Random Graph Models

Bartosz Pankratz

Introduction

- **Random graphs theory** is one of the most interesting research areas on the intersection of network science and probability.
- It provides a better understanding of underlying mechanisms that create networks.
- Random graphs are used to benchmark the network’s algorithms (e.g. clustering algorithms)
- They are also building blocks of synthetic networks that closely resemble real-world networks.

Introduction

- Before we proceed with random models, let us briefly introduce some useful notation.
- While discussing random graphs, we mostly focus on their asymptotical behavior, namely what happens when the number of nodes $n \to \infty$.
- We say that an event in a given probability space holds **asymptotically almost surely** (*a.a.s.*), if its probability tends to one as $n \to \infty$.

Introduction

- For describing the properties of random graphs, we will use **asymptotic notation.** Given two functions $f = f(n)$ and $g = g(n)$:
  - $f(n) = O(g(n))$ if there exists an absolute constant c such that $f(n) \leq c\, g(n)$ for all n.
  - $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$.
  - $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.
  - $f(n) = o(g(n))$ or $f(n) \ll g(n)$ if the limit

    $\lim_{n\to\infty} f(n)/g(n) = 0.$
  - $f(n) = \omega(g(n))$ or $f(n) \gg g(n)$ if $g(n) = o(f(n))$.
  - $f(n) \sim g(n)$ if $f(n) = (1 + o(1))g(n)$.

Binomial Random Graphs

Binomial Random Graphs

- There are three commonly used random graph models:
  - Binomial random graphs
  - Uniform random graphs
  - Erdős-Rényi random graphs
- They are in many cases asymptotically equivalent, thus selection of a specific model depends on its theoretical and computational properties.

Binomial random graphs

The binomial random graph $\mathcal{G}(n, p)$ can be generated by starting with the empty graph on the set of nodes $[n] = \{1, 2, \ldots, n\}$. For each pair of nodes i, j such that$1 \leq i < j \leq n$, we independently introduce an edge $i\ j$ in G with probability p.

Binomial random graphs

The binomial random graph $\mathcal{G}(n,p)$ can be generated by starting with the empty graph on the set of nodes $[n] = \{1, 2, \ldots, n\}$. For each pair of nodes i, j such that $1 \leq i < j \leq n$, we independently introduce an edge $i\,j$ in G with probability p.

- $\mathcal{G}(n,p)$ is defined as the probability distribution over a family of labeled graphs on n nodes.
- There are $\binom{N}{m}$ labeled graphs on n nodes and m edges, where $N = \binom{n}{2}$ is the number of pairs of nodes.
- For a given labeled graph G on n nodes and m edges:

$$\mathbb{P}(G) = p^m(1-p)^{N-m}$$

Uniform Random Graph

Let Ω be the family of all labeled graphs on the set of nodes $[n]$ and exactly m edges, where $0 \leq m \leq N$, $N = \binom{n}{2}$. **The uniform random graph** $\mathcal{G}(n, m)$ assigns to every graph $G \in \Omega$ the same probability, that is,

$$\mathbb{P}(G) = \frac{1}{|\Omega|} = \binom{N}{m}^{-1}$$

Uniform Random Graph

Let Ω be the family of all labeled graphs on the set of nodes $[n]$ and exactly m edges, where $0 \le m \le N$, $N = \binom{n}{2}$. **The uniform random graph** $\mathcal{G}(n,m)$ assigns to every graph $G \in \Omega$ the same probability, that is,

$$\mathbb{P}(G) = \frac{1}{|\Omega|} = \binom{N}{m}^{-1}$$

- $\mathcal{G}(n,p) \approx \mathcal{G}(n,m)$, provided $m \approx \binom{n}{2}p$.

