = w_n$.

Results

Chung, Lu, 2002 For a random graph G with admissible expected degree sequence $(w_1, \ldots, w_n)$, the average distance is almost surely $(1+o(1))\frac{\log n}{\log \tilde{d}}$.

Results

Chung, Lu, 2002 For a random graph G with admissible expected degree sequence $(w_1, \ldots, w_n)$, the average distance is almost surely $(1 + o(1))\frac{\log n}{\log \tilde{d}}$.

For a random graph G with strongly admissible expected degree sequence $(w_1, \ldots, w_n)$, the diameter is almost surely $\Theta(\frac{\log n}{\log \tilde{d}})$.

Results

Chung, Lu, 2002 For a random graph G with admissible expected degree sequence $(w_1, \ldots, w_n)$, the average distance is almost surely $(1+o(1))\frac{\log n}{\log \tilde{d}}$.

For a random graph G with strongly admissible expected degree sequence $(w_1, \ldots, w_n)$, the diameter is almost surely $\Theta(\frac{\log n}{\log \tilde{d}})$.

For $G(n,p)$, $\tilde{d} = d = np$. These results are consistent to results for $G(n,p)$.

Admissible condition

(i) $\log \tilde{d} \ll \log n$.

(ii) $d > 1 + \epsilon$. $w_i > \epsilon$ for all but $o(n)$ vertices.

(iii) $\exists$ a subset U:

$$\mathrm{vol}_2(U) = (1 + o(1))\mathrm{vol}_2(G) \gg \mathrm{vol}_3(U) \frac{\log \tilde{d} \log\log n}{\tilde{d} \log n}.$$

Admissible condition

(i) $\log \tilde{d} \ll \log n$.

(ii) $d > 1 + \epsilon$. $w_i > \epsilon$ for all but $o(n)$ vertices.

(iii) $\exists$ a subset U:

$$\mathrm{vol}_2(U) = (1 + o(1))\mathrm{vol}_2(G) \gg \mathrm{vol}_3(U)\frac{\log \tilde{d}\, \mathrm{loglog}\, n}{\tilde{d} \log n}.$$

Roughly speaking, G is close to $G(n, p)$. No dense subgraphs.

Admissible condition

(i) $\log \tilde{d} \ll \log n$.
(ii) $d > 1 + \epsilon$. $w_i > \epsilon$ for all but $o(n)$ vertices.
(iii) $\exists$ a subset U:

$$\mathrm{vol}_2(U) = (1 + o(1))\mathrm{vol}_2(G) \gg \mathrm{vol}_3(U)\frac{\log \tilde{d} \log\log n}{\tilde{d} \log n}.$$

Roughly speaking, G is close to $G(n, p)$. No dense subgraphs.
Example: Power law graphs with $\beta > 3$ and $G(n, p)$.

Strongly admissible condition

(i') $\log \tilde{d} = O(\log d)$.

(ii) $d > 1 + \epsilon$. $w_i > \epsilon$ for all but $o(n)$ vertices.

(iii') $\exists$ a subset U: $\mathrm{Vol}_3(U) = O(\mathrm{Vol}_2(G))\frac{\tilde{d}}{\log \tilde{d}}$, and $\mathrm{Vol}_2(U) > d\mathrm{Vol}_2(G)/\tilde{d}$.

Example: Power law graphs with $\beta > 3$ and $G(n, p)$.

Lower bound

- Random graph $G(w_1, \ldots, w_n)$
- u, v: two vertices

With probability at least $1 - \frac{w_u w_v}{\tilde{d}(\tilde{d}-1)} e^{-c}$,

$$d(u,v) \geq \lfloor \frac{\log \mathrm{vol}(G) - c}{\log \tilde{d}} \rfloor.$$

Lower bound

- Random graph $G(w_1, \ldots, w_n)$
- u, v: two vertices

With probability at least $1 - \frac{w_u w_v}{\tilde{d}(\tilde{d}-1)} e^{-c}$,

$$d(u,v) \geq \lfloor \frac{\log \mathrm{vol}(G) - c}{\log \tilde{d}} \rfloor.$$

It implies the average distance is at least

$$(1 - o(1)) \frac{\log n}{\log \tilde{d}}.$$

Proof of lower bound

- P_j: the set of all possible pathes from u to v with length j in K_n.
- For any $\pi = uv_{i_1} \ldots v_{i_{j-1}} v \in P_j$, the probability that π is not a path of G is exactly

$$1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j.$$

- For any $\pi \in P_j$, "π is not a path of G" is a monotone decreasing graph property. FKG inequality applies. (You can treat them as independent events).

