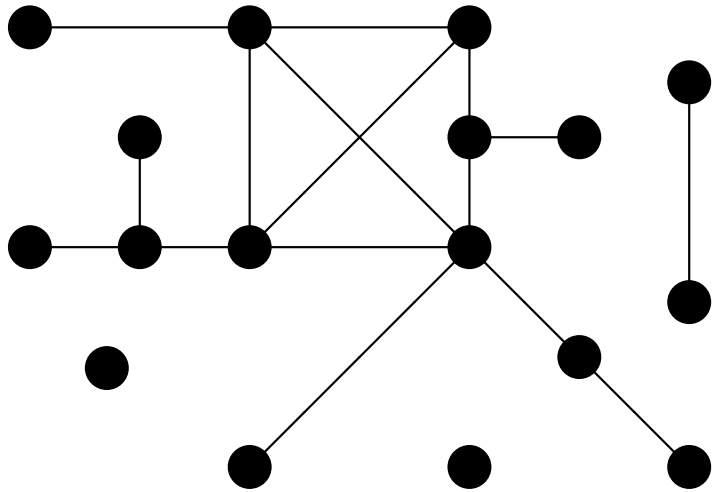
Random Graphs

Svante Janson

Stockholm, 29 January 2014

Graphs

A graph is a set of *nodes* (or *vertices*) together with *edges* (or *links*), where each edge connects two nodes. Sometimes the nodes are directed from one node to the other, but for simplicity we will ignore that possibility today.



Random graphs

A *random graph* is a graph where nodes or edges or both are created by some random procedure.
First example: (classical random graphs studied by Erdős and Rényi and many others from 1959 and until today – often called Erdős–Rényi graphs)

Fix two (large) numbers n (number of nodes) and m (number of edges). Number the nodes $1, \dots, n$. Draw two nodes at random and join them by an edge. Repeat m times. Denoted $G(n, m)$.

Random graphs

A *random graph* is a graph where nodes or edges or both are created by some random procedure.
First example: (classical random graphs studied by Erdős and Rényi and many others from 1959 and until today – often called Erdős–Rényi graphs)

Fix two (large) numbers n (number of nodes) and m (number of edges). Number the nodes $1, \dots, n$. Draw two nodes at random and join them by an edge. Repeat m times. Denoted $G(n, m)$.

A variant: Fix n (number of nodes) and a probability p. For each pair of nodes, make a random choice and connect the nodes by an edge with probability p. (Toss a biased coin, throw dice, get a random number, or use some other random procedure.) Denoted $G(n, p)$.

Why?

Some reasons why a pure mathematician (like me) might want to study random graphs:

- Random graphs provide many interesting problems in probabilistic combinatorics.
  There are many nice theorems with results that are easy to state, and sometimes surprising. Some of them are easy to prove, while others are quite difficult. Many different methods from probability theory or combinatorics are used, and there is plenty of opportunity to develop new methods.

Why?

Some reasons why a pure mathematician (like me) might want to study random graphs:

- Random graphs provide many interesting problems in probabilistic combinatorics.
  There are many nice theorems with results that are easy to state, and sometimes surprising. Some of them are easy to prove, while others are quite difficult. Many different methods from probability theory or combinatorics are used, and there is plenty of opportunity to develop new methods.
- Some colleagues claim that they are useful for applications.

Some reasons why an applied mathematician might want to study random graphs:

- Graphs are used to describe possible infection routes for an infectious disease. Typically, the graph is not known in detail (and even if it is, it will be different tomorrow), and a suitable random graph may be used as a model.

Some reasons why an applied mathematician might want to study random graphs:

- Graphs are used to describe possible infection routes for an infectious disease. Typically, the graph is not known in detail (and even if it is, it will be different tomorrow), and a suitable random graph may be used as a model.
- Graphs and random graphs are used to describe the structure of the Internet. (In several different ways.) Again a suitable random model may be useful.