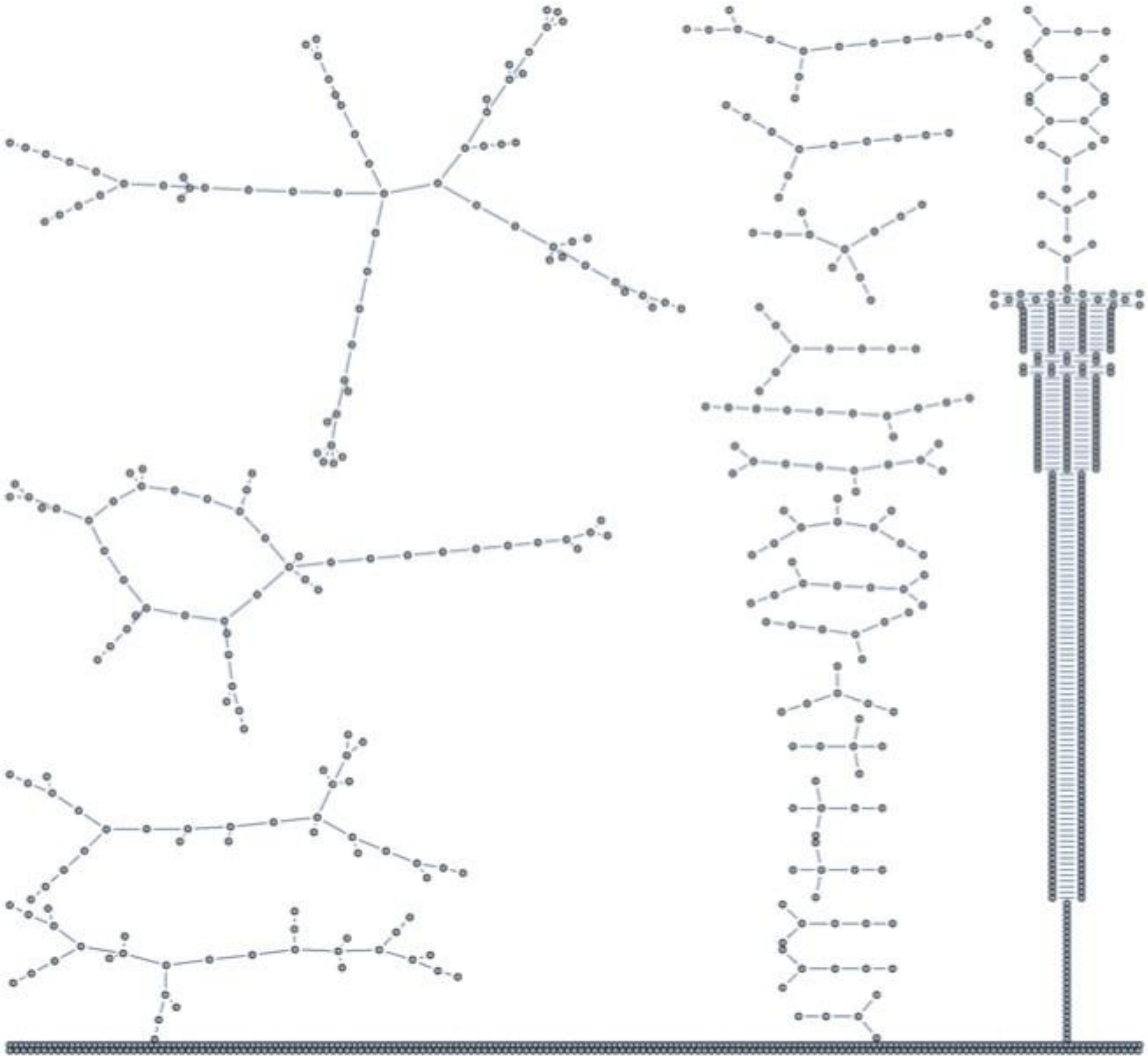
Erdős-Rényi Random Graph

The Erdős-Rényi random graph process is a stochastic process that starts with n labeled nodes and no edges, and at each step adds one new edge chosen uniformly at random from the set of missing edges. Formally, let $N = \binom{n}{2}$ and let $e_1, e_2, \ldots, e_N$ be a random permutation of the edges of the complete graph K_n. The graph process consists of the sequence of random graphs $(\mathcal{G}(n,m))_{m=0}^{N}$, where $\mathcal{G}(n,m) = ([n], E_m)$ and $Em = \{e_1, e_2, \ldots, e_m\}$. It is clear that $\mathcal{G}(n,m)$ is a graph taken uniformly at random from the set of all graphs on n nodes and m edges.

