On random graph models

Null models are most informative

Mehdi Naima[1]

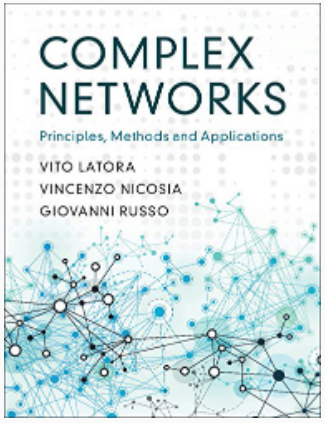
October 18, 2024

Sorbonne Université, LIP6.

Table of contents

References

- Complex networks: principles, methods and applications by Latora, Vito and Nicosia [LNR17].
- Random Graphs by Bollobás [Bol01].
- Introduction to random graphs by Frieze and Karoński [FK16].



Introduction

Graph basics

- We will consider laballed undirected graphs.
- A graph is a pair $G = (V, E)$ where $V = \{1, 2, \ldots, n\}$ node set and $E = \{\{i, j\} \mid i, j \in V, i \neq j\}$ edge set.
- The graphs we consider are called simple graphs, they contain no loops or multiple edges.
- We denote $n = |V|$ and $m = |E|$.

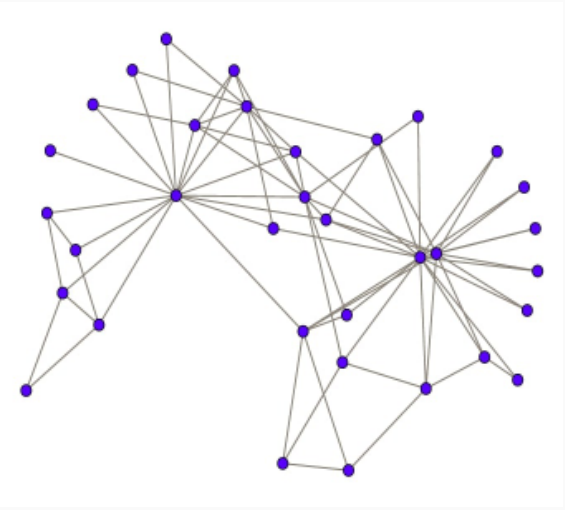


Figure 1: 34 members of a karate club, links between people who interacted outside the club.

Graph characteristics

- Degree distribution
- Connected components
- Characteristic path length
- Clustering coefficient
- Automorphism group

Degree Distribution

Let d_i be the degree of node i.

Degree Histogram: For $k \geq 0$:

$$n_k = |\{i \mid d_i = k\}|$$

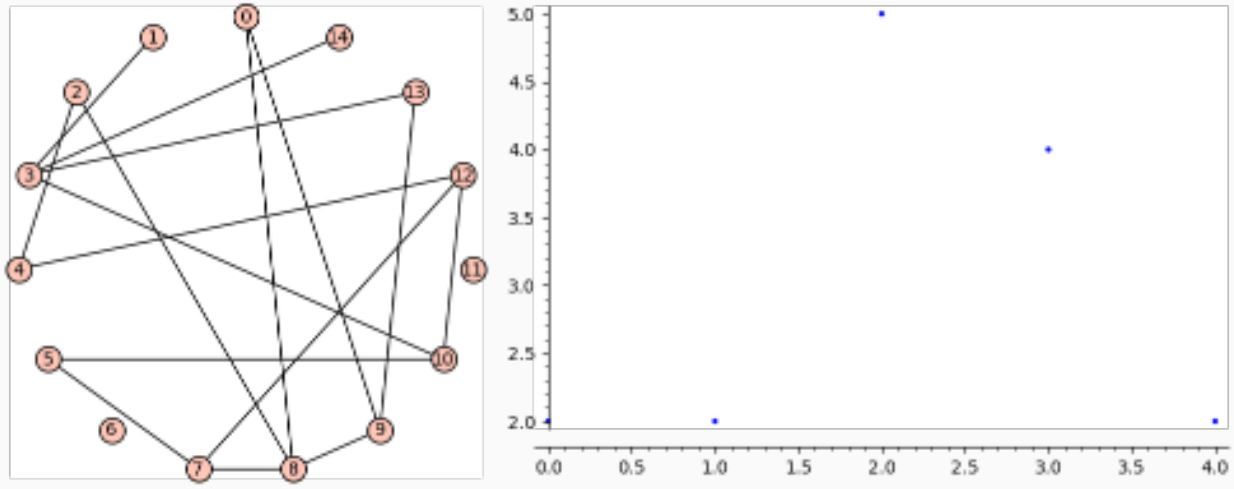


Figure 2: Random graph on 15 nodes and its degree histogram

Degree Distribution: For $k \geq 0$:

$$p_k = \frac{n_k}{n}$$

Connected Component

A Path: from u to v is a sequence of edges $(u_1, v_1), \ldots, (u_i, v_i)$ with $u_1 = u$ and $v_i = v$ and $u_i = v_{i-1}$.

Connected component:
Is a subset of nodes S such that $\forall u, w \in S$, $\exists$ a path between them. Moreover, S can not be extended (S is maximal).

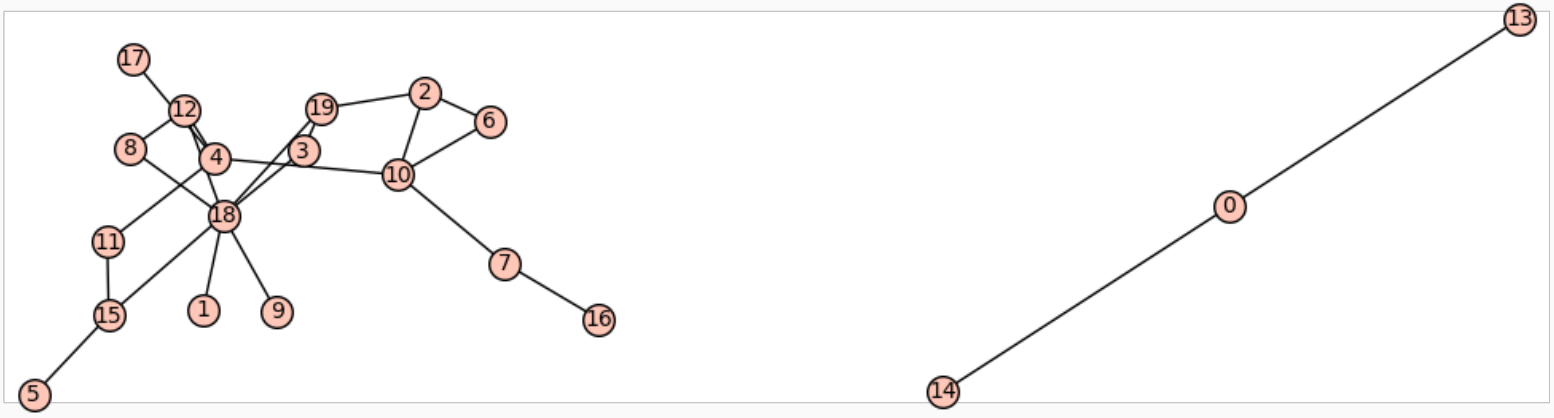


Figure 3: The graph has 2 Connected components

Characteristic path length

Distance: For $u, v \in V$, the distance between u and v is length of the shortest path connecting them. Let $\mathcal{P}(u, v)$ set of paths between u and v we have:

$$d(u, v) = \min_{p \in \mathcal{P}(u,v)} (|p|)$$

Distance distribution:

$$dd_k = \frac{|\{(u, v) \subseteq V \mid d(u, v) = k\}|}{\binom{n}{2}}$$

Characteristic path length:
Is the average distance between nodes: $ch = \mathbb{E}(dd)$
$ch = (10 \times 1 + 12 \times 2 + 11 \times 3 + 8 \times 4 + 4 \times 5)/45) = 2.64$

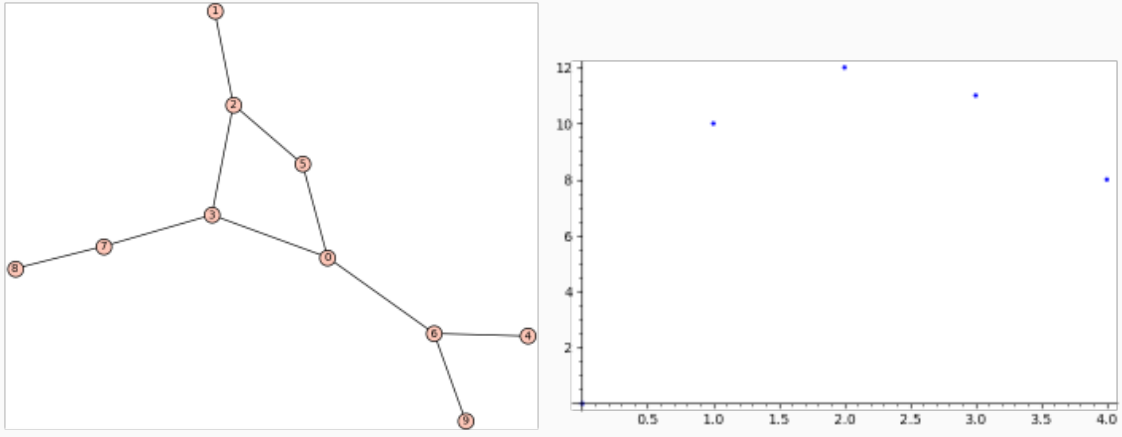


Figure 4: Distance histogram $0, 10, 12, 11, 8, 4, 0, 0, 0, 0$

Clustering coefficient

Global clustering coefficient

$$C = \frac{3\,|\text{triangles}|}{|\text{pairs of distinct neighbors}|}$$

For each v its number of pairs of distinct neighbors is $\binom{d}{2}$. Thus:

$$C = \frac{3\,|\text{triangles}|}{\sum_{v \in V} \binom{d_v}{2}}$$

Local clustering coefficient

$$C_i = \frac{|\text{triangles of node i}|}{|\text{pairs of distinct neighbors of i}|} = \frac{|\text{triangles of node i}|}{\binom{d_i}{2}}$$

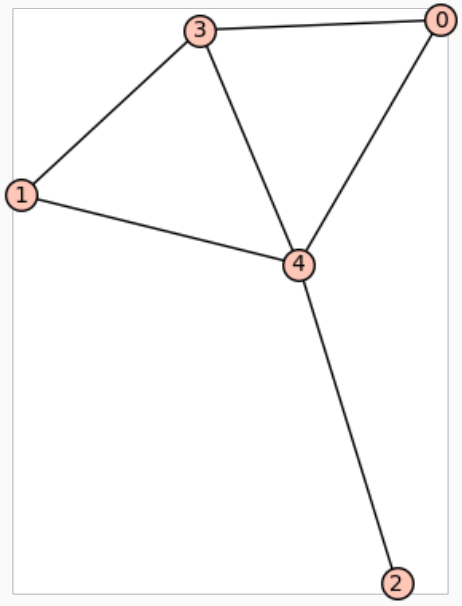


Figure 5: Local clustering $\{0:1, 1:1, 2:0, 3:2/3, 4:1/3\}$. Global clustering $6/11$.

Real-world graphs tend to have a high clustering coefficient

Automorphism Group

An Automorphism of a graph:
An Automorphism f of a graph $G = (V, E)$ is a permutation of its vertices such that:

$$\forall (u, v) \in V^2, ((u, v) \in E \Leftrightarrow (f(u), f(v)) \in E)$$

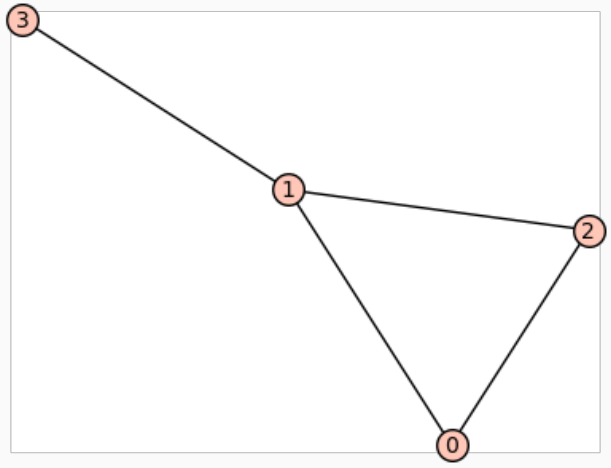


Figure 6: The automorphisms of G are the identity $f = id$ and $f : \{(0, 2), (1, 1), (2, 0), (3, 3)\}$

Importance of generating random grpahs

- Model real world phenomena (spread dicease, social interactions, . . .).
- Better understanding of real data (detect when a process has inherent randomness).
- Generating datasets to test algorithms.