Some reasons why an applied mathematician might want to study random graphs:

- Graphs are used to describe possible infection routes for an infectious disease. Typically, the graph is not known in detail (and even if it is, it will be different tomorrow), and a suitable random graph may be used as a model.
- Graphs and random graphs are used to describe the structure of the Internet. (In several different ways.) Again a suitable random model may be useful.
- Graphs are used to describe a lot of things, for example references between scientific papers, collaborations (joint publications) between scientists, metabolic reactions in *E. coli*, interactions between proteins in yeast, telephone calls in a given day, ... A suitable random model may be useful.

- We may study a random graph in order to compare its properties with known data from a real graph.
  Perhaps in order to adapt the parameters.

- We may study a random graph in order to compare its properties with known data from a real graph.
  Perhaps in order to adapt the parameters.
- If a simple random model reproduces some interesting properties of a graph, that is a strong warning that we should not give too much significance to those features and waste time inventing complicated explanations for them.

- We may study a random graph in order to compare its properties with known data from a real graph.
  Perhaps in order to adapt the parameters.
- If a simple random model reproduces some interesting properties of a graph, that is a strong warning that we should not give too much significance to those features and waste time inventing complicated explanations for them.
- Conversely, if a random graph model fails to reproduce some properties, this shows that there are other reasons behind these properties, and it is meaningful to study them further.

- We may study a random graph in order to compare its properties with known data from a real graph.
  Perhaps in order to adapt the parameters.
- If a simple random model reproduces some interesting properties of a graph, that is a strong warning that we should not give too much significance to those features and waste time inventing complicated explanations for them.
- Conversely, if a random graph model fails to reproduce some properties, this shows that there are other reasons behind these properties, and it is meaningful to study them further.
- If we believe in our model, we can use it to predict e.g. the spread of an epidemic.

- We may study a random graph in order to compare its properties with known data from a real graph.
  Perhaps in order to adapt the parameters.
- If a simple random model reproduces some interesting properties of a graph, that is a strong warning that we should not give too much significance to those features and waste time inventing complicated explanations for them.
- Conversely, if a random graph model fails to reproduce some properties, this shows that there are other reasons behind these properties, and it is meaningful to study them further.
- If we believe in our model, we can use it to predict e.g. the spread of an epidemic.

Back to theory!

Node degrees

The *degree* of a node is the number of links connecting the node to other nodes.
The classical Erdős–Rényi random graphs have node degrees that are random, but with a rather small random dispersion and very small probability of having a degree that is much larger than the average. (The distribution is Binomial or Hypergeometric, and asymptotically Poisson.)

Small subgraphs

Let H be a fixed graph, for example K_3. Let X_n be the number of copies of H as subgraphs of $G(n,p)$.
Let H have v_H vertices and e_H edges, and let $d_H = e_H/v_H$.

$$\mathbb{E}\, X_n \sim c n^{v_H} p^{e_H}$$

Thus $\mathbb{E}\, X_n \to \infty \iff p \gg n^{-v_H/e_H} = n^{-1/d_H}$.

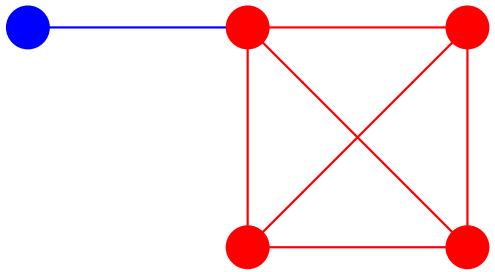
Small subgraphs

Let H be a fixed graph, for example K_3. Let X_n be the number of copies of H as subgraphs of $G(n,p)$.
Let H have v_H vertices and e_H edges, and let $d_H = e_H/v_H$.

$$\mathbb{E}\, X_n \sim c n^{v_H} p^{e_H}$$

Thus $\mathbb{E}\, X_n \to \infty \iff p \gg n^{-v_H/e_H} = n^{-1/d_H}$.
If $d_F \le d_H$ for all subgraphs $F \subset H$ (H is balanced), this is also necessary and sufficient for $\mathbb{P}(X_n > 0) \to 1$. ($X_n > 0$ w.h.p.)