Proof of lower bound

$$Pr(d(u,v) \geq k) \;\geq\; \prod_{j=1}^{k-1} \prod_{i_1 \ldots i_{j-1}} (1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j)$$

Proof of lower bound

$$Pr(d(u,v) \geq k) \geq \prod_{j=1}^{k-1} \prod_{i_1 \ldots i_{j-1}} (1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j)$$

$$\approx \prod_{j=1}^{k-1} e^{-w_u w_v \rho^j \sum_{w_1,\ldots,w_{j-1}} w_1^2 \cdots w_{j-1}^2}$$

Proof of lower bound

$$
\begin{aligned}
Pr(d(u,v) \geq k) &\geq \prod_{j=1}^{k-1} \prod_{i_1 \ldots i_{j-1}} (1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j) \\
&\approx \prod_{j=1}^{k-1} e^{-w_u w_v \rho^j \sum_{w_1,\ldots,w_{j-1}} w_1^2 \cdots w_{j-1}^2} \\
&\approx e^{-w_u w_v \sum_{j=1}^{k-1} \rho^j (\sum_{i=1}^n w_i^2)^{j-1}}
\end{aligned}
$$

Proof of lower bound

$$
\begin{aligned}
Pr(d(u,v) \geq k) &\geq \prod_{j=1}^{k-1} \prod_{i_1 \ldots i_{j-1}} (1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j) \\
&\approx \prod_{j=1}^{k-1} e^{-w_u w_v \rho^j \sum_{w_1,\ldots,w_{j-1}} w_1^2 \cdots w_{j-1}^2} \\
&\approx e^{-w_u w_v \sum_{j=1}^{k-1} \rho^j (\sum_{i=1}^n w_i^2)^{j-1}} \\
&\approx e^{-w_u w_v \rho((\sum_i w_i^2 \rho)^k - 1)/(\sum_i w_i^2 \rho - 1)}
\end{aligned}
$$

Proof of lower bound

$$
\begin{aligned}
Pr(d(u,v) \ge k) &\ge \prod_{j=1}^{k-1} \prod_{i_1 \ldots i_{j-1}} (1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j) \\
&\approx \prod_{j=1}^{k-1} e^{-w_u w_v \rho^j \sum_{w_1,\ldots,w_{j-1}} w_1^2 \cdots w_{j-1}^2} \\
&\approx e^{-w_u w_v \sum_{j=1}^{k-1} \rho^j (\sum_{i=1}^n w_i^2)^{j-1}} \\
&\approx e^{-w_u w_v \rho((\sum_i w_i^2 \rho)^k - 1)/(\sum_i w_i^2 \rho - 1)} \\
&\ge 1 - \frac{w_u w_v}{\tilde{d}(\tilde{d}-1)} e^{-c}
\end{aligned}
$$