Erdös-Rényi random graphs

Paul Erdös and Alfred Rényi

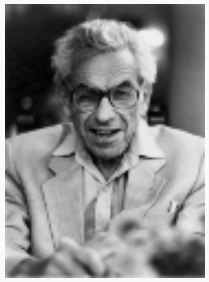


Figure 7: Paul Erdös (1913 - 1996)

- Hungarian mathematician.
- One of the most prolific mathematicians of the 20th century.
- Published around 1,500 mathematical papers during his lifetime.
- Contributions mainly concerns discrete mathematics, graph theory, probability theory and number theory.

Figure 8: Alfred Rényi (1921 - 1970)

- Hungarian mathematician.
- made contributions in combinatorics, graph theory and number theory.

History of random graphs

- Introduced by Erdös to give proba. construction of a graph with large chromatic number in 1959 [Erd59b].
- Later Erdös and Rényi began a systematic study of random graph in a series of papers [Erd59a, ER$^+$60, ER61, ER68].
- They introduced random graphs with fixed number of edges $\mathbb{G}_{n,m}$.
- Gilbert introduced random graphs where the number of edges is variable $\mathbb{G}_{n,p}$ in [Gil59].
- Béla Bollobás published a book on random graphs in 1985.

ON THE EVOLUTION OF RANDOM GRAPHS

by

P. ERDŐS and A. RÉNYI

Dedicated to Professor P. Turán on his 50th birthday.

Introduction

Our aim is to study the probable structure of a random graph $\Gamma_{n,N}$ which has n given labelled vertices $P_1, P_2, \ldots, P_n$ and N edges; we suppose that these N edges are chosen at random among the $\binom{n}{2}$ possible edges, so that all $\binom{\binom{n}{2}}{N} = C_{n,N}$ possible choices are supposed to be equiprobable. Thus if $G_{n,N}$ denotes any one of the $C_{n,N}$ graphs formed from n given labelled points and having N edges, the probability that the random graph $\Gamma_{n,N}$ is identical with $G_{n,N}$ is $\frac{1}{C_{n,N}}$. If A is a property which a graph may or may not possess, we denote by $\mathbf{P}_{n,N}(A)$ the probability that the random graph $\Gamma_{n,N}$ possesses the property A, i. e. we put $\mathbf{P}_{n,N}(A) = \frac{A_{n,N}}{C_{n,N}}$ where $A_{n,N}$ denotes the number of those $G_{n,N}$ which have the property A.

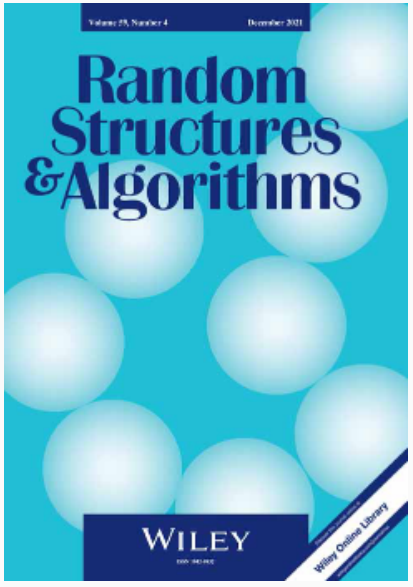
An other equivalent formulation is the following: Let us suppose that n labelled vertices $P_1, P_2, \ldots, P_n$ are given. Let us choose at random an edge among the $\binom{n}{2}$ possible edges, so that all these edges are equiprobable. After this let us choose an other edge among the remaining $\binom{n}{2} - 1$ edges, and continue this process so that if already k edges are fixed, any of the remaining $\binom{n}{2} - k$ edges have equal probabilities to be chosen as the next one. We shall study the "evolution" of such a random graph if N is increased. In this investigation we endeavour to find what is the "typical" structure at a given stage of evolution (i. e. if N is equal, or asymptotically equal, to a given function $N(n)$ of n). By a "typical" structure we mean such a structure the probability of which tends to 1 if $n \to +\infty$ when $N = N(n)$. If A is such a property that $\lim_{n=\infty} \mathbf{P}_{n,N(n)}(A) = 1$, we shall say that „almost all" graphs $G_{n,N(n)}$ possess this property.

17

2 A Matematikai Kutató Intézet Közleményei 7. A/1–2.

History of random graphs

- Journal of Random Structures and Algorithms in 1990.
- Followed by Combinatorics, Probability and computing.
- Other contributors include Janson, Luczak and many others.



Model A (Uniform Random Graph)

Let $\mathcal{G}_{n,m}$ be the class of labelled graphs on vertex set $V = \{1, 2, \ldots, n\}$ having m edges:

$$|\mathcal{G}_{n,m}| = \binom{\binom{n}{2}}{m}$$

Since $\sum_{m=0}^{n} \binom{m}{n} = 2^n$, the total number of graphs $\mathcal{G}_n$ on n vertices is

$$\begin{aligned} |\mathcal{G}_n| &= \sum_{m=0}^{n} |\mathcal{G}_{n,m}| \\ &= \sum_{m=0}^{n} \binom{\binom{n}{2}}{m} \\ &= 2^{\binom{n}{2}} \end{aligned}$$

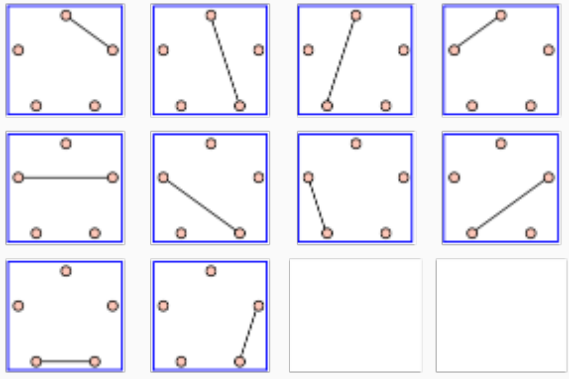


Figure 9: Ensemble $\mathcal{G}_{5,1}$

Model A (Uniform Random Graph)

Give to each graph $G \in \mathcal{G}_{n,m}$ the same probability:

$$\mathbb{P}(G) = \binom{\binom{n}{2}}{m}^{-1}$$

Denote such a random graph with $\mathbb{G}_{n,m}$.
Remember the number of possible graphs is large, for $n = 15$ and $m = 15$

$$|\mathcal{G}_{15,15}| = \binom{\binom{15}{2}}{15} = \binom{105}{15} \approx 5.5 \times 10^8$$

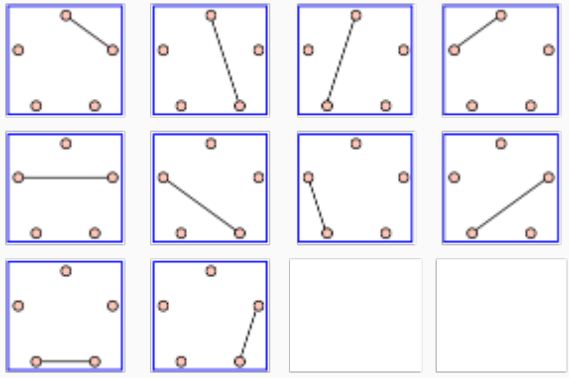


Figure 10: Ensemble $\mathcal{G}_{5,1}$

Model B (Binomial Random Graph)

- Let $0 \leq p \leq 1$
- The random graph $\mathbb{G}_{n,p}$ has n nodes and is obtained by connecting each pair of nodes with a probability p
- Each graph G having n nodes and k edges is generated by $\mathbb{G}_{n,p}$ with probability

$$\mathbb{P}(G) = p^k(1-p)^{\binom{n}{2}-k}$$

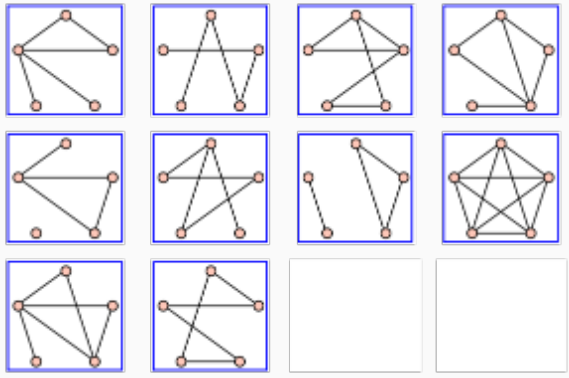


Figure 11: Few graphs of $\mathbb{G}_{5,0.5}$

Model B Number of edges

The number of edges is variable.

Probability of $\mathbb{G}_{n,p}$ having exactly K edges

$$P(k) = \binom{M}{k} p^k (1-p)^{M-k}$$

where $M = \binom{n}{2}$
$P(k)$ corresponds to $Bin(M, p, k)$ a binomial distribution.

Average nb. of edges in a graph of $\mathbb{G}_{n,p}$

$$\langle m \rangle = \binom{n}{2} p$$

Since $\sum_{v \in V} d_v = 2|E|$.

Average node degree in a graph of $\mathbb{G}_{n,m}$:

$$\langle k \rangle = \frac{\sum_{v \in V} d_v}{|V|} = \frac{2m}{n}$$

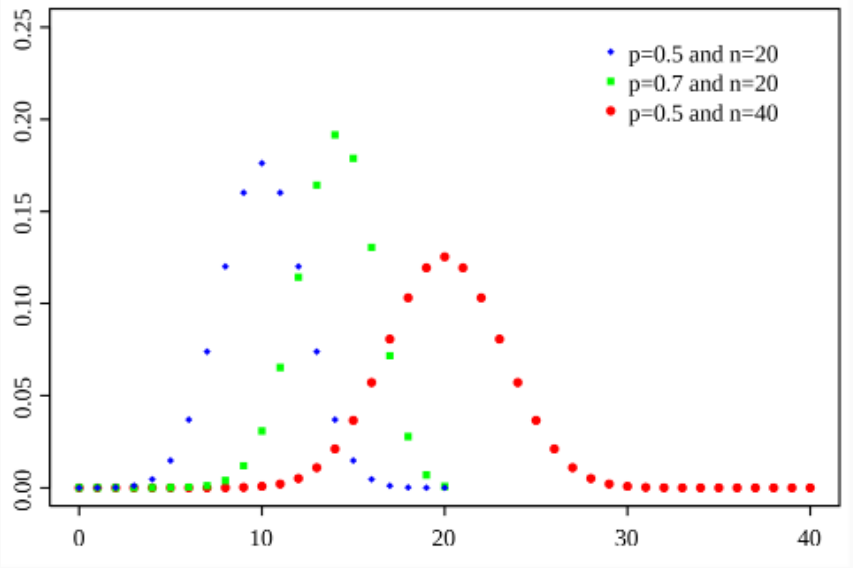


Figure 12: Binomoial Distribution from Wikipedia

Model B Degree distribution

Probability of node i in $\mathbb{G}_{n,p}$ having degree k: for $0 \leq k \leq n-1$

$$P(d_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

- Degree distribution $p_k = \frac{n_k}{n}$
- Extending this definition to an ensemble of graphs gives: $p_k = \frac{\bar{n_k}}{n}$, $\bar{n_k}$ being the mean of the number of nodes of degree k.
- $\bar{n_k} = \sum_{i=1}^{n} P(d_i = k) = nP(d_i = k)$
- Therefore, d_k (degree distribution) has the same distribution as $P(d_i = k)$.

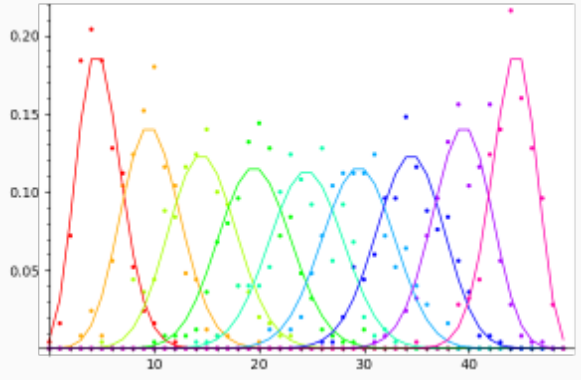


Figure 13: Degree distribution on graphs of $\mathbb{G}_{n,p}$ with different values of p. The points correspond to the real degree distribution and the lines to the theoretical binomial one.