Emergence of the giant component

Subcritical phase:
$\langle k \rangle < 1 - \epsilon$ for some $\epsilon > 0$.

A.a.s. $\mathcal{G}(n, p)$ consists of small trees and unicyclic components; the size of the largest component is $O(\ln n) = o(n)$.



$\mathcal{G}(1000, 0.95/1000)$

Emergence of the giant component

Critical phase: $\langle k \rangle \sim 1$

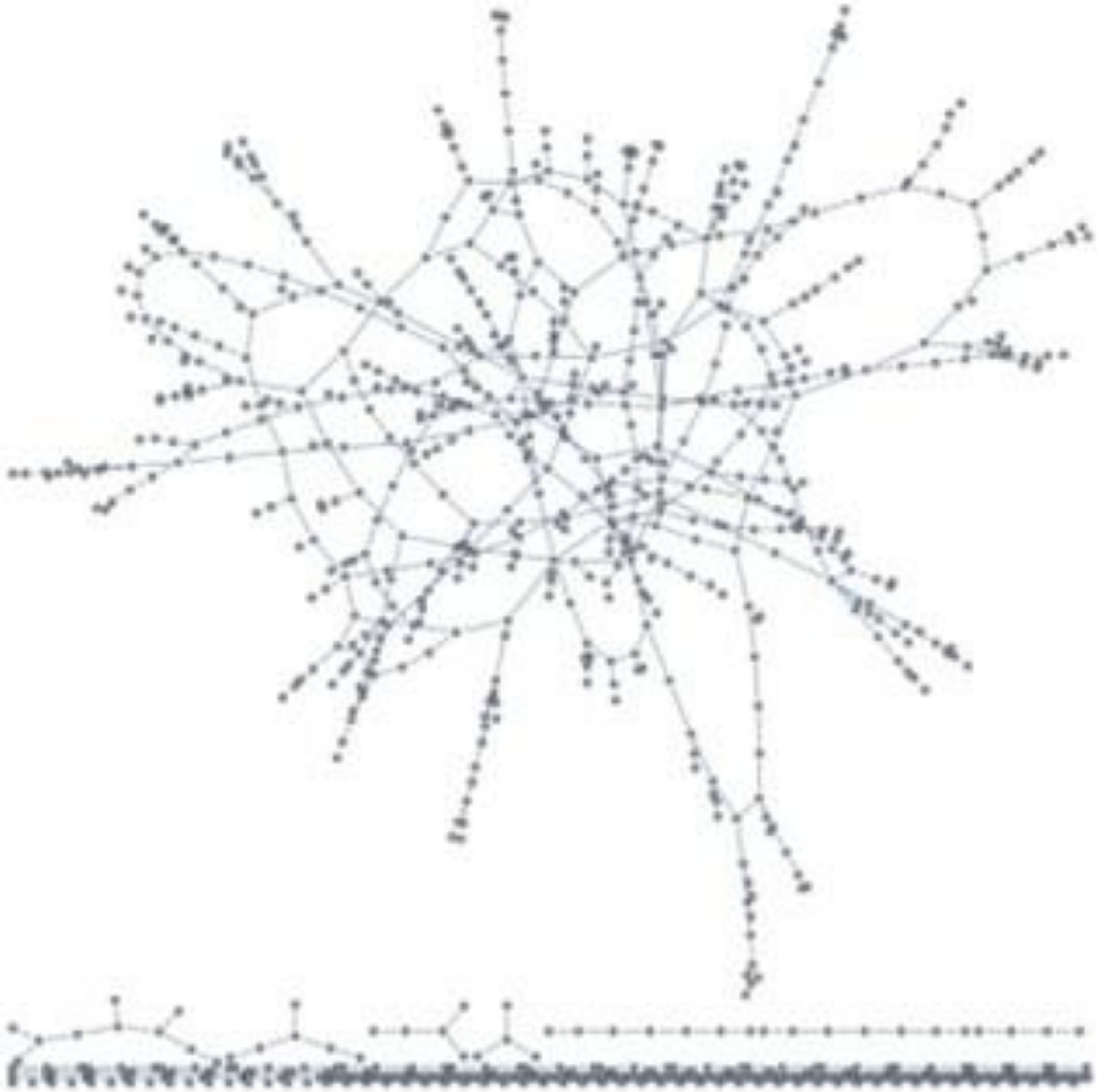
The giant component is formed. During that phase, the size of the largest component keeps growing reaching $\Theta(n^{2/3})$ nodes at precisely $\langle k \rangle = 1$ a.a.s.