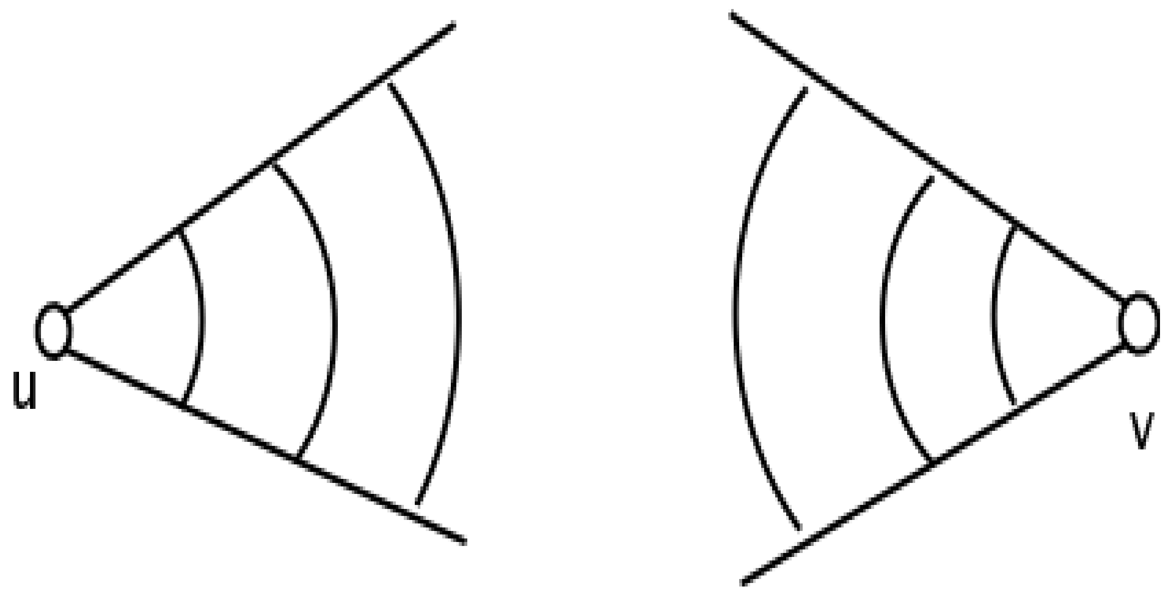
Proof of lower bound

$$
\begin{aligned}
Pr(d(u,v) \geq k) &\geq \prod_{j=1}^{k-1} \prod_{i_1 \ldots i_{j-1}} (1 - w_u w_v w_{i_1}^2 \cdots w_{i_{j-1}}^2 \rho^j) \\
&\approx \prod_{j=1}^{k-1} e^{-w_u w_v \rho^j \sum_{w_1,\ldots,w_{j-1}} w_1^2 \cdots w_{j-1}^2} \\
&\approx e^{-w_u w_v \sum_{j=1}^{k-1} \rho^j (\sum_{i=1}^{n} w_i^2)^{j-1}} \\
&\approx e^{-w_u w_v \rho((\sum_i w_i^2 \rho)^k - 1)/(\sum_i w_i^2 \rho - 1)} \\
&\geq 1 - \frac{w_u w_v}{\tilde{d}(\tilde{d}-1)} e^{-c}
\end{aligned}
$$

Here we choose $k = \lfloor \frac{\log \mathrm{vol}(G) - c}{\log \tilde{d}} \rfloor$. □

Upper bound

To construct a path from u to v, expand u and v's neighborhoods simultaneously.



The neighborhood of S:

$$\Gamma(S) = \{v \ : \ v \sim u \in S \text{ and } v \notin S\}.$$

Neighborhood expansion

Lemma 1: *In a random graph* $G(w_1, \ldots, w_n)$, *for any two subsets* S *and* T *of vertices, we have*

$$\mathrm{vol}(\Gamma(S) \cap T) \geq (1 - 2\epsilon)\mathrm{vol}(S)\frac{\mathrm{vol}_2(T)}{\mathrm{vol}(G)}$$

with probability at least $1 - e^{-c}$, *provided* $\mathrm{vol}(S)$ *satisfies*

$$\frac{2c\mathrm{vol}_3(T)\mathrm{vol}(G)}{\epsilon^2\mathrm{vol}_2^2(T)} \leq \mathrm{vol}(S) \leq \frac{\epsilon\mathrm{vol}_2(T)\mathrm{vol}(G)}{\mathrm{vol}_3(T)}$$

Early neighborhood expansion

Lemma 2: *Suppose that G is admissible. For any fixed vertex v in the giant component, if $\tau = o(\sqrt{n})$, then there is an index $i_0 \leq c_0\tau$ so that with probability at least $1 - \frac{c_1\tau^{3/2}}{e^{c_2\tau}}$, we have*

$$\mathrm{vol}(\Gamma_{i_0}(v)) \geq \tau$$

where c_i's are constants depending only on c and d.
Proof will be omitted.

Time to stop neighborhood expansion

Lemma 3: *For any two disjoint subsets* S *and* T *with* $\mathrm{vol}(S)\mathrm{vol}(T) > c\mathrm{vol}(G)$, *we have*

$$Pr(d(S,T) > 1) < e^{-c}$$

where $d(S,T)$ *denotes the distance between* S *and* T.

Time to stop neighborhood expansion

Lemma 3: *For any two disjoint subsets* S *and* T *with* $\mathrm{vol}(S)\mathrm{vol}(T) > c\mathrm{vol}(G)$, *we have*

$$Pr(d(S,T) > 1) < e^{-c}$$

where $d(S,T)$ *denotes the distance between* S *and* T.

Proof:

$$\begin{aligned} Pr(d(S,T) > 1) &= \prod_{v_i \in S, v_j \in T} (1 - w_i w_j \rho) \\ &\leq e^{-\mathrm{vol}(S)\mathrm{vol}(T)\rho} \\ &< e^{-c}. \end{aligned}$$

Sketched proof of the theorem

It is sufficient to construct a path from u to v with target length $(1+o(1))\frac{\log n}{\log \tilde{d}}$.

Sketched proof of the theorem

It is sufficient to construct a path from u to v with target length $(1+o(1))\frac{\log n}{\log \tilde{d}}$.

- By lemma 2, there is a $i_0 \leq C\epsilon \frac{\log n}{\log \tilde{d}}$ satisfying almost surely

$$\mathrm{vol}(\Gamma_{i_0}(v)) \geq \epsilon \frac{\log n}{\log \tilde{d}}.$$

Sketched proof of the theorem

It is sufficient to construct a path from u to v with target length $(1+o(1))\frac{\log n}{\log \tilde{d}}$.