Model B (Binomial Random Graph)

- Interested to analyse $n \to \infty$ with $\mathbf{p = p(n) \approx c\, n^{-1}}$, the average node degree gives :

$$\langle k \rangle = \frac{\binom{n}{2} p}{n} = \frac{c(n-1)}{2n} \approx c/2$$

- The variance in binomial distribution is $Mp(1-p)$ keeping constant the average node degree

$$\mathbb{V}[m] = \binom{n}{2} \frac{c}{n}(1 - \frac{c}{n}) \underset{n\to\infty}{=} \frac{c}{2} n, \quad \sigma_m = \sqrt{\mathbb{V}[m]}$$

- Average number of edges:

$$\langle m \rangle = \binom{n}{2} p \underset{n\to\infty}{=} \frac{c}{2} n \quad \Longrightarrow \quad \frac{\sigma_m}{\bar{m}} \underset{n\to\infty}{=} 0$$

$\Longrightarrow$ In large graphs, the fluctuations in the value of k of model B can be neglected

Comparison between the 2 models

- In model A the number of edges is fixed, it is variable in model B.
- The probability of an edge between nodes in model A is m/M where m is the number of edges and $M = \binom{n}{2}$.
- To get a similar graph in model B we can take $p = m/M$.
- Same can be done for $\mathbb{G}_{n,m}$ by taking $m = \binom{n}{2} p$.

Correlations in the model A

In model B the probability of edge $\{i, j\}$ does not depend on existence of an edge $\{r, s\}$. In model A, let $g \in \mathbb{G}_{n,m}$, with $g = (V, E)$, then

$$\mathbb{P}(\{i, j\} \in E) = \frac{m}{M}$$

However if the edge $\{r, s\}$ exists, this becomes

$$\mathbb{P}(\{i, j\} \in E \,|\, \{r, s\} \in E) = \frac{m-1}{M-1}$$

Comparison between the 2 models

The average node degree in a graph of $\mathbb{G}_{n,m}$ is $\langle m \rangle = \frac{2m}{n}$, by keeping $\langle m \rangle$ constant and letting $n \to \infty$, we get

Correlations can be neglected in large graphs

$$\frac{m/M}{(m-1)/(M-1)} \underset{n\to\infty}{=} 1$$

We have $\frac{m/M}{(m-1)/(M-1)} = \frac{m}{M}\frac{M-1}{m-1} = \frac{m(M-1)}{M(m-1)} = \frac{m}{m-1} - \frac{m}{M(m-1)}$. Now $\frac{m}{m-1} = 1 + \frac{1}{k-1} = 1 + \frac{1}{(\langle m \rangle n/2)-1} = 1 + \frac{2}{\langle m \rangle n - 2} \underset{n\to\infty}{=} 1$ and immediately $\frac{m}{M(m-1)} \underset{n\to\infty}{=} 0$

Comparison between the 2 models

Definition

Let graph $G = (V, E)$ and $e \notin E$, we call a graph property $\mathcal{P}$ **monotone increasing** if $G \in \mathcal{P}$ implies that $G' = (V, E \cup \{e\})$ is also in $\mathcal{P}$ that is $G' \in \mathcal{P}$.

For example, connectivity and Hamiltonicity, containing a subgraph are monotone increasing properties

Theorem from [Bol01, Chap I]

Let $\mathcal{P}$ be a monotone increasing graph property and $p = m/n$ where $m = m(n)$. Then, for large n and $p = o(1)$ such that $np, n(1-p)(np)^{1/2} \to \infty$,

$$\mathbb{P}(\mathbb{G}_{n,m} \in \mathcal{P}) \leq 3\mathbb{P}(\mathbb{G}_{n,p} \in \mathcal{P})$$

Phase transitions

Phase transition

- Phase transitions are common in many branches of physics.
- For example boiling (transition from the liquid to the gaseous phase) of water at a transition temperature of 100° C

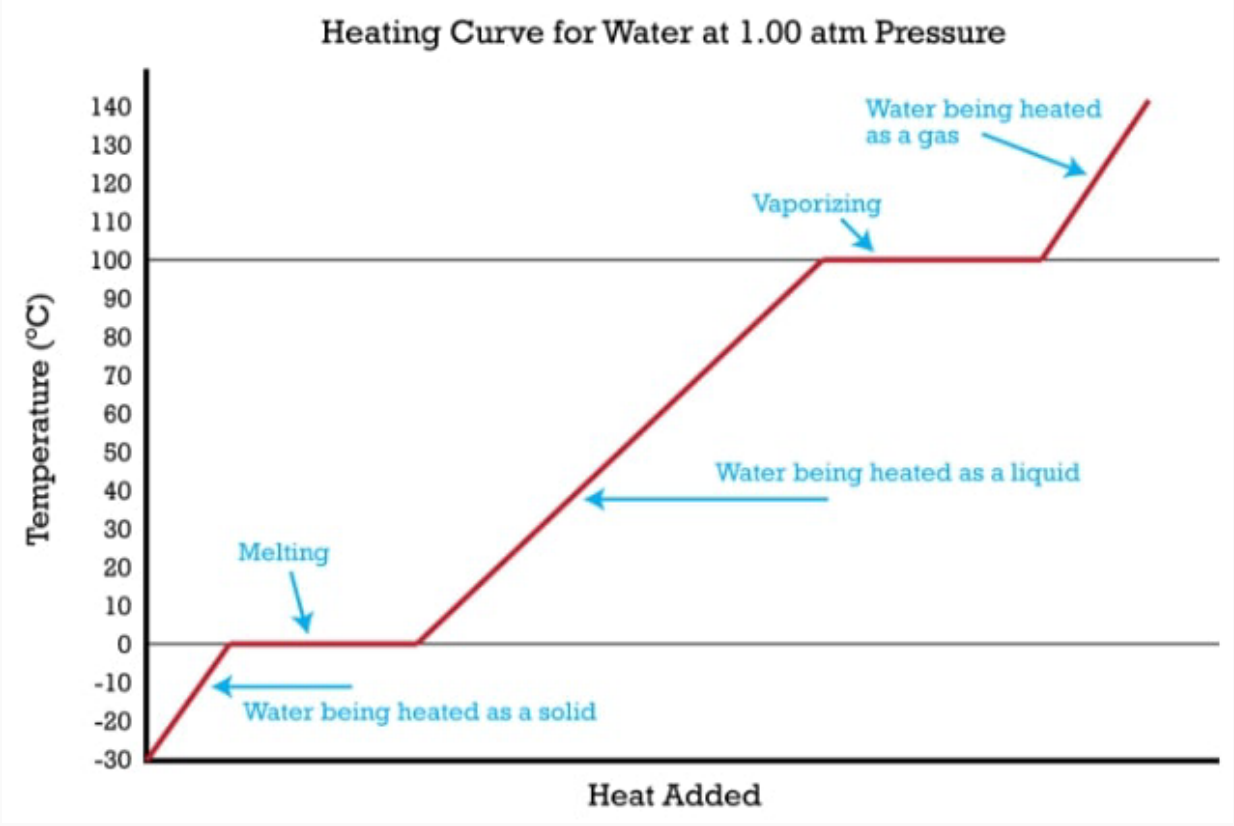


Figure 14: Water phase transition

Phase transition

Definition

Almost every graph (a.e.g.) has the property $\mathcal{Q}$ means that the probability that a graph in the ensemble has the property $\mathcal{Q}$ tends to 1 as $n \to \infty$.

Theorem from [LNR17, Chap 3]

Let $k \geq 2$, $k-1 \leq l \leq \binom{k}{2}$ and let $F \equiv F_{k,l}$ be a connected graph on k nodes and l edges.

If $p(n)/n^{-k/l} \to 0$ then a.e.g in $\mathbb{G}_{n,p(n)}$ does not contain F, while if $p(n)/n^{-k/l} \to \infty$ then a.e.g in $\mathbb{G}_{n,p(n)}$ contains F.

Average number of a subgraph $F_{k,l}$

$$\bar{n}_F = \mathbb{E}[n_F] = \binom{n}{k} \frac{n!}{a_F} p^l \approx \frac{n^k p^l}{a_F},$$

where a_F is the size of the automorphism group of F.

Phase transition for cycles, trees, and complete graphs

- A tree of order n has $l = n - 1$ edges $\longrightarrow p_c(n) = c\, n^{-n/(n-1)}$.
- A cycle of order n has $l = n$ edges $\longrightarrow p_c(n) = c\, n^{-1}$.

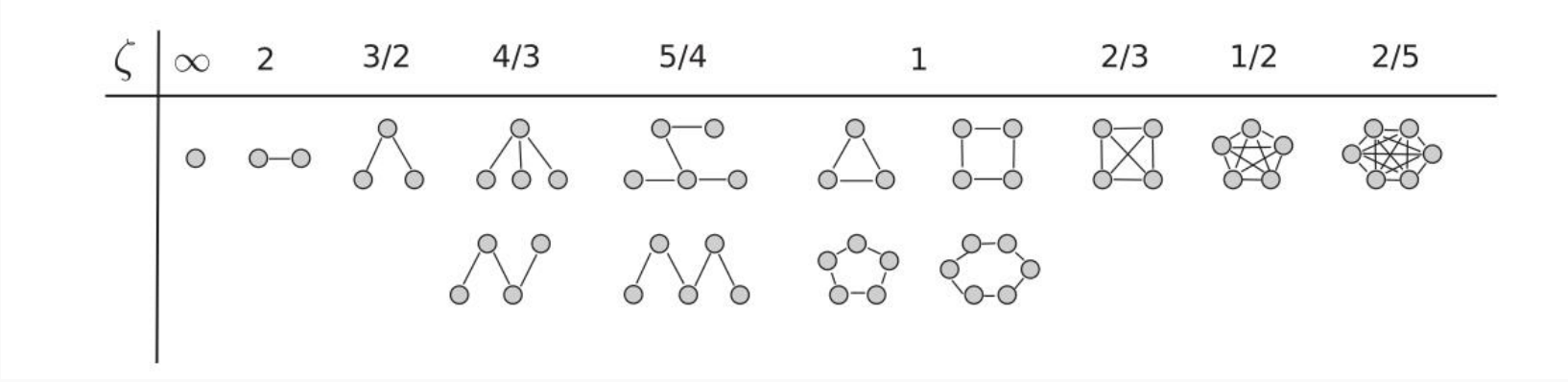


Figure 15: Threshold probabilities $p(n) \sim n^{-\zeta}$ for the appearance of subgraphs.

Giant component

- In simple words, the probability $1/n$ is a threshold value for the size of the largest component.
- It is possible to get threshold probabilities for other properties, for instance the threshold on connectedness is $p_c(n) = \ln n/n$.
- The same phenomena happens in directed graphs or hypergraphs, however the threshold happens at different values.

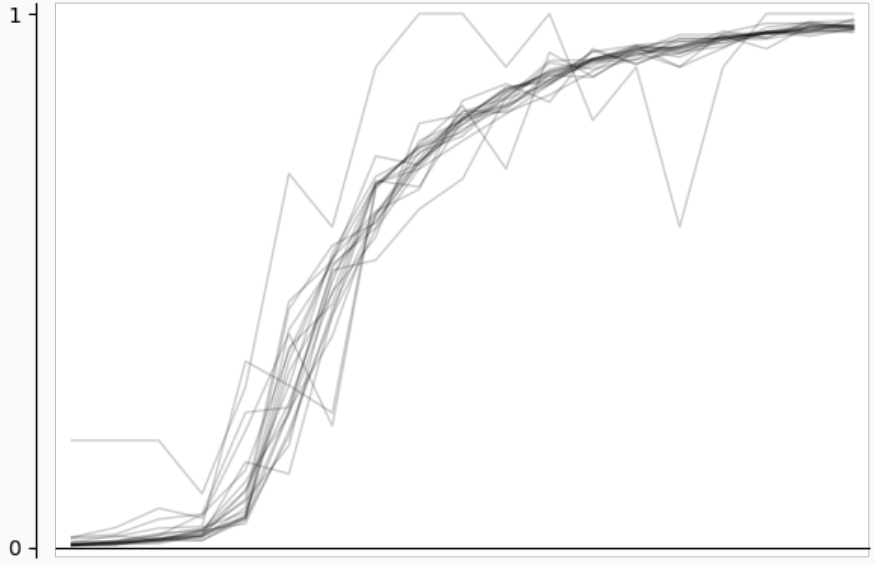


Figure 16: Phase transition for Giant Component

Giant component

- In simple words, the probability $1/n$ is a threshold value for the size of the largest component.
- It is possible to get threshold probabilities for other properties, for instance the threshold on connectedness is $p_c(n) = \ln n/n$.
- The same phenomena happens in directed graphs or hypergraphs, however the threshold happens at different values.