Example:



An unbalanced graph. $d_H = 7/5 = 1.4$, $d_F = 6/4 = 1.5$.

Theorem

(A typical threshold result.)

Let $m_H = \max\{d_F : F \subseteq H\}$. *Then*

$X_H > 0$ *w.h.p.* $\iff p \gg n^{-1/m_H}$.

$X_H = 0$ *w.h.p.* $\iff p \ll n^{-1/m_H}$.

Theorem

(A typical threshold result.)

Let $m_H = \max\{d_F : F \subseteq H\}$. *Then*

$X_H > 0$ *w.h.p.* $\iff p \gg n^{-1/m_H}$.

$X_H = 0$ *w.h.p.* $\iff p \ll n^{-1/m_H}$.

Theorem (Rucinski (1988))

If $p \gg n^{-1/m_H}$ *and* $n^2(1-p) \to \infty$, *then* X_H *is asymptotically normal.*

Theorem

(A typical threshold result.)

Let $m_H = \max\{d_F : F \subseteq H\}$. *Then*

$X_H > 0$ *w.h.p.* $\iff p \gg n^{-1/m_H}$.

$X_H = 0$ *w.h.p.* $\iff p \ll n^{-1/m_H}$.

Theorem (Rucinski (1988))

If $p \gg n^{-1/m_H}$ *and* $n^2(1-p) \to \infty$, *then* X_H *is asymptotically normal.*

Theorem

If $d_H > d_F$ *for all* $F \subset H$ *(*H *is strictly balanced), and* $p \sim cn^{-1/d_H}$, *then* $X_H \xrightarrow{\mathrm{d}} \mathrm{Po}(\lambda)$ *for some* $\lambda > 0$.

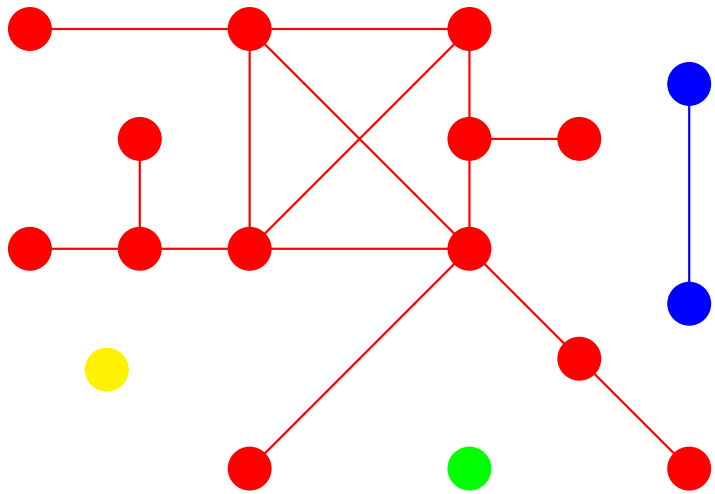
Proof.

$$X_H = \sum_\alpha I_\alpha$$

where α are the copies of H in the complete graph K_n.
The indicators are weakly dependent.
Etc. □

Components

A *component* in a graph is a connected part of the graph.
Example:



This graph has 4 components.
The graph is *connected* if there is only one component.

A typical case is that there is a giant component containing a large part of all nodes, together with many very small components with only one or a few nodes each.

Another case, typical for very sparse graphs, is that there are many small components but no really big one.

This is often interesting in applications. For example, in epidemiology, a component may represent the set of people that will become ill, if one vertex is infected from the outside. Hence a graph with a large component is likely to give large epidemics, but a graph with only small components is safer.

Theorem (Erdős and Rényi)

A classical Erdős–Rényi random graph $G(n, m)$ with n nodes and m edges has a giant component if $m > n/2$ but not otherwise.

More formally: If $n \to \infty$ and $m \sim cn$ for some constant c, and C_1 is the largest component of the random graph, then

$$\frac{|C_1|}{n} \xrightarrow{\mathrm{p}} \begin{cases} 0 & \text{if } c \le 1/2, \\ \rho(2c) > 0 & \text{if } c > 1/2. \end{cases}$$

If $c < 1/2$, then $|C_1| = O_p(\log n)$.
The same holds for $G(n, p)$ with $p \sim c'/n$, with $c' = 2c$ so the threshold is $c' = 1$, i.e. $p = 1/n$.