Emergence of the giant component

Supercritical phase: $\langle k \rangle > 1 + \epsilon$ for some $\epsilon > 0$.

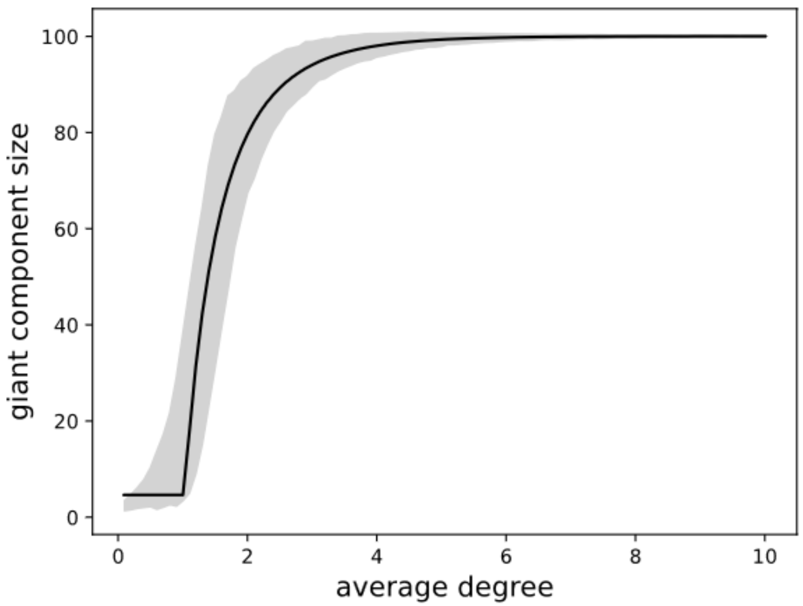
A.a.s. the size of the giant component is $(1 + o(1))\beta n$, where $\beta + e^{-\beta \cdot \langle k \rangle} = 1$.

The giant component is unique and the second-largest component is acyclic or unicyclic and has size $O(\ln n)$.