- By lemma 2, there is a $i_0 \leq C\epsilon\frac{\log n}{\log \tilde{d}}$ satisfying almost surely

$$\mathrm{vol}(\Gamma_{i_0}(v)) \geq \epsilon\frac{\log n}{\log \tilde{d}}.$$

- By lemma 1, almost surely $\mathrm{vol}(\Gamma_i(u))$ grows roughly by a factor of $(1-2\epsilon)\tilde{d}$.

Proof continues

- Therefore, almost surely, for some $i = (\frac{1}{2} + o(1))\frac{\log n}{\log \tilde{d}}$,

$$\mathrm{vol}(\Gamma_i(u)) \geq \sqrt{\mathrm{vol}(G)\log n}.$$

Proof continues

- Therefore, almost surely, for some $i = (\frac{1}{2} + o(1))\frac{\log n}{\log \tilde{d}}$,

$$\mathrm{vol}(\Gamma_i(u)) \geq \sqrt{\mathrm{vol}(G)\log n}.$$

- Similarly, with probability $1 - o(1)$, for some $j = (\frac{1}{2} + o(1))\frac{\log n}{\log \tilde{d}}$,

$$\mathrm{vol}(\Gamma_j(v)) \geq \sqrt{\mathrm{vol}(G)\log n}.$$

Proof continues

- Therefore, almost surely, for some $i = (\frac{1}{2} + o(1))\frac{\log n}{\log \tilde{d}}$,

$$\mathrm{vol}(\Gamma_i(u)) \geq \sqrt{\mathrm{vol}(G)\log n}.$$

- Similarly, with probability $1 - o(1)$, for some $j = (\frac{1}{2} + o(1))\frac{\log n}{\log \tilde{d}}$,

$$\mathrm{vol}(\Gamma_j(v)) \geq \sqrt{\mathrm{vol}(G)\log n}.$$

- Almost surely $\Gamma_i(u)$ and $\Gamma_j(v)$ are connected. Thus

$$d(u,v) \leq i + j + 1 = (1 + o(1))\frac{\log n}{\log \tilde{d}}. \qquad \square$$

A large deviation inequality

Lemma 4: *Let* $X_1, \ldots, X_n$ *be independent random variables with*

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

For $X = \sum_{i=1}^n a_i X_i$, *we have* $E(X) = \sum_{i=1}^n a_i p_i$ *and we define* $\nu = \sum_{i=1}^n a_i^2 p_i$. *Then we have*

$$Pr(X < E(X) - \lambda) \;\leq\; e^{-\lambda^2/2\nu}$$

A large deviation inequality

Lemma 4: *Let* $X_1, \ldots, X_n$ *be independent random variables with*

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

For $X = \sum_{i=1}^n a_i X_i$*, we have* $E(X) = \sum_{i=1}^n a_i p_i$ *and we define* $\nu = \sum_{i=1}^n a_i^2 p_i$*. Then we have*

$$Pr(X < E(X) - \lambda) \ \leq \ e^{-\lambda^2/2\nu}$$

With probability $1 - e^{-c}$*,*

$$X > E(X) - \sqrt{2c\nu}.$$

Proof of Lemma 1

X_j: the indicated random variable for $v_j \in T \cap \Gamma(S)$.

$$\begin{aligned} Pr(X_j = 1) &= 1 - \prod_{v_i \in S} (1 - w_i w_j \rho) \\ &\geq \mathrm{vol}(S) w_j \rho - \mathrm{vol}(S)^2 w_j^2 \rho^2. \end{aligned}$$

Proof of Lemma 1

X_j: the indicated random variable for $v_j \in T \cap \Gamma(S)$.

$$\begin{aligned} Pr(X_j = 1) &= 1 - \prod_{v_i \in S} (1 - w_i w_j \rho) \\ &\geq \mathrm{vol}(S) w_j \rho - \mathrm{vol}(S)^2 w_j^2 \rho^2. \end{aligned}$$

Since $\mathrm{vol}(\Gamma(S) \cap T) = \sum_{v_j \in T} w_j X_j$, the expected value of $\mathrm{vol}(\Gamma(S) \cap T)$ is at least $\mathrm{vol}(S)\mathrm{vol}_2(T)\rho - \mathrm{vol}(S)^2 \mathrm{vol}_3(T)\rho^2$.