Theorem (Erdős and Rényi)

A classical Erdős–Rényi random graph $G(n,m)$ with n nodes and m edges is connected if $m > n\log n/2$ but not otherwise.

More precisely: If $n \to \infty$ and $m = n\log n/2 + cn/2 + o(n)$ for some constant c, then

$$P(\text{connected}) \to e^{-e^{-c}}.$$

The number of isolated nodes is asymptotically $\mathrm{Po}(e^{-c})$.

The same holds for $G(n,p)$ with $p = (\log n + c + o(1))/n$.

Branching processes

One useful technique to study Erdős–Rényi graphs and many other random graph models is to start at some node, find first its neighbours, then their neighbours, and so on, until an entire component has been explored.
For some models (for example the Erdős–Rényi graphs), the number of new nodes found each time are (asymptotically) independent and with the same distribution, so this yields a *branching process*, more precisely a *Galton–Watson process*, where each individual gets a random number of children, and these numbers are i.i.d.

Branching processes

One useful technique to study Erdős–Rényi graphs and many other random graph models is to start at some node, find first its neighbours, then their neighbours, and so on, until an entire component has been explored.
For some models (for example the Erdős–Rényi graphs), the number of new nodes found each time are (asymptotically) independent and with the same distribution, so this yields a *branching process*, more precisely a *Galton–Watson process*, where each individual gets a random number of children, and these numbers are i.i.d.

Theorem

A Galton–Watson process has a positive probability of surviving for ever if and only if the expected number of children is > 1.

Branching processes

For an Erdős–Rényi graph $G(n,p)$ the expected number of neigbours is np, so the branching process approximation yields the condition $c = np > 1$.

It follows also that $\rho(c)$ is the survival probability in a Galton–Watson process with offspring distribution $\mathrm{Po}(c)$, which is given by

$$1 - \rho = e^{-c\rho}$$

Branching processes

For an Erdős–Rényi graph $G(n,p)$ the expected number of neigbours is np, so the branching process approximation yields the condition $c = np > 1$.

It follows also that $\rho(c)$ is the survival probability in a Galton–Watson process with offspring distribution $\mathrm{Po}(c)$, which is given by

$$1 - \rho = e^{-c\rho}$$

In some other random graph models, we obtain instead a *multi-type branching process* where individuals may be of different types, with different offspring distributions. (Typically an infinite number of types are needed.)

Susceptibility

The *susceptibility* or *mean cluster size* $\chi(G)$ is the expected size of the component containing a random node. Equivalently, it is n times the probability that two random nodes lie in the same component (and thus may be connected by a path in the graph). If the components are $C_1, C_2, \dots$, then

$$\chi(G) = \frac{\sum_i |C_i|^2}{n}.$$

Susceptibility

The *susceptibility* or *mean cluster size* $\chi(G)$ is the expected size of the component containing a random node. Equivalently, it is n times the probability that two random nodes lie in the same component (and thus may be connected by a path in the graph). If the components are $C_1, C_2, \dots$, then

$$\chi(G) = \frac{\sum_i |C_i|^2}{n}.$$

Theorem

For $G(n,p)$, as $n \to \infty$:

$$\chi(G(n,p)) \sim_p \begin{cases} \frac{1}{1-np}, & 1-np \gg n^{-1/3} \\ n\rho(np)^2, & np-1 \gg n^{-1/3}. \end{cases}$$

Distances and diameter

Given that two nodes are in the same component, we may ask for the *distance* between them, i.e., the shortest path between them in the graph. The maximum distance is the *diameter*. The *average distance* between two random nodes is often at least as interesting.

Distances and diameter

Given that two nodes are in the same component, we may ask for the *distance* between them, i.e., the shortest path between them in the graph. The maximum distance is the *diameter*. The *average distance* between two random nodes is often at least as interesting.