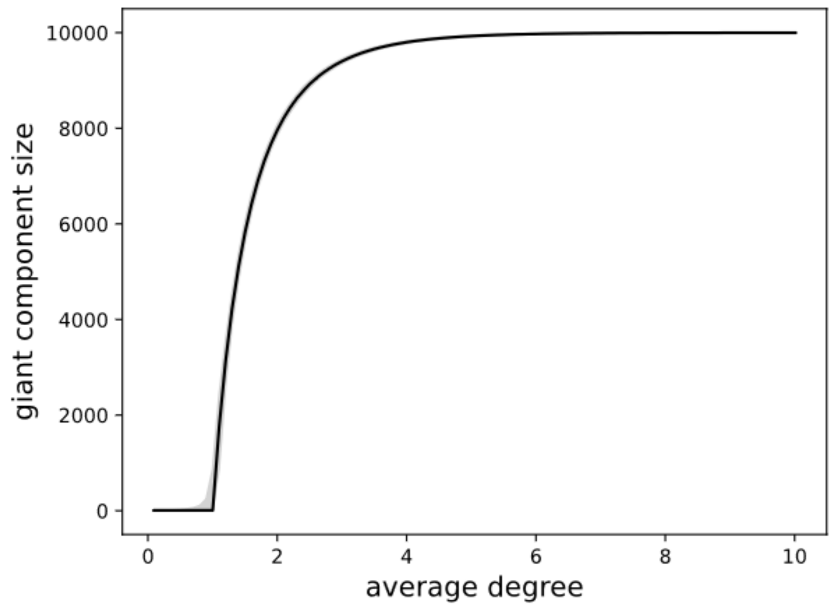


$\mathcal{G}(1000, 1.5/1000)$

Binomial Random Graphs



(a) $n = 100$



(b) $n = 10{,}000$

The order of the giant component: theoretical predictions and empirical results based on 1,000 independent runs for small graphs on (a) n = 100 nodes and larger graphs on (b) n = 10,000 nodes.

Connectivity of Binomial Random Graphs

- Assume that

$$p = p(n) = \frac{\ln n + c}{n} \sim \frac{\ln n}{n} \text{ for some } c \in \mathbb{R}$$

- then the expected number of isolated nodes is approximately equal to:

$$e^{-c}$$

- Clearly:
  - If $c \to -\infty$, then expect isolated nodes.
  - $c \to +\infty$, then expect the opposite.

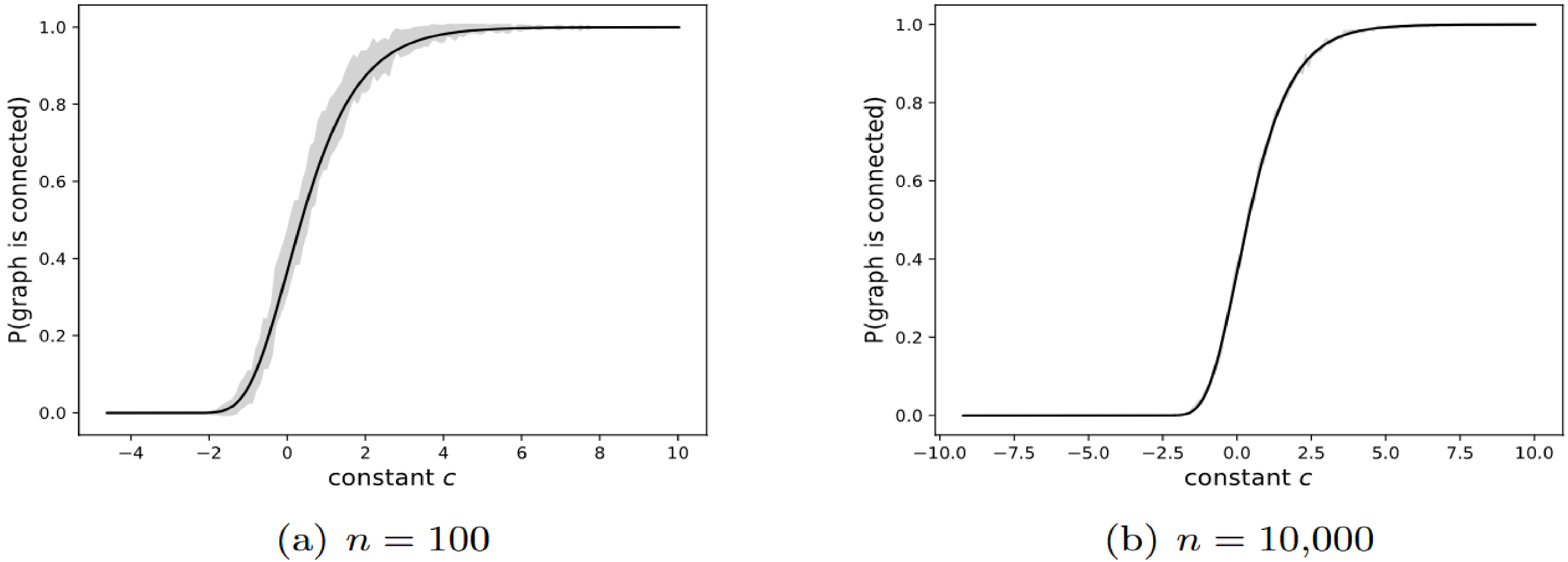
Connectivity of Binomial Random Graphs

- Once isolated nodes disappear, $\mathcal{G}(n,p)$ is connected a.a.s.