Proof of Lemma 1

By Lemma 4, with probability at least $1 - e^{-c}$, we have

$$\mathrm{vol}(\Gamma(S) \cap T) = \sum_{v_j \in T} w_j X_j$$

Proof of Lemma 1

By Lemma 4, with probability at least $1 - e^{-c}$, we have

$$\mathrm{vol}(\Gamma(S) \cap T) = \sum_{v_j \in T} w_j X_j$$

$$\geq \; \mathrm{vol}(S)\mathrm{vol}_2(T)\rho - \mathrm{vol}(S)^2\mathrm{vol}_3(T)\rho^2 - \sqrt{2c\mathrm{vol}(S)\mathrm{vol}_3(T)\rho}$$

Proof of Lemma 1

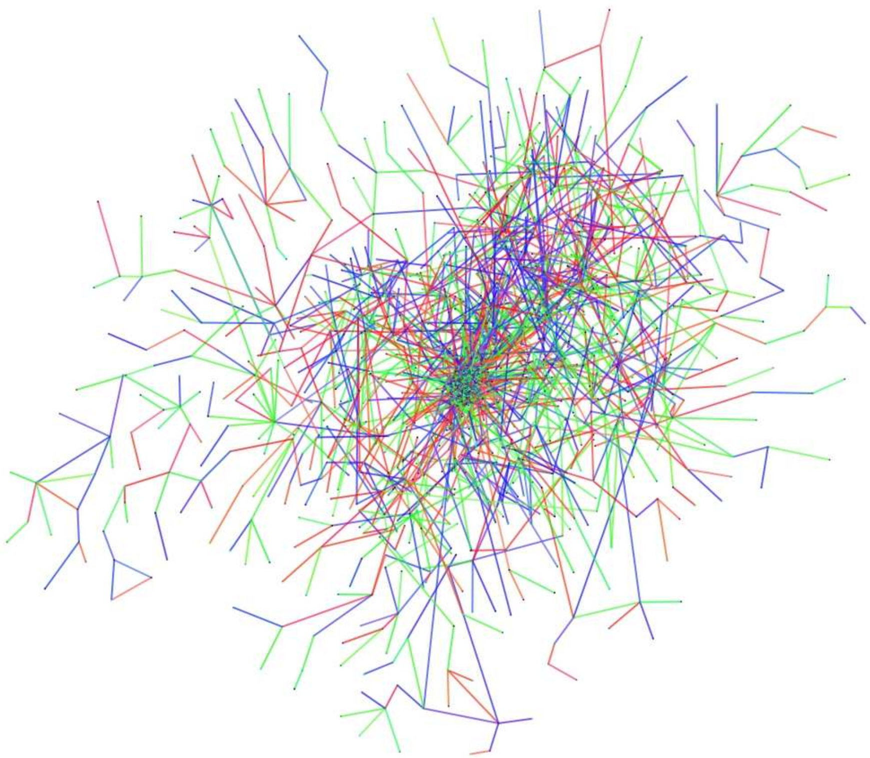
By Lemma 4, with probability at least $1 - e^{-c}$, we have

$$\mathrm{vol}(\Gamma(S) \cap T) = \sum_{v_j \in T} w_j X_j$$

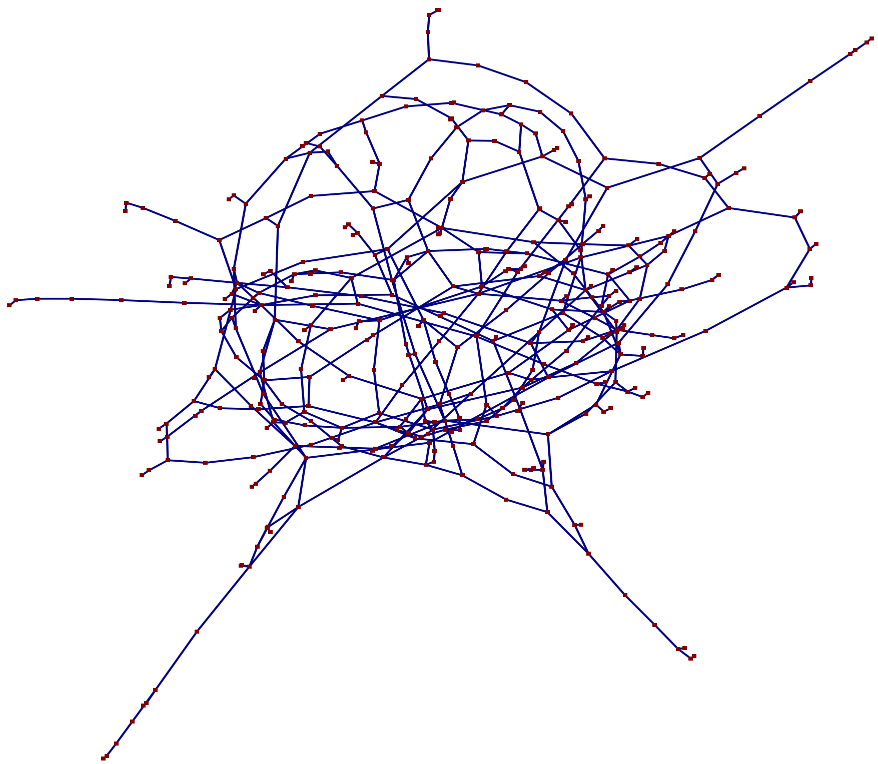
$$\begin{aligned} &\geq \mathrm{vol}(S)\mathrm{vol}_2(T)\rho - \mathrm{vol}(S)^2\mathrm{vol}_3(T)\rho^2 - \sqrt{2c\mathrm{vol}(S)\mathrm{vol}_3(T)\rho} \\ &\geq (1 - 2\epsilon)\mathrm{vol}(S)\mathrm{vol}_2(T)\rho \end{aligned}$$