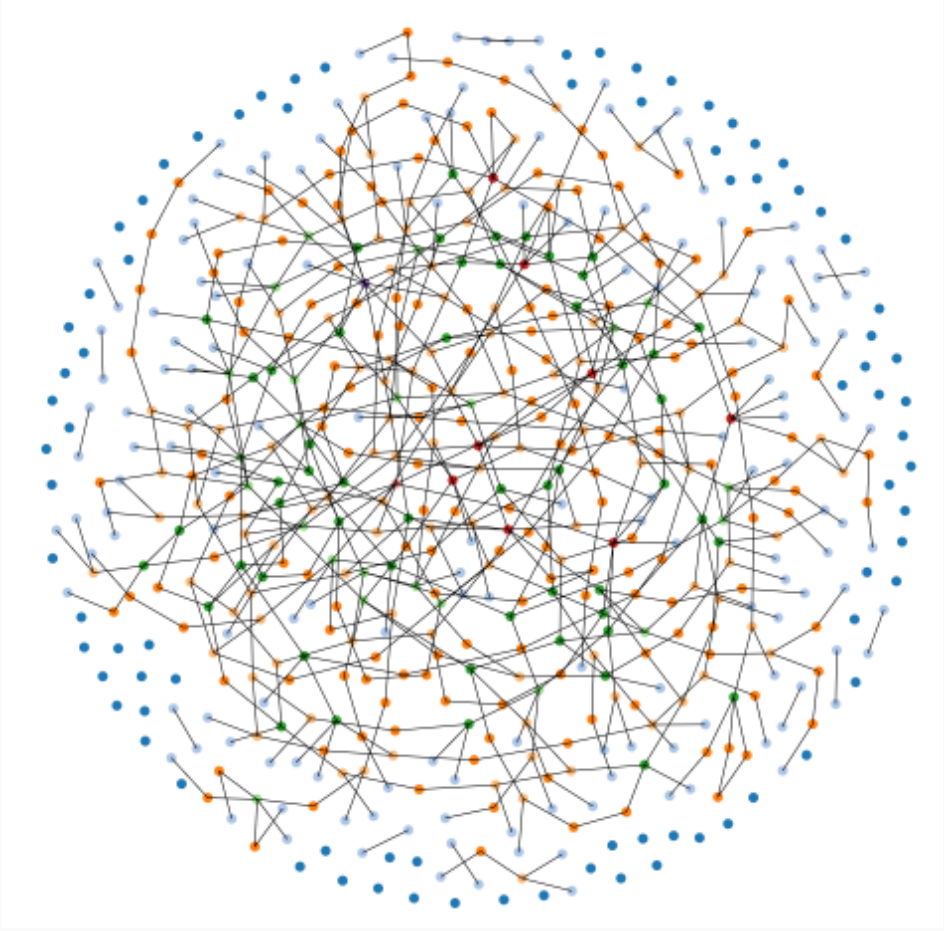


Figure 17: Random graph with 600 nodes and $p = 0.0033$, $c = 2$. fitzner.nl/simulator/index.html

Characteristic path length of $\mathbb{G}_{n,p}$

Theorem from [Bol01]

The charateristic path length of a graph $\mathbb{G}_{n,p}$:

$$L \approx \frac{\ln n}{\ln \langle k \rangle}, \quad \langle k \rangle \text{ is the average node degree}$$

An estimation of this quantity can be found by examining the average number of nodes reached by a node i in m steps which gives:

$$z_m = \langle k \rangle^m$$

Then L can be estimated as the value of m such that the total number of vertices at distance m from a given vertex i in the giant component is of the order of n, which gives:

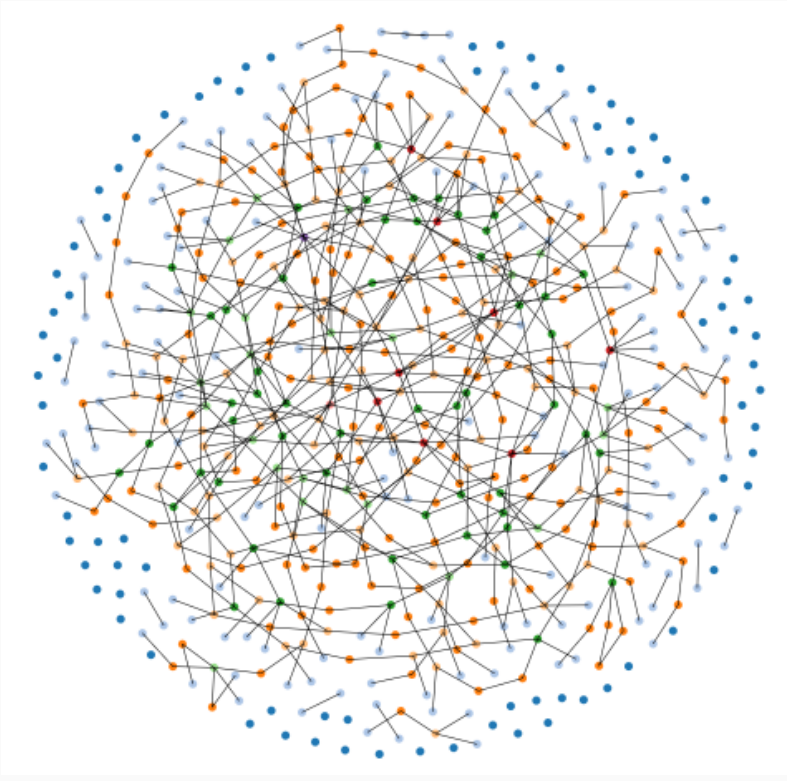
$$n \approx \langle k \rangle^L \implies L \approx \frac{\ln n}{\ln \langle k \rangle}$$

Summary on Erdös-Rényi graphs

- The simple process allows for analytic results to be derived.
- Phase transitions phenomena for giant component, connectedness, ...
- Exhibit small-world effect (small characteristic path length).

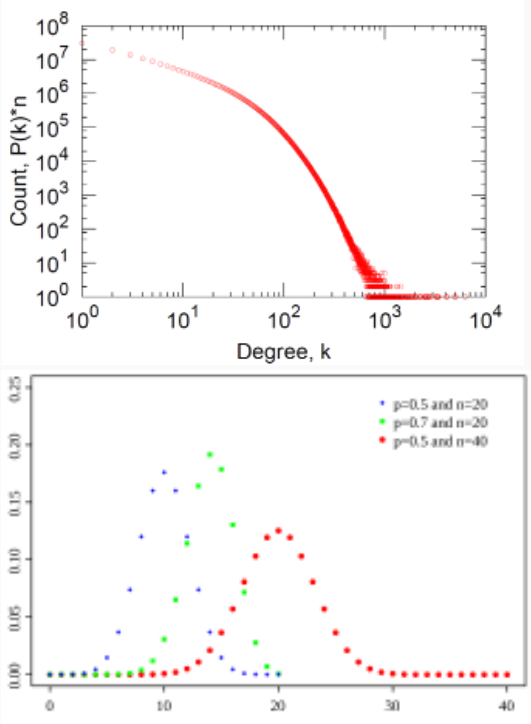
Model	$\mathbb{G}_{n,p}$
Degree Distribution	Binomial
Average node degree $\langle k \rangle$	$(n-1)p$
Characteristic path length	$\frac{\ln n}{\ln \langle k \rangle}$
Global Clustering Coefficient	$\frac{\langle k \rangle}{n}$

Table 1: Characteristics of Erdös-Rényi graphs

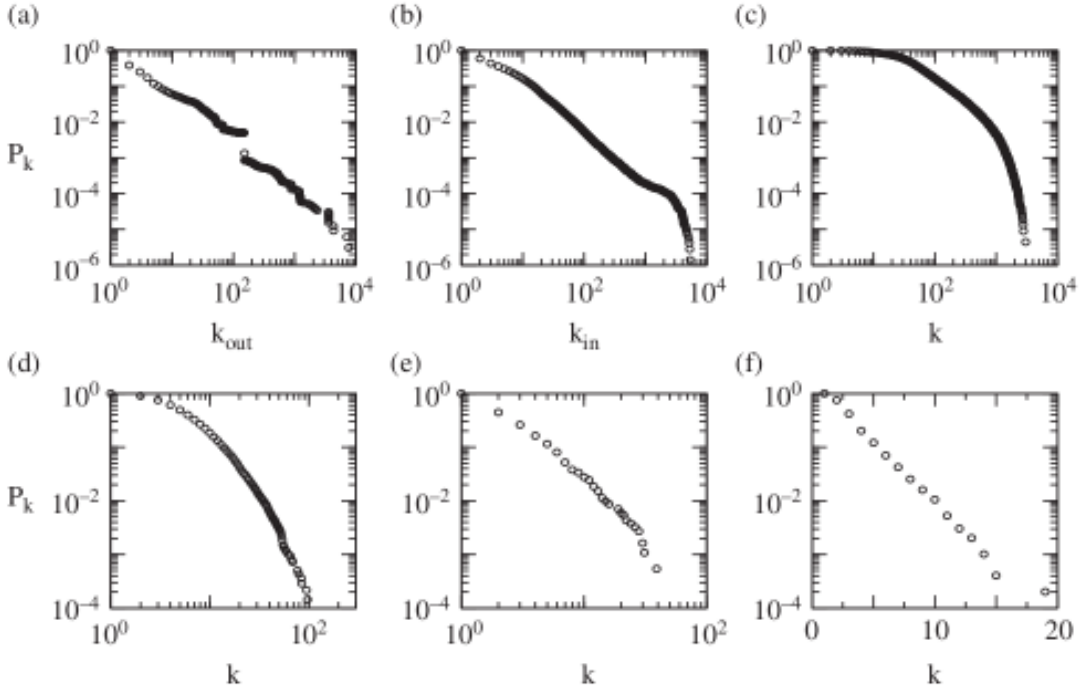


Conclusion on Erdös-Rényi graphs

- Starting point in the study of Random graphs.
- The simplicity implied precise estimation of many quantities.
- The simplicity called for more complex models (closer to real-world graphs) to obtain more varieties on degree distribution, clustering coefficient, average path length...



Degree distribution example



Cumulative degree distributions for a number of different networks, namely: (a) Notre Dame WWW (in-degree), (b) Google WWW (in-degree), (c) movie actor collaborations, (d) cond-mat coauthorships, (e) yeast protein interaction network and (f) the US power grid.

Real networks

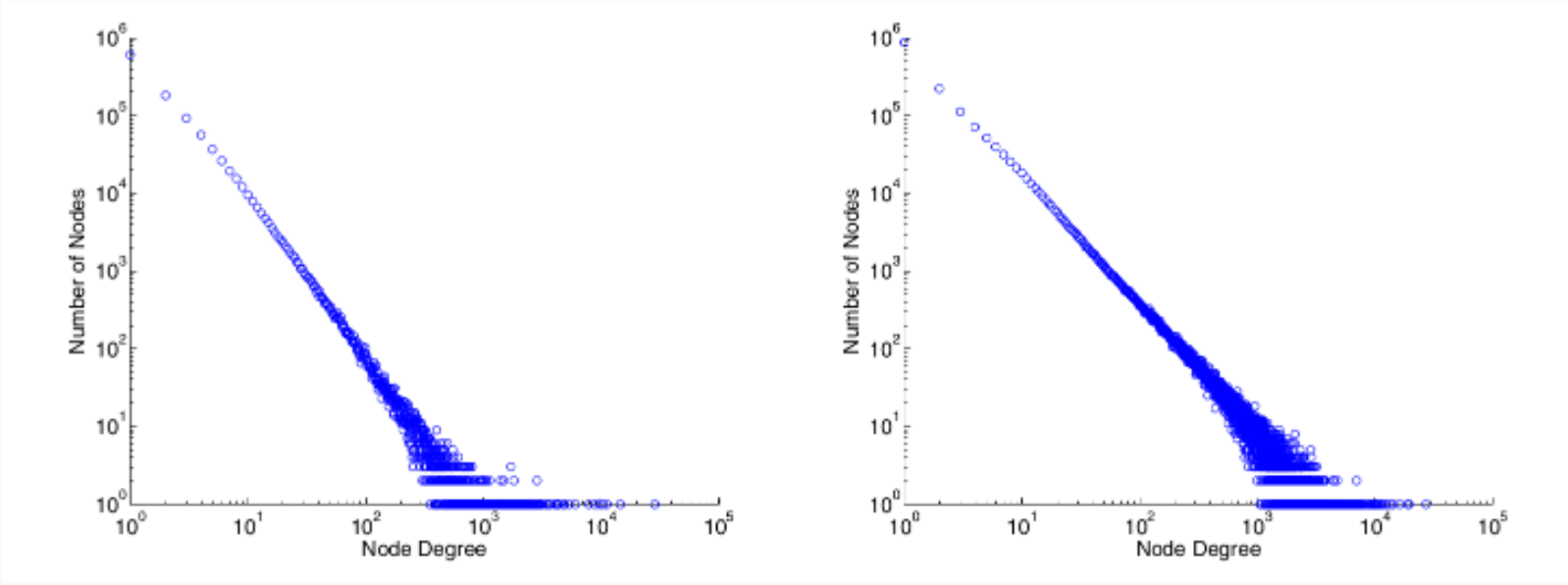


Figure 18: From [TL10], node distributions of a YouTube network with $1,138,499$ nodes and a Flickr network with $1,715,255$ edges (based on data from (Mislove et al., 2007)). For power law distributions, the scatter plot of node degrees is approximately a straight line.