$$\mathbb{P}(\mathcal{G}(n,p) \text{ is connected}) \sim \begin{cases} 0 \text{ if } c \to -\infty \\ e^{-c} \text{ if } c \in \mathbb{R} \\ 1 \text{ if } c \to +\infty \end{cases}$$

- For $any\ \epsilon > 0$ we have that:
  - If $pn < (1-\epsilon)\ln n$, then a.a.s. $\mathcal{G}(n,p)$ is **disconnected**.
  - If $pn > (1+\epsilon)\ln n$, then a.a.s. $\mathcal{G}(n,p)$ is **connected**.

Connectivity of Binomial Random Graphs



he probability that $\mathcal{G}(n,p)$) with $np = \ln n + c$ is connected: theoretical predictions and empirical estimations based on 1,000 independent runs for small graphs on (a) n = 100 nodes and larger graphs on (b) n = 10,000 nodes.

Degree Distribution of Binomial Random Graphs

- Consider $\mathcal{G}(n,p)$ with $p = p(n) = c/n$ for some constant $c \in \mathbb{R}+$ and large n. For any node $v \in [n]$,
$$\mathbb{E}[\boldsymbol{deg}(\boldsymbol{v})] = \boldsymbol{p} \cdot (\boldsymbol{n} - \mathbf{1}) \sim \boldsymbol{c}$$
- But $\mathcal{G}(n,p)$ is not a c-regular graph!
- For any $\ell \in \mathbb{N} \cup \{0\}$:
$$\mathbb{P}(\deg(v) = \ell) \sim \frac{c^{\ell}}{\ell!} e^{-c}$$

Degree Distribution of Binomial Random Graphs

- In the limit, the degree distribution of sparse $\mathcal{G}(n,p)$ can be approximated by the Poisson distribution, that is,

$$d_\ell \sim \frac{c^\ell}{\ell!} e^{-c}$$

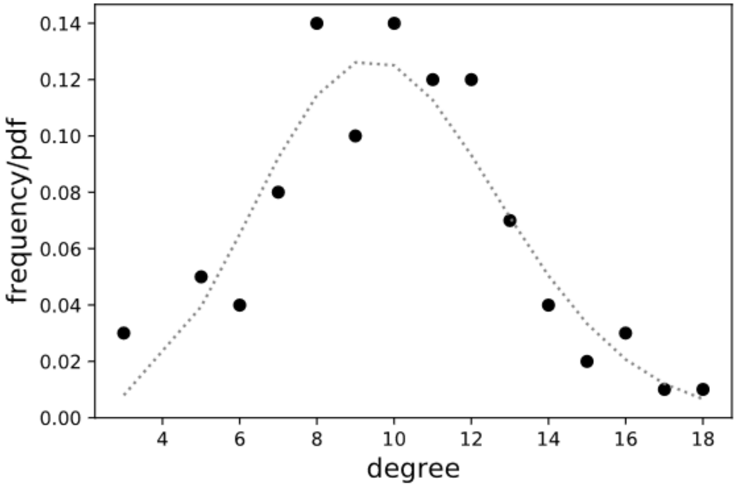
  where c is the asymptotic expected average degree.
- In particular, a.a.s. the maximum degree is $\boldsymbol{O(\ln n / \ln \ln n)}$.