In many graphs, the diameter and average distance are of the order $\log n$, and thus quite small even when the number n of nodes is large. This phenomenon is often called *Small Worlds*.
It says, essentially, that the number of vertices within distance x grows exponentially. (Roughly the same as Expander graphs.)
Example: The giant component of $G(n,p)$ with $p = c/n$, $c > 1$.

Other random graphs

The classical Erdős–Rényi random graphs have node degrees that are random, but with a rather small random dispersion and very small probability of having a degree that is much larger than the average. (The distribution is Binomial or Hypergeometric, and asymptotically Poisson, with an exponentially decreasing tail..)

Power laws

Many graphs from "reality" seem to have node degrees that are distributed according to a *power law*, i.e., there are constants γ and C_1 such that

$$\text{number of nodes with degree } k \approx C_1 k^{-\gamma}$$

or, which is roughly equivalent, with another constant C_2,

$$\text{number of nodes with degree at least } k \approx C_2 k^{-(\gamma-1)}.$$

(Since the graphs are finite (although large), this can of course hold only in some (large) range and not for all k.)
Graphs with a power law are often called *scale-free*.

The degree of a random node has (asymptotically) finite mean
$\iff \gamma > 2$.

The degree of a random node has (asymptotically) finite variance
$\iff \gamma > 3$.

It has during the last decade been popular to study large graphs "in real life" and find such power laws for them, often with a value of γ between 2 and 3.

It has during the last decade been popular to study large graphs "in real life" and find such power laws for them, often with a value of γ between 2 and 3.

As always when something becomes popular, it is easy to overinterpret data, and see power laws also when the evidence really is weak. (*Seek and ye shall find.*)

I'm not an expert on any example, but I think that one should be suspicious.

Random graph models

A classical (Erdős–Rényi) random graph is thus too homogeneous to be a good model in many applications. (In others it is excellent.)

Many other random graph models have been proposed and studied, especially the last 10 years, often with power laws for the node degrees.

This has given new input and new life to the theory of random graph, and has stimulated the mathematical development. (Independently of whether the models are good models for anything or not.)

Some proposed random graph models

1. (Inequality depends on other properties.)
Fix n (number of nodes). Give each node i a number a_i which we call *activity*. Let $A = \sum_{i=1}^{n} a_i$ be the sum of all activities. For each pair i, j of nodes, make a random choice and join them by an edge ij with probability $p_{ij} = \frac{a_i a_j}{A}$.
(Chung and Lu (2002), and several others.)

Some proposed random graph models

1. (Inequality depends on other properties.)
Fix n (number of nodes). Give each node i a number a_i which we call *activity*. Let $A = \sum_{i=1}^{n} a_i$ be the sum of all activities. For each pair i, j of nodes, make a random choice and join them by an edge ij with probability $p_{ij} = \frac{a_i a_j}{A}$.
(Chung and Lu (2002), and several others.)

- If we choose all activities a_i equal, say c, we get a classical random graph $G(n, p)$ with $p = c/n$.

Some proposed random graph models

1. (Inequality depends on other properties.)
Fix n (number of nodes). Give each node i a number a_i which we call *activity*. Let $A = \sum_{i=1}^{n} a_i$ be the sum of all activities. For each pair i, j of nodes, make a random choice and join them by an edge ij with probability $p_{ij} = \frac{a_i a_j}{A}$.
(Chung and Lu (2002), and several others.)

- If we choose all activities a_i equal, say c, we get a classical random graph $G(n, p)$ with $p = c/n$.
- If we choose a_i according to a power law, we get (asymptotically) the same power law for the degrees.

Some proposed random graph models

1. (Inequality depends on other properties.)
Fix n (number of nodes). Give each node i a number a_i which we call *activity*. Let $A = \sum_{i=1}^{n} a_i$ be the sum of all activities. For each pair i, j of nodes, make a random choice and join them by an edge ij with probability $p_{ij} = \frac{a_i a_j}{A}$.
(Chung and Lu (2002), and several others.)