The **average clustering coefficients are** 0.08 **and** 0.18, whereas if the connections are uniformly random, the expected coefficient in a random graph is 4.6×10^{-6} and 10^{-5}

Generalized random graphs

Configuration models

- Introduced by Edward Bender and Rodney Canfield in 1978 [BC78].
- It was later refined by Bollobás in [Bol01].
- There exists 2 models that generalise the models A and B.
- The graphs are parameterised by their degree sequence.

$\Longrightarrow$ This model allows for arbitrary degree distributions

Configuration model A

Definition

Let $n > 0$, $k > 0$, and $\mathbf{k} = (k_1, k_2, \ldots, k_n)$, then $\mathcal{G}_{n,\mathbf{k}}$ is the class of all graphs of n nodes and k edges, and is such that node i has the specified degree k_i. Then, $\mathbb{G}_{n,\mathbf{k}}$ generates a random graph in $\mathcal{G}_{n,\mathbf{k}}$ such that every graph has the same probability.

Configuration model A

Definition

Let $n > 0$, $k > 0$, and $\mathbf{k} = (k_1, k_2, \ldots, k_n)$, then $\mathcal{G}_{n,\mathbf{k}}$ is the class of all graphs of n nodes and k edges, and is such that node i has the specified degree k_i. Then, $\mathbb{G}_{n,\mathbf{k}}$ generates a random graph in $\mathcal{G}_{n,\mathbf{k}}$ such that every graph has the same probability.

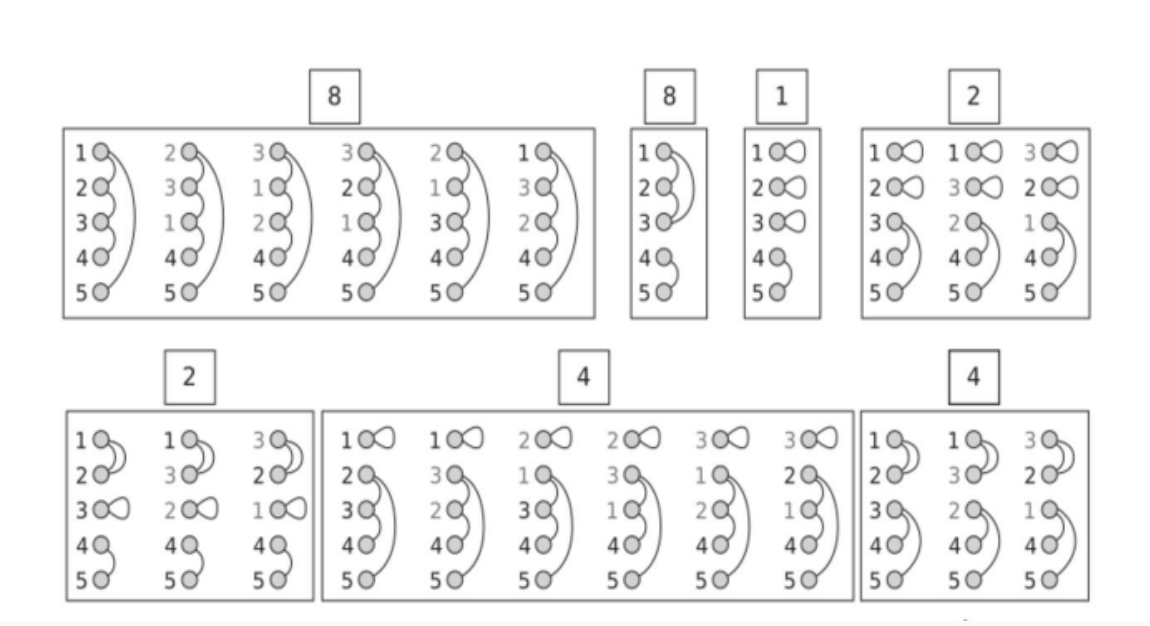


Figure 19: From [LNR17]. The 23 different graphs with $n = 5$ nodes and $k = 4$ links with degree sequence $K = (2, 2, 2, 1, 1)$.

Configuration model A

Simple way to generate all graphs is to assign to each node i, k_i half-edges, then a graph is constructed by connecting half-edges together uniformly randomly.

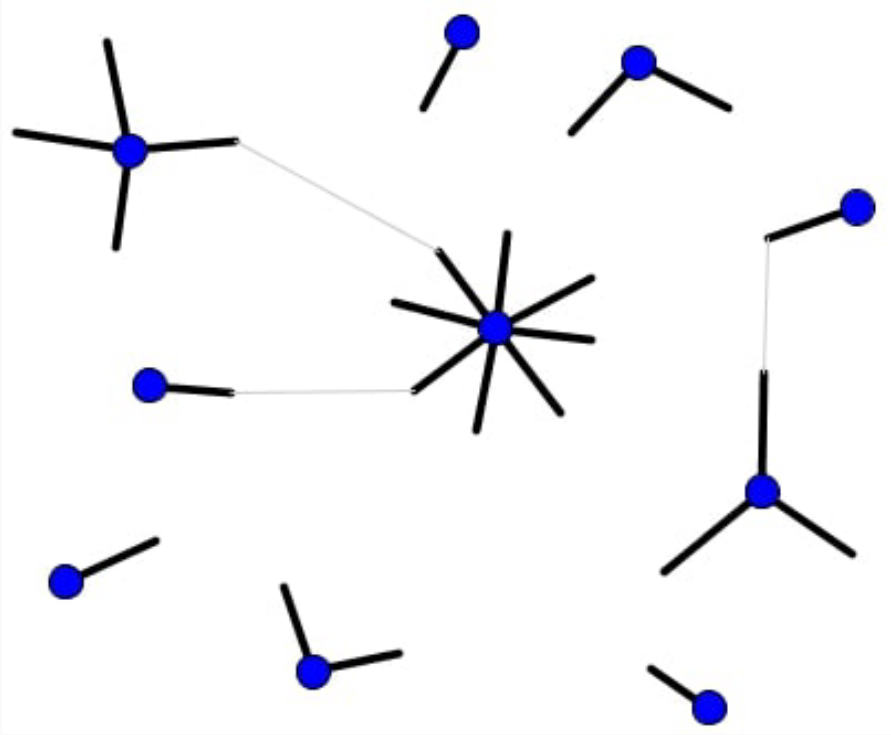


Figure 20: A graph with $n = 10$, $k = 12$ and a degree sequence $\mathbf{k} = (1, 1, 1, 1, 1, 2, 2, 3, 4, 8)$

Matchings in Configuration model A

- Let $\mathbf{k} = (2, 2, 2, 1, 1)$, there are 8 half edges
- We will write ij for the j-th half edges of node i
- **A matching** is then an association of each half edge with another half edge
- Example : the 8 half edges are $\{11, 12, 21, 22, 31, 32, 41, 51\}$ and $\{\{11, 21\}, \{12, 51\}, \{22, 32\}, \{31, 41\}\}$ is a matching
- Each different matching creates a multigraph
- A multigraph is a graph where multiple edges between the same pair of nodes are allowed as well as self-loops
- A **simple graph** is a multigraph having no self-loops nor multiple edges.

Matchings in Configuration model A

- Sampling all matchings with the same probability does not imply that all graphs have the same probability.
- We can put an order on the half-edges of a vertex and call ij, the j-th half-edge of vertex i. Then the first graph in the table has 8 matchings:

1. $\{\{11, 21\}, \{12, 51\}, \{22, 32\}, \{31, 41\}\}$
2. $\{\{11, 22\}, \{12, 51\}, \{21, 32\}, \{31, 41\}\}$
3. $\{\{11, 21\}, \{12, 51\}, \{22, 31\}, \{32, 41\}\}$
4. $\{\{11, 22\}, \{12, 51\}, \{21, 31\}, \{32, 41\}\}$
5. $\{\{12, 21\}, \{11, 51\}, \{22, 31\}, \{32, 41\}\}$
6. $\{\{12, 22\}, \{11, 51\}, \{21, 32\}, \{31, 41\}\}$
7. $\{\{12, 21\}, \{11, 51\}, \{22, 32\}, \{31, 41\}\}$
8. $\{\{12, 22\}, \{11, 51\}, \{21, 31\}, \{32, 41\}\}$

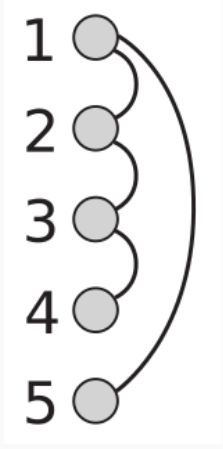


Figure 21: Graph with $n = 5$ nodes and $k = 4$ links with degree sequence $K = (2, 2, 2, 1, 1)$.

Matchings in Configuration model A

- Sampling all matchings with the same probability does not imply that all graphs have the same probability.
- We can put an order on the half-edges of a vertex and call ij, the j-th half-edge of vertex i. while this graph with the same degree sequence has only 2 matchings :

1. $\{\{11, 12\}, \{21, 22\}, \{31, 41\}, \{32, 51\}\}$
2. $\{\{11, 12\}, \{21, 22\}, \{31, 51\}, \{32, 41\}\}$

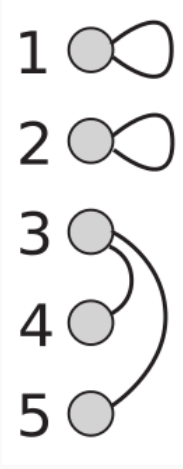


Figure 22: Graph with $n = 5$ nodes and $k = 4$ links with degree sequence $K = (2, 2, 2, 1, 1)$.

Matchings in Configuration model A

- The graphs may contain multiple edges or self-loops.
- It is possible to make a canonical ordering on the couples of halfedges and on the unordered pairs to get $[[51, 32], [41, 31], [22, 21], [12, 11]]$.
- The number of matchings is the number of possible canonical orderings.

Matchings in Configuration model A

$$\{11, 12, 21, 22, 31, 32, 41, 51\}$$

- In a matching start with the highest half-edge 51. Then, there are $(2k-1)$ possible half-edges to join and create the first edge.
- The same process is repeated, there will be $(2k-3)$ half-edges to join second half-edge.
- Then the number of possible matchings of a degree sequence with k edges is:

$$(2k-1)!! = (2k-1)(2k-3)\dots 1$$

Configuration model A

How many different matchings does a multigraph have?

A multigraph G with n nodes and degree distribution $\mathbf{k}$ whose adjacency matrix x_{ij} gives the number of edges between vertices i and j.