by the assumption. □

Non-admissible graph versus admissible graph



A random subgraph of the Collaboration Graph.



A Connected component of $G(n, p)$ with $n = 500$ and $p = 0.002$.

A random power law graph

For $\beta > 2$, $d > 1$, and $m >> d$, a random power law graph with the exponent β, the average degree d, and the maximum degree m is defined as $G(w_{i_0}, \ldots, w_{n+i_0-1})$ where

- $c = \frac{\beta-2}{\beta-1} d n^{\frac{1}{\beta-1}}$
- $i_0 = n(\frac{d(\beta-2)}{m(\beta-1)})^{\beta-1}$
- $w_i = c i^{-\frac{1}{\beta-1}}$, for $i_0 \leq i < n + i_0$.

Power law graphs with β in $(2, 3)$

Chung, Lu (2002)

- Examples: the WWW graph, Collaboration graph, etc.

Power law graphs with β in $(2, 3)$

Chung, Lu (2002)

- Examples: the WWW graph, Collaboration graph, etc.
- Non-admissible.

Power law graphs with β in $(2, 3)$

Chung, Lu (2002)

- Examples: the WWW graph, Collaboration graph, etc.
- Non-admissible.
- Containing a dense core, with diameter $\log \log n$.

Power law graphs with β in $(2, 3)$

Chung, Lu (2002)

- Examples: the WWW graph, Collaboration graph, etc.
- Non-admissible.
- Containing a dense core, with diameter $\log \log n$.
- Mostly vertices are within the distance of $O(\log \log n)$ from the core.

Power law graphs with β in $(2, 3)$

Chung, Lu (2002)

- Examples: the WWW graph, Collaboration graph, etc.
- Non-admissible.
- Containing a dense core, with diameter $\log\log n$.
- Mostly vertices are within the distance of $O(\log\log n)$ from the core.
- There are some vertices at the distance of $O(\log n)$.

Power law graphs with β in $(2, 3)$

Chung, Lu (2002)

- Examples: the WWW graph, Collaboration graph, etc.
- Non-admissible.
- Containing a dense core, with diameter $\log \log n$.
- Mostly vertices are within the distance of $O(\log \log n)$ from the core.
- There are some vertices at the distance of $O(\log n)$.

The diameter is $\Theta(\log n)$, while the average distance is $O(\log \log n)$.

The small world phenomenon

Small distance Between any pair of nodes, there is a short path.

Clustering effect Two nodes are more likely to be adjacent if they share a common neighbor.

The small world phenomenon

Small distance Between any pair of nodes, there is a short path.

Clustering effect Two nodes are more likely to be adjacent if they share a common neighbor.

$$\text{A hybrid model} = \text{a local graph} + \text{a random power law graph}$$

Local connectivity

For two fixed integers $k \geq 2$ and $l \geq 2$, a graph L is said to be "locally (k, l)-connected" if for any edge uv, there are at least k edge-disjoint paths with length at most l joining u to v (including the edge uv).

Local connectivity

For two fixed integers $k \geq 2$ and $l \geq 2$, a graph L is said to be "locally (k, l)-connected" if for any edge uv, there are at least k edge-disjoint paths with length at most l joining u to v (including the edge uv).

For example, the grid graph $C_n \Box C_n$ is locally $(3, 3)$-connected as well as locally $(4, 9)$-connected.

Local connectivity

For two fixed integers $k \geq 2$ and $l \geq 2$, a graph L is said to be "locally (k, l)-connected" if for any edge uv, there are at least k edge-disjoint paths with length at most l joining u to v (including the edge uv).

For example, the grid graph $C_n \square C_n$ is locally $(3, 3)$-connected as well as locally $(4, 9)$-connected.

By this definition, the union of two locally (k, l)-connected graphs is locally (k, l)-connected.