- If we choose all activities a_i equal, say c, we get a classical random graph $G(n, p)$ with $p = c/n$.
- If we choose a_i according to a power law, we get (asymptotically) the same power law for the degrees.
- In general: the degrees of most nodes are close to their activities, so by choosing the activities, we can obtain almost any desired distribution of the node degrees.

Example. Let a_i be proportional to $\sqrt{n}/\sqrt{i}$; this yields $p_{ij} \approx c/\sqrt{ij}$ for some constant $c > 0$. In this case, the node degrees follow a power law with exponent $\gamma = 3$.

Theorem

With high probability, there exist a giant component if $\frac{1}{A}\sum_{i=1}^{n} a_i^2 > 1$, *but not otherwise.*

Example. Let a_i be proportional to $\sqrt{n}/\sqrt{i}$; this yields $p_{ij} \approx c/\sqrt{ij}$ for some constant $c > 0$. In this case, the node degrees follow a power law with exponent $\gamma = 3$.

Theorem

With high probability, there exist a giant component if $\frac{1}{A}\sum_{i=1}^{n} a_i^2 > 1$, *but not otherwise.*

Example. Power law with $\gamma \leq 3$ (as in the example above): since the harmonic series diverges, there is always a giant component even if many edges are removed. (Good for communication network. Bad for epidemics.)

Proof.

Branching process approximation, with the links to the neighbours as individuals in the branching process.

The expected number of children of a link is

$$\sum_i \frac{a_i}{\sum_j a_j} a_i = \frac{\sum_i a_i^2}{\sum_i a_i}.$$

Variants

Variants

- Determine the probabilities p_{ij} by some other formula. Example 2b below is the case $p_{ij} = c/\max(i,j)$. Bollobás, Janson and Riordan (2006) have studied the case $p_{ij} = \kappa(i/n, j/n)/n$ for a given function κ on $[0,1]^2$ (and more generally). ("*Inhomogeneous random graph*")

Theorem (Bollobás, Janson and Riordan)

This random graph has a giant component if and only if the integral operator with kernel κ has norm on $L^2[0,1]$ (or, equivalently, spectral radius) greater than 1.

- Fix the node degrees as a given sequence $d_1, \dots, d_n$, and take a random graph, uniformly among all possible graphs with these degrees.
("*Configuration model*". Bender and Canfield (1978), Bollobás (1985), Molloy and Reed (1995, 1998), and others.)

- Fix the node degrees as a given sequence $d_1, \dots, d_n$, and take a random graph, uniformly among all possible graphs with these degrees.
("*Configuration model*". Bender and Canfield (1978), Bollobás (1985), Molloy and Reed (1995, 1998), and others.)

Theorem (Molloy and Reed)

The random graph with given node degrees $d_1, d_2, \dots, d_n$ has a giant component if and only if $\sum_i d_i(d_i - 2) > 0$.

Proof.

Branching process approximation.

The expected number of children of a link is

$$\sum_i \frac{d_i}{\sum_j d_j}(d_i-1) = \frac{\sum_i d_i(d_i-1)}{\sum_i d_i}$$

and

$$\frac{\sum_i d_i(d_i-1)}{\sum_i d_i} > 1 \iff \sum_i d_i(d_i-2) > 0.$$

□

More examples: (Inequality depends on who is first.)

More examples: (Inequality depends on who is first.)
Let c be a constant.

2a Add new nodes one by one. After adding a node, add c edges connecting it to c randomly chosen old nodes. (c must be an integer.)

More examples: (Inequality depends on who is first.)
Let c be a constant.

2a Add new nodes one by one. After adding a node, add c edges connecting it to c randomly chosen old nodes. (c must be an integer.)

2b Add new nodes one by one. If there already are n nodes, make a random choice for each old node and connect it to the new node with probability c/n. (c can be any positive real number.)

More examples: (Inequality depends on who is first.)
Let c be a constant.

2a Add new nodes one by one. After adding a node, add c edges connecting it to c randomly chosen old nodes. (c must be an integer.)

2b Add new nodes one by one. If there already are n nodes, make a random choice for each old node and connect it to the new node with probability c/n. (c can be any positive real number.)