- If G is simple, $x_{ii} = 0$ and $i \neq j, x_{ij} = 0$ or $x_{i,j} = 1$.
- For a vertex i of degree k_i permuting its half-edges will generate a different matching of the same graph $\implies$ a vertex i creates $k_i!$ different matchings
- The number of matchings is then $\prod_{i=1}^{n} k_i!$.
- Each self-loop divides the number of matchings by 2.
- Multiple edges between 2 vertices are counted by both $k_i!$ and need to be factored out with $\prod_{1 \leq i \leq j \leq n} x_{ij}!$.

$$\#\text{different matchings of } G = \frac{\prod_{i=1}^{n} k_i!}{\left(\prod_{i=1}^{n} 2^{x_{ii}}\right)\left(\prod_{1 \leq i \leq j \leq n} x_{ij}!\right)}$$

Configuration model A

- From the previous result the probability of drawing a multigraph G with n nodes and degree distribution $\mathbf{k}$ is

$$\frac{\prod_{i=1}^{n} k_i!}{(2k-1)!! \left(\prod_{i=1}^{n} 2^{x_{ii}}\right) \left(\prod_{1 \leq i \leq j \leq n}^{n} x_{ij}!\right)}$$

Result

All simple graphs are such that for all $i, x_{ii} = 0$ and $x_{i,j} = 0$ or $x_{ij} = 1$ and therefore are equiprobable

$$\frac{\prod_{i=1}^{n} k_i!}{(2k-1)!!}$$

Matchings in Configuration model A

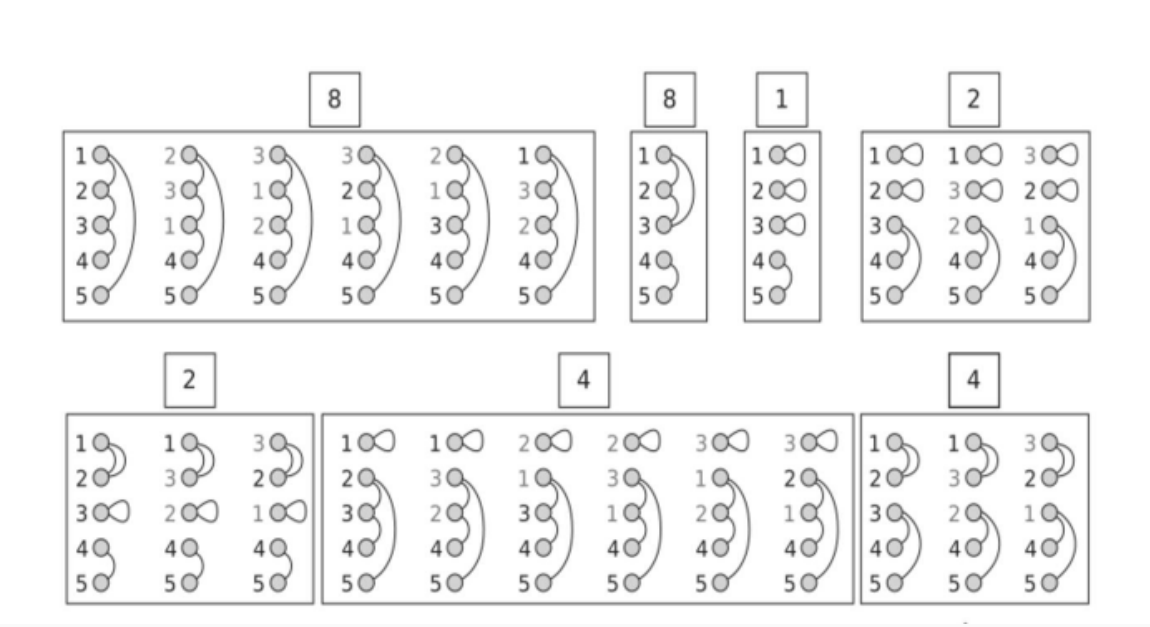


Figure 23: From [LNR17]. The 23 different graphs with $n = 5$ nodes and $k = 4$ links with degree sequence $\mathbf{k} = (2, 2, 2, 1, 1)$.

The simple graphs are the ones in the first 2 rectangles and they all have 8 different matchings.

Number of multiple edges and self-loops

- We will calculate how many links we have on average between a given pair of nodes in a model with self-loops and multiple edges.
- If we have nodes i and j, having, respectively, degrees k_i and k_j. For $1 \leq a \leq k_i$ and $1 \leq b \leq k_j$. Let $X_{i,a}^{j,b}$ be a R.V which is 1 if half-edge ia is connected to jb and 0 otherwise. Let X_i^j be the number of edges between i and j, then

$$X_i^j = \sum_{a=1}^{k_i} \sum_{b=1}^{k_j} X_{i,a}^{j,b}$$

and

$$\mathbb{E}[X_i^j] = \sum_{a=1}^{k_i} \sum_{b=1}^{k_j} \mathbb{E}[X_{i,a}^{j,b}]$$

Number of multiple edges and self-loops

$$\mathbb{E}[X_i^j] = \sum_{a=1}^{k_i} \sum_{b=1}^{k_j} \mathbb{E}[X_{i,a}^{j,b}]$$

- The average value of $\mathbb{E}[X_{i,a}^{j,b}]$ corresponds to the probability that halfedges ia and jb are connected which corresponds to $\frac{1}{2k-1}$.
- Thus the average number of links between i and j with $i \neq j$:

$$\mathbb{E}[X_i^j] = \frac{k_i k_j}{(2k-1)}$$

Exercice

A similar argument can be made for self-loops and find

$$\mathbb{E}[Y_i] = \frac{k_i(k_i-1)}{2(2k-1)}$$

Configuration model A

Let $\mathbb{E}[Y]$ be the average number of self-loops in a graph $G \in \mathbb{G}_{n,\mathbf{k}}$, then:[1]

$$\mathbb{E}[Y] = \sum_{i \in V} \mathbb{E}[Y_i] = \frac{1}{2(2k-1)} \sum_{i \in V} k_i(k_i - 1) \overset{n \to \infty}{=} \frac{1}{2} \left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right)$$

In a similar way, we can compute the average number of multiple edges and get:

$$\mathbb{E}[X] = \sum_{i,j \in V} \mathbb{E}[X_i^j] \overset{n \to \infty}{=} \frac{1}{4} \left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right)^2$$

the expected number of multiple edges depends on the ratio $\frac{\langle k^2 \rangle}{\langle k \rangle}$. If the ratio is finite when $n \to \infty$, then the number of multiple edges goes to 0 as n grows. The same apply for the number of loops.

[1] The formulas $\langle k \rangle = \frac{\sum_i k_i}{n}$ and $\langle k^2 \rangle = \frac{\sum_i k_i^2}{n}$

Giant components of $\mathbb{G}_{n,\mathbf{k}}$

The following Theorem generalises the results presented earlier on Erdös-Rényi graphs

Theorem (Molloy and Reed)

Given a degree distribution $\mathbf{k}$, if

$$\langle k^2 \rangle - 2\langle k \rangle > 0$$

then a.e.g a uniform configuration graph on $\mathbf{k}$ will have a giant component. If the quantity < 0 then all components are small.

Is this result consistent with the one on $\mathbb{G}_{n,p}$?

- $\langle k \rangle = (n-1)p$
- $\langle k^2 \rangle = (n-1)p + (n-1)(n-2)p^2$

Giant components of $\mathbb{G}_{n,\mathbf{k}}$

The following Theorem generalises the results presented earlier on Erdös-Rényi graphs

Theorem (Molloy and Reed)

Given a degree distribution $\mathbf{k}$, if

$$\langle k^2 \rangle - 2\langle k \rangle > 0$$

then a.e.g a uniform configuration graph on $\mathbf{k}$ will have a giant component. If the quantity < 0 then all components are small.

Is this result consistent with the one on $\mathbb{G}_{n,p}$?

- $\langle k \rangle = (n-1)p$
- $\langle k^2 \rangle = (n-1)p + (n-1)(n-2)p^2$
- $\langle k^2 \rangle - 2\langle k \rangle = (n-1)(n-2)(c/n)^2 - (n-1)(c/n) \sim c(c-1)$

Giant components of $\mathbb{G}_{n,\mathbf{k}}$

The following Theorem generalises the results presented earlier on Erdös-Rényi graphs

Theorem (Molloy and Reed)

Given a degree distribution $\mathbf{k}$, if

$$\langle k^2 \rangle - 2\langle k \rangle > 0$$

then a.e.g a uniform configuration graph on $\mathbf{k}$ will have a giant component. If the quantity < 0 then all components are small.

Is this result consistent with the one on $\mathbb{G}_{n,p}$?

- $\langle k \rangle = (n-1)p$
- $\langle k^2 \rangle = (n-1)p + (n-1)(n-2)p^2$
- $\langle k^2 \rangle - 2\langle k \rangle = (n-1)(n-2)(c/n)^2 - (n-1)(c/n) \sim c(c-1)$
- In $\mathbb{G}_{n,p}$ the giant component transition happens at $c = 1$, which is consistent

Characteristic path length of $\mathbb{G}_{n,\mathbf{k}}$

Theorem

The charateristic path length of a graph $\mathbb{G}_{n,\mathbf{k}}$:

$$L \approx \frac{\ln\left(\frac{n}{\langle k\rangle}\right)}{\ln\left(\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right)} + 1$$

An estimation of this quantity can be found by examining the average number of nodes reached by a node i in m steps which gives:

$$z_m = \rho^{m-1} z_1, \quad \rho = \frac{\langle k^2\rangle}{\langle k\rangle} - 1$$

where $z_1 = \langle k\rangle$.

Then L can be estimated as the value of m such that the total number of vertices at distance m from a given vertex i in the giant component is of the order of n, which gives:

$$n \approx z^{\ell} = \rho^{\ell-1} z_1$$

Random Generation configuration model A

Depending on what we want to do we can have different random generation algorithms:

- Pragmatic developer $\rightarrow$ concerned with efficiency not uniformity
- Logician $\rightarrow$ concerned only of uniformity
- Probabilist $\rightarrow$ using markov chains and random walks.
- Biologist $\rightarrow$ with genetic like methods.

Configuration model A : Pragmatic Developer generator

Matching method

- Generate a matching uniformly and build its corresponding graph.
- The graph might contains multiple-edges and self loops.
- In some simulations it does not matter.

Advantages and inconveniences

- Efficient generation.
- Uniformity among simple graphs if the generated graph is simple.

Configuration model A : Pragmatic Developer generator 2

Matching method

- Generate a uniform matching.
- (variant) If the generated graph is a multigraph just erase the self-loops and multiedges!

Advantages and inconveniences

- Efficient generation.
- We do not get the exact initial distribution
- No uniformity among simple graphs.

Configuration model A : Pragmatic Developer generator 3

Matching method

- Generate a uniform matching.
- (variant II) During the process if the selected half-edges form a self-loop or a multiple edge we discard them and pick another pair of stubs.

Advantages and inconveniences

- The method generates a biased sample of possible simple graphs.
- Efficient algorithm but not uniform.

Configuration model A : Logician generator

Matching method. Uniformity at all price.

- Generate uniformly a matching and build the corresponding graph.
- Simple graphs have the same probability.
- Therefore, if the graph is simple we just return it.
- If it is not the case, we sample another matching until we get a simple graph.

Advantages and inconvenients

- Uniformity guaranteed
- It is possible to compute the probability of a simple graph thus knowing number of sampling needed on average.
- Depeding on the distribution it might take a long time.

Configuration model A : Probabilist generator

For the probabilsts models we refer to [MKI+03]. This method is known as **Switching method**

- Start from a given network with the right distribution (Havel-Hakimi algorithm).
- Make a series of switching where $\{A, B\}, \{C, D\}$ is transformed to $\{A, D\}, \{C, B\}$.
- The exchange is only made if it does not create a multiedge or a self-loop.
- The process is repeated some number Qk times, where k is the number of edges in the graph and Q is chosen large enough that the Markov chain shows good mixing.

Advantages and inconveniences

- The large Q was not quite known and used to be determined experimentally. Recent studies showed $O((\ln n)^2)$ rewirings needed
- We eventually get a unifrom simple graph!

Configuration model A : Biologist genetic generator

- Consider a colony of M graphs.
- Start with the appropriate number of half-edges for each vertex and repeatedly choose at random 2 half-edges from the graph and link them together to create an edge.
- If a multiple edge or self-edge is generated, the graph containing it is removed from the colony and discarded (it dies).
- To compensate for decline in the colony, its size is periodically doubled by cloning each of the surviving graphs (survival of the fittest).
- The process is repeated until all half-edges have been linked, then one network is chosen uniformly.

Advantages and inconveniences

- Generation slowier than switching and matching methods.

Uniformity tests

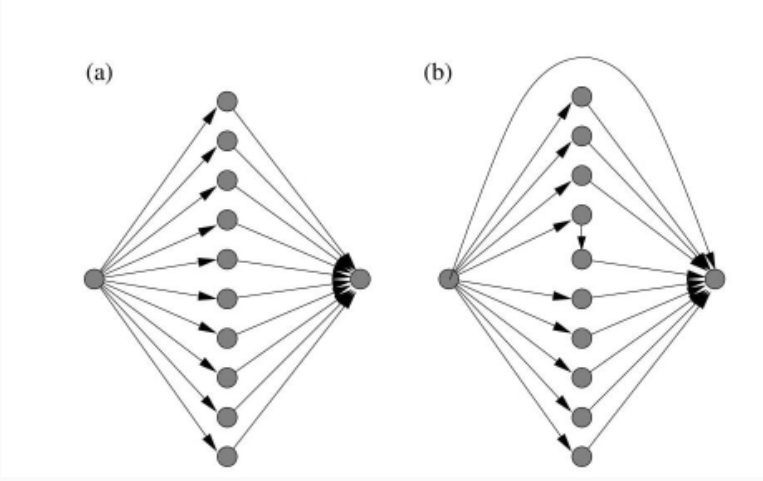


Figure 24: From [MKI+03],The two types of graphs possible with the degree distribution. One of them appears 90 times and the other only one time.