Local connectivity

For two fixed integers $k \geq 2$ and $l \geq 2$, a graph L is said to be "locally (k, l)-connected" if for any edge uv, there are at least k edge-disjoint paths with length at most l joining u to v (including the edge uv).

For example, the grid graph $C_n \Box C_n$ is locally $(3, 3)$-connected as well as locally $(4, 9)$-connected.

By this definition, the union of two locally (k, l)-connected graphs is locally (k, l)-connected.

The maximum locally (k, l)-connected subgraph H is the union of all locally (k, l)-connected subgraphs of G.

Algorithm(k, l):

For each edge $e = uv$, check whether there are k edge-disjoint paths with length at most l connecting u and v in the current graph G. If not, delete the edge e from G. Then iterate the procedure until no edge can be removed.

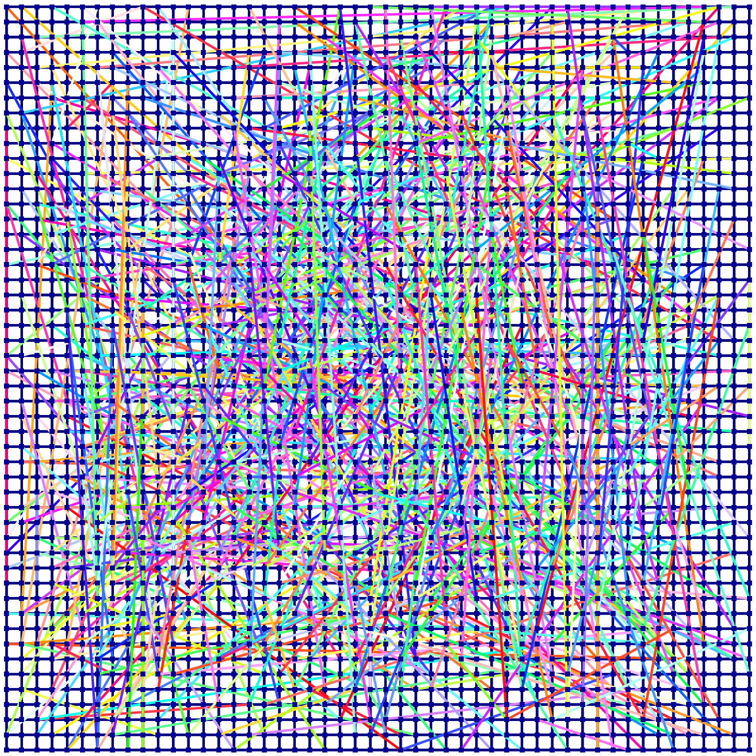
Algorithm(k, l):

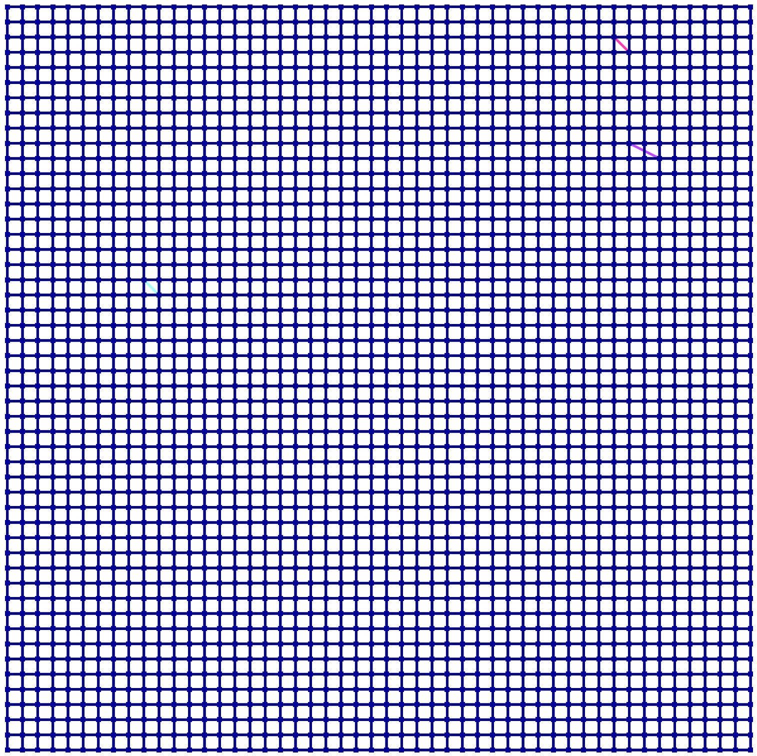
For each edge $e = uv$, check whether there are k edge-disjoint paths with length at most l connecting u and v in the current graph G. If not, delete the edge e from G. Then iterate the procedure until no edge can be removed.