Theorem

In this case, there will be a giant component if $c > 1/4$, but not otherwise.

2c Add new node one by one. Each time, add with probability c also a new edge, with two endpoints chosen at random among all nodes (old and new). (c must be less than 1.) Known as the *CHKNS model*, after Callaway, Hopcroft, Kleinberg, Newman and Strogatz.

2c Add new node one by one. Each time, add with probability c also a new edge, with two endpoints chosen at random among all nodes (old and new). (c must be less than 1.) Known as the *CHKNS model*, after Callaway, Hopcroft, Kleinberg, Newman and Strogatz.

Theorem

For the CHKNS model, there is a giant component if $c > 1/8$, but not otherwise.

More precisely, if $c = \frac{1}{8} + \epsilon$, the proportion of the nodes belonging to the giant component is (asymptotically)

$$\exp\Bigl(-\frac{\pi}{2\sqrt{2}}\epsilon^{-1/2} + O(\log\epsilon)\Bigr).$$

2c Add new node one by one. Each time, add with probability c also a new edge, with two endpoints chosen at random among all nodes (old and new). (c must be less than 1.) Known as the *CHKNS model*, after Callaway, Hopcroft, Kleinberg, Newman and Strogatz.

Theorem

For the CHKNS model, there is a giant component if $c > 1/8$, but not otherwise.

More precisely, if $c = \frac{1}{8} + \epsilon$, the proportion of the nodes belonging to the giant component is (asymptotically)

$$\exp\Bigl(-\frac{\pi}{2\sqrt{2}}\epsilon^{-1/2} + O(\log\epsilon)\Bigr).$$

These random graphs do not have power laws for the degrees.

More examples: (Inequality depends on chance, and *"For whosoever hath, to him shall be given"*.)

3a Add nodes one by one. Join every node by an edge to one old node; this node is chosen with probability proportional to the number of edges that node already has.

More examples: (Inequality depends on chance, and *"For whosoever hath, to him shall be given"*.)

3a Add nodes one by one. Join every node by an edge to one old node; this node is chosen with probability proportional to the number of edges that node already has.

3b (More generally.) Let c be a positive integer. Add nodes one by one. Join every new node by c edges to c old nodes; these nodes are chosen with probabilities proportional to the number of edges that the nodes already have.

More examples: (Inequality depends on chance, and *"For whosoever hath, to him shall be given"*.)

3a Add nodes one by one. Join every node by an edge to one old node; this node is chosen with probability proportional to the number of edges that node already has.

3b (More generally.) Let c be a positive integer. Add nodes one by one. Join every new node by c edges to c old nodes; these nodes are chosen with probabilities proportional to the number of edges that the nodes already have.

This model is called *preferential attachment*, or the *Barabási–Albert model*.

Theorem

In the Barabási–Albert model, the node degrees have a power law distribution with exponent 3.

Theorem

In the Barabási–Albert model, the node degrees have a power law distribution with exponent 3.

Minor variations of the attachment probabilities yield power laws with other exponents.

Theorem

In the Barabási–Albert model, the node degrees have a power law distribution with exponent 3.

Minor variations of the attachment probabilities yield power laws with other exponents.

Theorem

Let $a > -1$ be fixed and let the probability of attaching a new edge to an old node of degree k be proportional to $k + a$. Then the node degrees have a power law distribution with exponent $3 + a$.

Theorem (Oliveira and Spencer)

Let $\alpha > 1$ be fixed and let the probability of attaching a new edge to an old node of degree k be proportional to k^α. Then only a finite number of nodes will ever have degrees larger than $1/(\alpha - 1)$, so the proportion of such vertices tends to 0.
If $\alpha > 2$, then moreover, after some random time, all new nodes will attach to the same node.

Stochastic processes

Another useful tool is to introduce a time parameter and study suitable stochastic processes that describe, for example, the exploration of the successive neighbourhoods of a vertex.

Standard theorems about convergence of a stochastic process to a deterministic function, or to a Gaussian process, then can be applied.

Example: The existence and size of a giant component in a random graph with given vertex degrees.