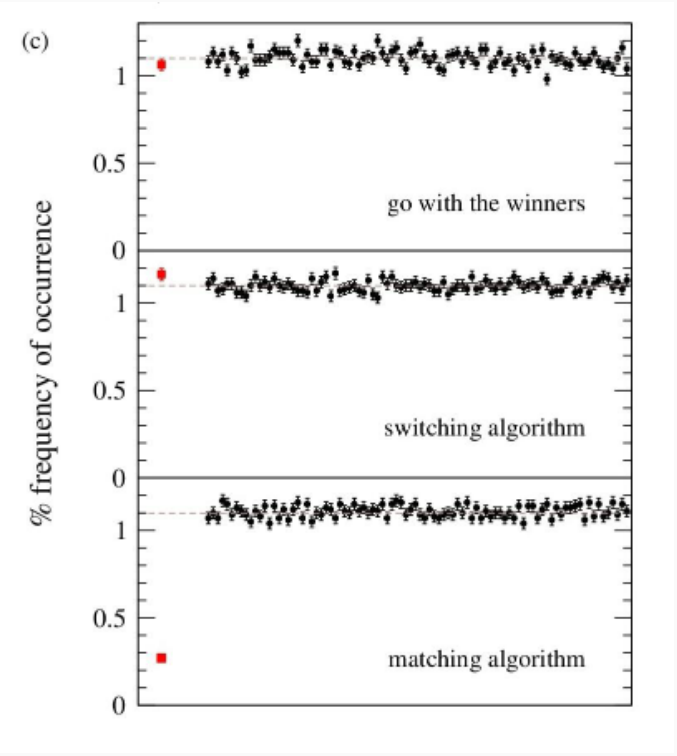


Figure 25: From [MKI+03], Uniformity tests of the three algorithms, it shows the frequency with which each configuration is sampled by our three algorithms. 100 000 graphs were generated with each algorithm.

Summary on generators

	Developer1	Developer2	Logician	Switching	Developer3	Biologist
Simple graphs	no	yes	yes	yes	yes	yes
Distribution node degrees	exact	approximate	exact	exact	exact	exact
Uniformity over simple graphs	yes	no	yes	yes	no	yes
Possible infinite loops	no	no	yes	yes	yes	no
Time complexity	$O(k)$	$O(k)$	$O(k)^*$	$O(m \ln n)^{**}$	$O(k)^*$	$O(Mk)^{***}$

Table 2: Comparison between generators

* : If prob. self or multiple edges is fixed on average we will reject a constant number of graphs or edges.

** : $O(m \ln n)$ Experimental studies switching.

*** : The constant in the $O(k)$ is large, the switching and biologist algorithms are faster experimentally.

Configuration model B

The configuration model A, is parameterised with a fix number of edges, in a sense it generalises model A in Erdös-Rényi graphs. We want to introduce a configuration model B that generalises model B of Erdös-Rényi graphs.

Let $n > 0$, and $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ where $w_i \geq 0$ and $\max_i w_i^2 < \sum_i w_i$. Let m be defined as $\sum_i w_i = 2m$. Then $\mathbb{G}_{n,\mathbf{w}}$ is a graph on n nodes in which expected number of links between different pairs of nodes is given by:

$$p_{i,j} = \frac{w_i w_j}{2m}$$

The degree of a node can fluctuate, the expextation of the degree of node i gives:

$$\bar{k}_i = \sum_j \frac{w_i w_j}{2m} = w_i$$

Configuration model B

Then, the average value of the total number of edges is

$$\bar{k} = \frac{1}{2}\sum_i \bar{k}_i = \frac{1}{2}\sum_i w_i = m$$

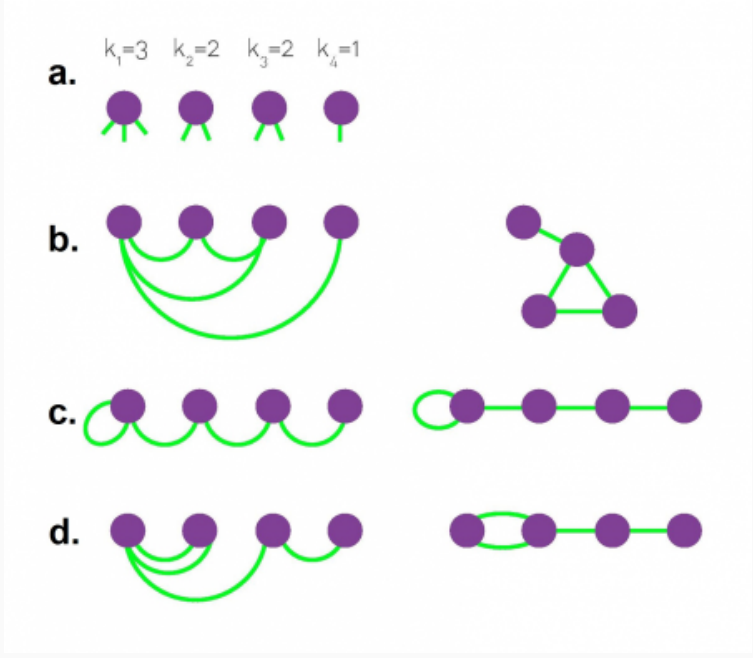
- This model is called Chung–Lu model.
- Generalises model B of Erdös-Rényi graphs take $\forall i, w_i = pn$.

Sampling :

- Take each possible node pairs and generate a link or not following to the probability.
- The Chung-Lu model does not produce exactly $\mathbf{w}$, and instead generates networks whose degrees are $\mathbf{w}$ in expectation.

Summary on Generalized random graphs

Model	$\mathbb{G}_{n,p}$	$\mathbb{G}_{n,\mathbf{k}}$
Degree Distribution	Binomial	Arbitrary
Average node degree $\langle k \rangle$	$(n-1)p$	k_i
Characteristic path length	$\frac{\ln n}{\ln \langle k \rangle}$	$\frac{\ln\left(\frac{n}{\langle k \rangle}\right)}{\ln\left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}\right)} + 1$
Global Clustering Coefficient	$\frac{\langle k \rangle}{n}$	$\frac{\langle k \rangle}{n}\left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle^2}\right)^2$



Growing networks

Power laws

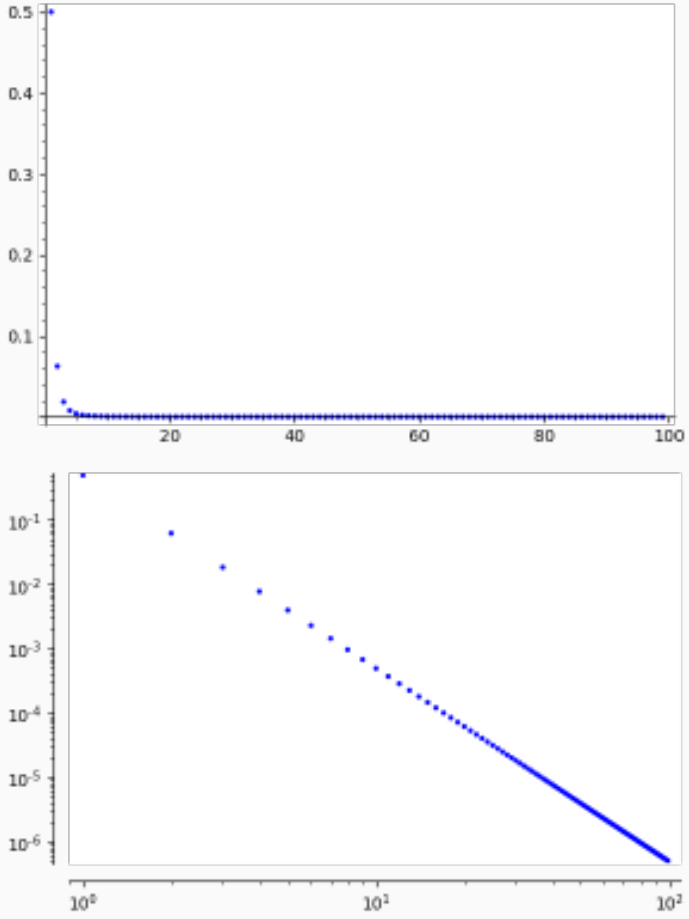
A discrete power law is in the form of:

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma)}, \quad k > 0, \gamma \in \mathbb{R},$$

Scale-free or scale-invariant property:

$$f(\lambda x) = a(\lambda) f(x)$$

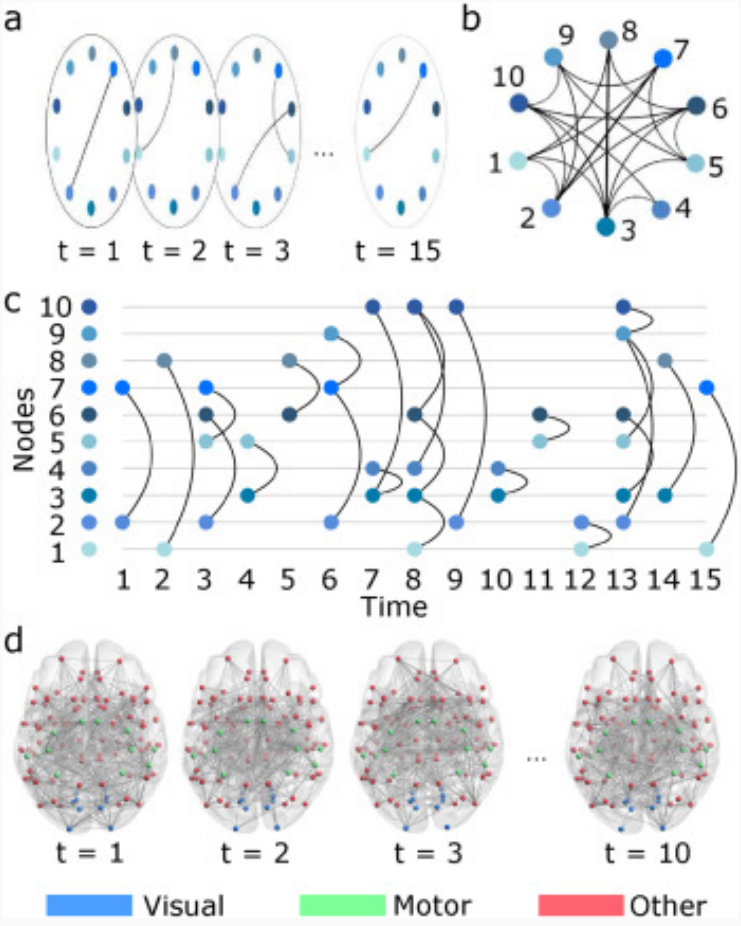
$\implies$ Power law distributions are scale-free



$f(x) = 0.5x^{-3}$ (up) normal scale. (down) log scale. The exponent can be better seen in log scale

Need for evolving networks

- Networks around us are dynamic in nature
- Examples include World Wide Web and Networks of citations, ...
- Different formalisms have been proposed to model dynamicity
- We are interested in evolving networks in the form of a power law



Case study : Scientometrics dataset

- Scientometrics is a monthly peer-reviewed academic journal
- Papers from Scientometrics 1978–2004
- Number of papers 1655 (nodes)
- 3904 citation links among the papers where found