Theorem: *For any graph G,* Algorithm(k, l) *finds the unique maximum locally (k, l)-connected subgraph regardless of the order of edges chosen.*

Recovering the local graph

A hybrid graph, which contains the grid graph $C_{50} \Box C_{50}$ as the local graph, and 528 additional random edges.





The local graph is almost perfect recoverd after applying the algorithm with $k = l = 3$.

Hybrid graph model $H(n, \beta, d, m, L)$

- n: the number of vertices.

Hybrid graph model $H(n, \beta, d, m, L)$

- n: the number of vertices.
- L: a locally (k, l)-connected graph with bounded degrees.

Hybrid graph model $H(n, \beta, d, m, L)$

- n: the number of vertices.
- L: a locally (k, l)-connected graph with bounded degrees.
- β: the target power law exponent.

Hybrid graph model $H(n, \beta, d, m, L)$

- n: the number of vertices.
- L: a locally (k, l)-connected graph with bounded degrees.
- β: the target power law exponent.
- d : the target average degree.

Hybrid graph model $H(n, \beta, d, m, L)$

- n: the number of vertices.
- L: a locally (k, l)-connected graph with bounded degrees.
- β: the target power law exponent.
- d : the target average degree.

The hybrid graph is the union of the local graph L and the random power law graph with parameter n, β, d, and m.

Result 1

Chung Lu *For any fixed constants* M, $k \geq 3$, *and* $l \geq 2$, *suppose* L *is a connected and locally* (k, l)*-connected graph with degrees bounded by* M. *Let* L' *be the maximum locally* (k, l)*-connected subgraph in the hybrid graph* $H(n, \beta, d, m, L)$ *with the maximum degree* m *satisfying* $m = o(n^{\frac{1-1/(2k)}{l+1}})$. *Then the following holds:*

1. $L \subset L'$. *The expected number of edges in* $L' \setminus L$ *is small, i.e.,* $e(L') - e(L) = O(m) = o(n^{\frac{1-1/(2k)}{l+1}})$.

Continue

2. *Almost surely, for all vertices v, the degree of v in L' can increase at most by 1 if $l \geq 3$ (and by 2 if $l = 2$).*

$$d_{L'}(v) \leq \begin{cases} d_L(v) + 2 & \text{if } l = 2; \\ d_L(v) + 1 & \text{if } l \geq 3. \end{cases}$$

Continue

2. *Almost surely, for all vertices v, the degree of v in L' can increase at most by 1 if $l \geq 3$ (and by 2 if $l = 2$).*

$$d_{L'}(v) \leq \begin{cases} d_L(v) + 2 & \text{if } l = 2; \\ d_L(v) + 1 & \text{if } l \geq 3. \end{cases}$$

3. *The diameter $D(L')$ of L' is almost surely $(1 + o(1))D(L)$ if the diameter $D(L)$ is sufficiently large.*

Diameter and average distance

Chung Lu (2004) For a hybrid graph $H(n, \beta, d, m, L)$, almost surely, we have

Case $\beta > 3$, the average distance is $(1 + o(1))\frac{\log n}{\log \tilde{d}}$ and the diameter is $O(\log n)$.

Diameter and average distance

Chung Lu (2004) For a hybrid graph $H(n, \beta, d, m, L)$, almost surely, we have

Case $\beta > 3$, the average distance is $(1 + o(1))\frac{\log n}{\log \tilde{d}}$ and the diameter is $O(\log n)$.

Case $2 < \beta < 3$, the average distance is $O(\log \log n)$ and the diameter is $O(\log n)$.

Diameter and average distance

Chung Lu (2004) For a hybrid graph $H(n, \beta, d, m, L)$, almost surely, we have

Case $\beta > 3$**,** the average distance is $(1 + o(1))\frac{\log n}{\log \tilde{d}}$ and the diameter is $O(\log n)$.

Case $2 < \beta < 3$**,** the average distance is $O(\log \log n)$ and the diameter is $O(\log n)$.

Case $\beta = 3$**,** the average distance is $O(\log n / \log \log n)$ and the diameter is $O(\log n)$.

References

- Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees, *Proc. Natl. Acad. Sci.* **99** (2002), 15879–15882. Full version is in *Internet Mathematics* **1**, No. 1, (2003), 91–114.
- Fan Chung and Linyuan Lu, The small world phenomenon in hybrid power law graphs, *Lect. Notes Phys.* **650** (2004), 89-104.
- Reid Andersen, Fan Chung, and Linyuan Lu, Modeling the small-world phenomenon with local network flow, *Internet Mathematics*, **2** No. 3, (2005),

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