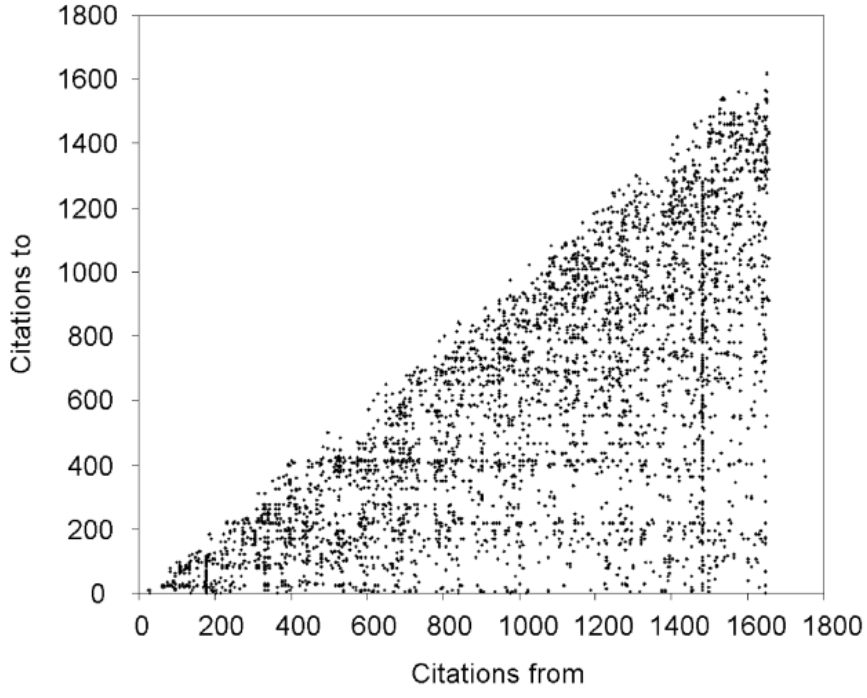
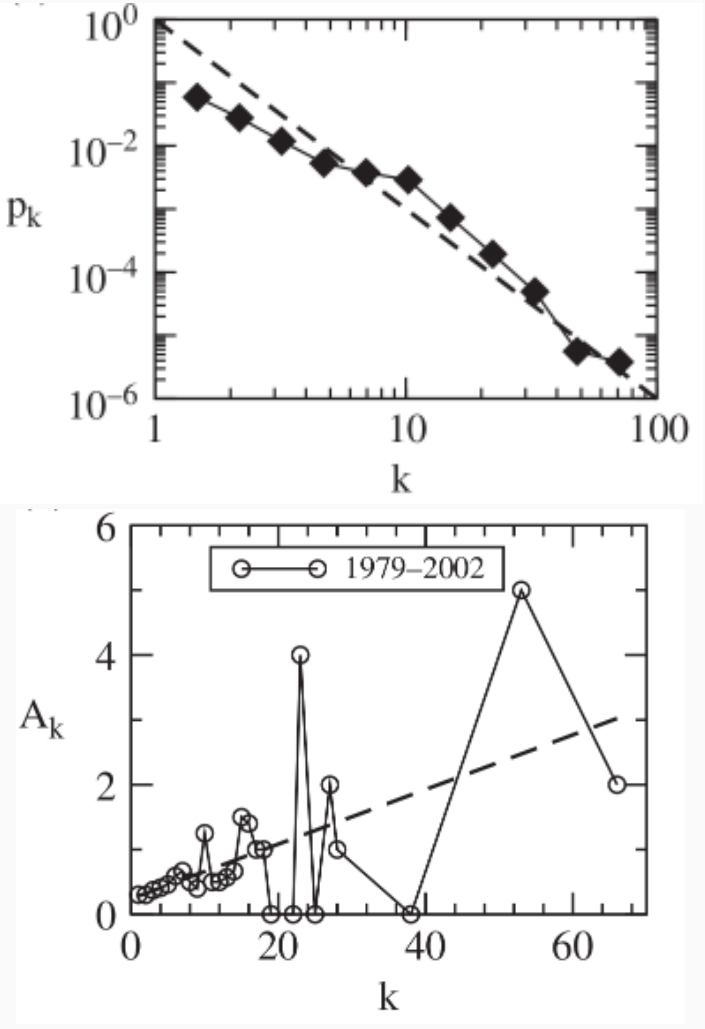


Figure 1.The reference-network graph for *Scientometrics* 1978–2004

From [Per06]

Case study : Scientometrics dataset

- Highly cited papers appear as horizontal lines of dots
- Vertical lines of dots are typical review papers
- Distribution of dots denser close to diagonal
- authors tend to give recent papers more attention than older ones
- Network accumulates new links proportionally to their link number



(up) Degree distribution, and dashed line $\sim k^{-3}$.
(down) Growth rate according to citations

Albert-László Barabási and Réka Albert

Figure 26: Albert-László Barabási

- Hungarian American physicist at Northeastern University, USA
- Introduced the concept of scale-free network
- One of the founders of network science field
- (link) Video link of Barabási on network science

Figure 27: Réka Albert

- Romanian Hungarian physicist at Pennsylvania State University, USA
- Introduced the concept of scale-free network
- Made contributions to biology systems

The Barabási–Albert (BA) Model

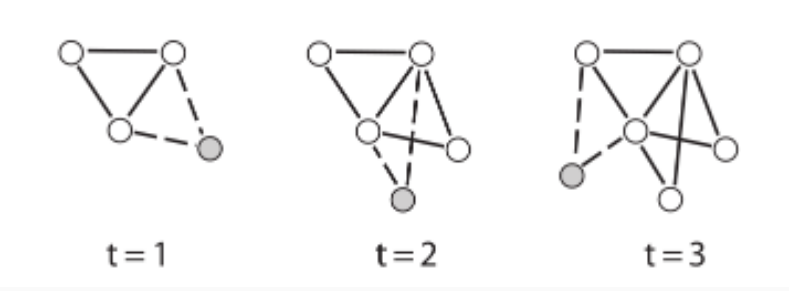
- Linear preferential attachment "Rich gets richer"

Definition:

- Given n_0 and $m \leq n_0 \leq n$, let n_t, l_t number of nodes and links at time t.
- At time 0 start with a complete graph on n_0 nodes
- At time $t > 0$, a new node is added labelled $n = n_0 + t$, and add m edges according from the new node to already existing ones according to

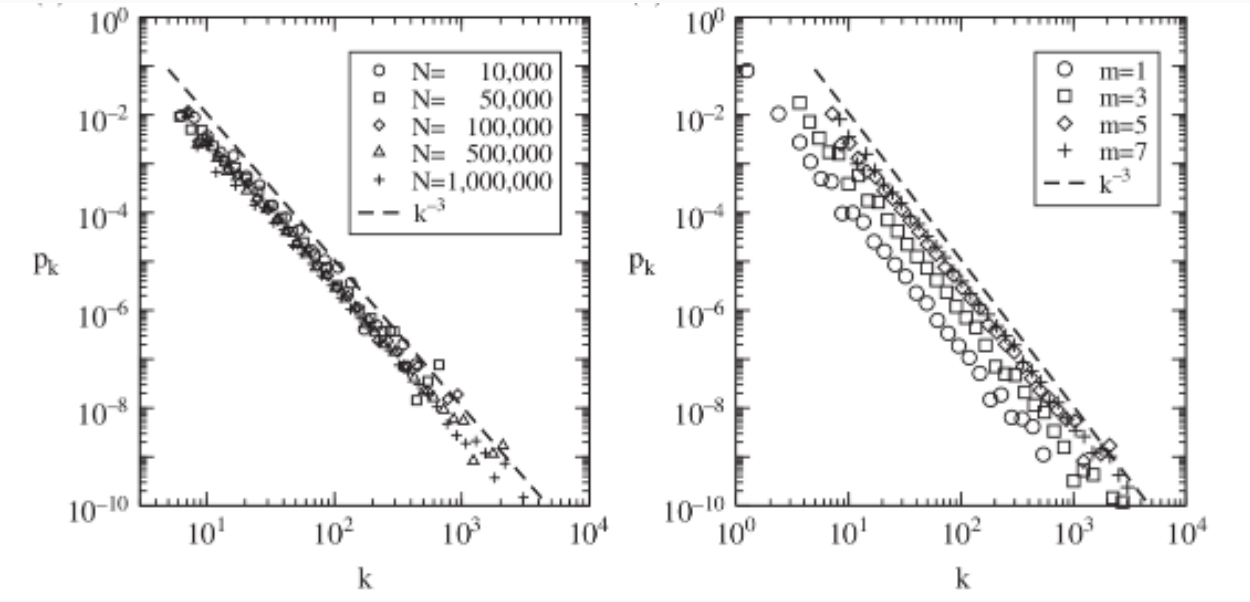
$$P(n \to i) = \frac{d_{i,t-1}}{\sum_j d_{j,t-1}} = \frac{d_{i,t-1}}{2l_{t-1}}$$

where $d_{i,k}$ is the degree of node i at time k.



$n_0 = 3, m = 2$, new node shaded, new edges dashed

Experiment Degree Distribution of BA



- The distribution experimentally behaves like a power law of exponent 3
- The distribution or exponent are not affected by larger N or different m

Degree Distribution

Let $\bar{n}_{k,t}$ be the average number of nodes of degree k at time t

$$\bar{n}_{k,t} = \bar{n}_{k,t-1} - LOSS + GAIN, \quad k \geq m$$

After some calculation:

$$\bar{n}_{k,t} - \bar{n}_{k,t-1} = m\frac{(k-1)\bar{n}_{k-1,t-1} - k\bar{n}_{k,t-1}}{2l_{t-1}} + \delta_{k,m}$$

Since we are interested in $p_{k,t} = \frac{\bar{n}_{k,t}}{n_t}$

$$n_{t-1}(p_{k,t} - p_{k,t-1}) = -p_{k,t} + \frac{mn_{t-1}}{l_{t-1}}\left(\frac{k-1}{2}p_{k-1,t-1} - \frac{k}{2}p_{k,t-1}\right) + \delta_{k,m} \quad k \geq m$$

when $t \to \infty$, and since $n_t \approx t$ and $l_t \approx mt$ we find

$$(t-1)(p_{k,t} - p_{k,t-1}) = Q_k, \quad Q_k = -p_k + (\frac{k-1}{2} - \frac{k}{2}p_k) + \delta_{k,m}$$

Giving

$$p_{k,t} = p_{k,t-1} + \frac{Q_k}{t-1}$$

To avoid divergence Q_k must be 0 implying

$$p_k = \frac{k-1}{2}p_{k-1} + \frac{k}{2}p_k + \delta_{k,m}, k \geq m$$

Solving for p_k:

Result

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)} \simeq 2m(m+1)k^{-3}$$

Properties of BA

We can show that

$$\bar{k}_{i,t} = m \left(\frac{t}{t_i} \right)^{1/2}$$

where $\bar{k}_{i,t}$ is the average degree of node i at time t and t_i the time at which node i was introduced.

$\Longrightarrow$ Older nodes have higher degrees

Characteristic path length:

- BA graphs have small-world property
- L scales logarithmically with N and is smaller than ER graphs

Clustering coefficient:

- $C \sim N^{-0.75}$ which is slower than ER graphs $C \sim N^{-1}$
- The clustering coefficient still goes to 0

Summary on Random Graph models

- Erdös-Rényi graphs
- Generalized random graphs (Configuration model)
- Barabási–Albert graphs

Model	$\mathbb{G}_{n,p}$	$\mathbb{G}_{n,\mathbf{k}}$	$BA_{n,m}$
Degree Distribution	Binomial	Arbitrary	Power Law
Average node degree $\langle k \rangle$	$(n-1)p$	k_i	m
Characteristic path length	$\frac{\ln n}{\ln \langle k \rangle}$	$\frac{\ln\left(\frac{n}{\langle k \rangle}\right)}{\ln\left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}\right)} + 1$	$c \log n$?
Global Clustering Coefficient	$\frac{\langle k \rangle}{n}$	$\frac{\langle k \rangle}{n}\left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle^2}\right)^2$	$c' \, n^{-0.75}$

Null models more details

- Random graphs are remarkably useful as a null model for investigating the structure of a real graph G.
- They allow to answer the question How much of an observed pattern is explained by edge density or degrees alone, under randomness?
- The idea is to generate a random graph of the studied model and looking for the desired parameters if they are still the same in the random model or not, if it is the case then we can say that it is "explained" by the null's underlying assumptions.

Conclusion

- Basic notions on graphs (component, clustering, automorphism, . . .)
- Erdös-Rényi graphs
  - Parameters
  - Phase transitions
- Graphs of fixed degree distribution (Configuration model)
  - Parameters
  - Random generation
- Barabási-Albert Model
  - Parameters
  - Evolution in time

All models had clustering coefficient $\to 0$.

$\Longrightarrow$ Still need other models for instance Watts and Strogatz models

Some material has been borrowed from the following links

for good slides on random graphs and network analysis :

- http://lioneltabourier.fr/documents/course1.pdf.
- https://aaronclauset.github.io/teaching.htm
- https://www.ndsu.edu/pubweb/ novozhil/Teaching
- https://math.stackexchange.com/questions/3676422/edge-probability-and-expected-number-of-edges-in-the-configuration-model
- https://www.fitzner.nl/simulator/

References

[BC78] Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

[Bol01] Béla Bollobás. *Random graphs*. Number 73. Cambridge university press, 2001.

[ER+60] Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[ER61] Paul Erdös and Alfréd Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

References ii

[ER68] P Erdös and A Rényi. On random matrices ii. *Studia Sci. Math. Hungar*, 3:459–464, 1968.

[Erd59a] P Erd. os and a. r enyi,\on random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[Erd59b] Paul Erdös. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

[FK16] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[Gil59] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[LNR17] Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex networks: principles, methods and applications*. Cambridge University Press, 2017.

References iii

[MKI+03] Ron Milo, Nadav Kashtan, Shalev Itzkovitz, Mark EJ Newman, and Uri Alon. On the uniform generation of random graphs with prescribed degree sequences. *arXiv preprint cond-mat/0312028*, 2003.

[Per06] Olle Persson. Exploring the analytical potential of comparing citing and cited source items. *Scientometrics*, 68(3):561–572, 2006.

[TL10] Lei Tang and Huan Liu. Community detection and mining in social media. *Synthesis lectures on data mining and knowledge discovery*, 2(1):1–137, 2010.