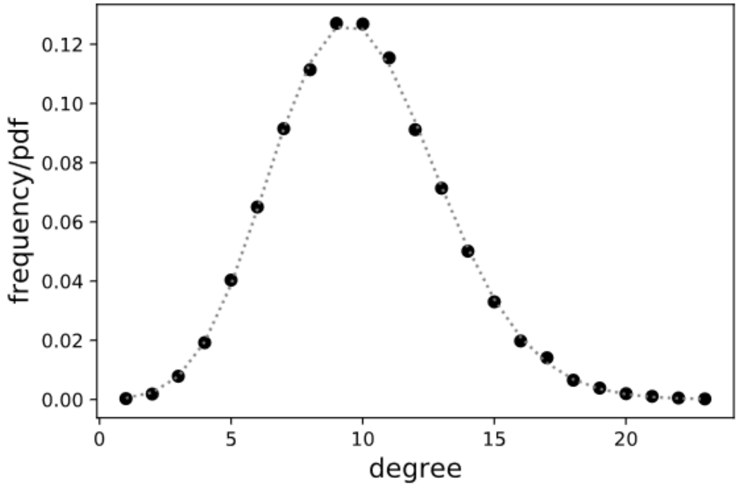
Degree Distribution of Binomial Random Graphs

- In the limit, the degree distribution of sparse $\mathcal{G}(n,p)$ can be approximated by the Poisson distribution, that is,

$$d_\ell \sim \frac{c^\ell}{\ell!} e^{-c}$$

  where c is the asymptotic expected average degree.
- In particular, a.a.s. the maximum degree is $\boldsymbol{O(\ln n / \ln \ln n)}$.





Chung-Lu Model

Power-law distribution

- Real-world networks typically do not have the Poisson distribution.
- Typically, degree distribution follows power law:

$$d_\ell \approx c * \ell^\gamma$$

  where $\gamma > 0$ is a **degree exponent** and $c > 0$ is some normalizing constant.
- Degree distribution can be easily rewritten as:

$$\ln d_\ell \approx -\gamma \ln \ell + \ln c$$

  (straight line with the slope $-\gamma$ on a log-log plot).

Power-law distribution

- To generate a power-law degree distribution with a given degree exponent γ and minimum degree $\delta \geq 1$ we assume:

$$d_\ell \approx c \int_{\ell}^{\ell+1} x^{-\gamma} dx \approx c * \ell^{\gamma}$$

- Note that:

$$1 = \sum_{l=\delta}^{\infty} d_\ell = c \int_{\delta}^{\infty} x^{-\gamma} dx = \frac{c(-\delta^{1-\gamma})}{1-\gamma}$$

- Finally:

$$d_\ell \approx (\gamma - 1)\delta^{\gamma-1}\ell^{-\gamma}$$

Power-law distribution

- **Average degree**:

$$\langle k \rangle = \frac{\gamma - 1}{\gamma - 2}\delta$$

  provided that $\gamma > 2$.

- **Maximum degree**:
  - Assuming that the number of nodes of degree at least Δ is close to 1 we get that:

$$\Delta = \delta n^{1/(\gamma-1)}$$

  which is often referred to as the **natural cut-off** of the graph.

Chung-Lu Model

Let $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ be any vector of n positive real numbers and let $W = \sum_{i=1}^{n} w_i$. Random graph $G(\boldsymbol{w})$ is generated as follows:

1. Each pair of nodes i, j such that $1 \leq i \leq j \leq n$ is independently sampled as an edge (or loop if $i = j$) with probability given by:

$$p_{i,j} \begin{cases} \dfrac{w_i w_j}{W} \text{ for } i \neq j \\ \dfrac{w_i^2}{W} \text{ for } i = j \end{cases}$$

Chung-Lu Model

- For any any $i \in [n]$:
$$\mathbb{E}[\deg(i)] = w_i$$
- The model is also well-studied but the behavior and results are more complex; for example, for any $\epsilon > 0$:
  - if $\langle k \rangle \leq \langle k^2 \rangle / \langle k \rangle < 1 - \epsilon$, then a.a.s. $G(\boldsymbol{w})$ has no giant,
  - if $\langle k^2 \rangle / \langle k \rangle \geq \langle k \rangle > 1 + \epsilon$, then a.a.s. there is one.

Chung-Lu Model

- Generating power-law graphs with degree exponent γ:

$$w_i = c \cdot (i + i_0 - 1)^{-1/(\gamma-1)}$$

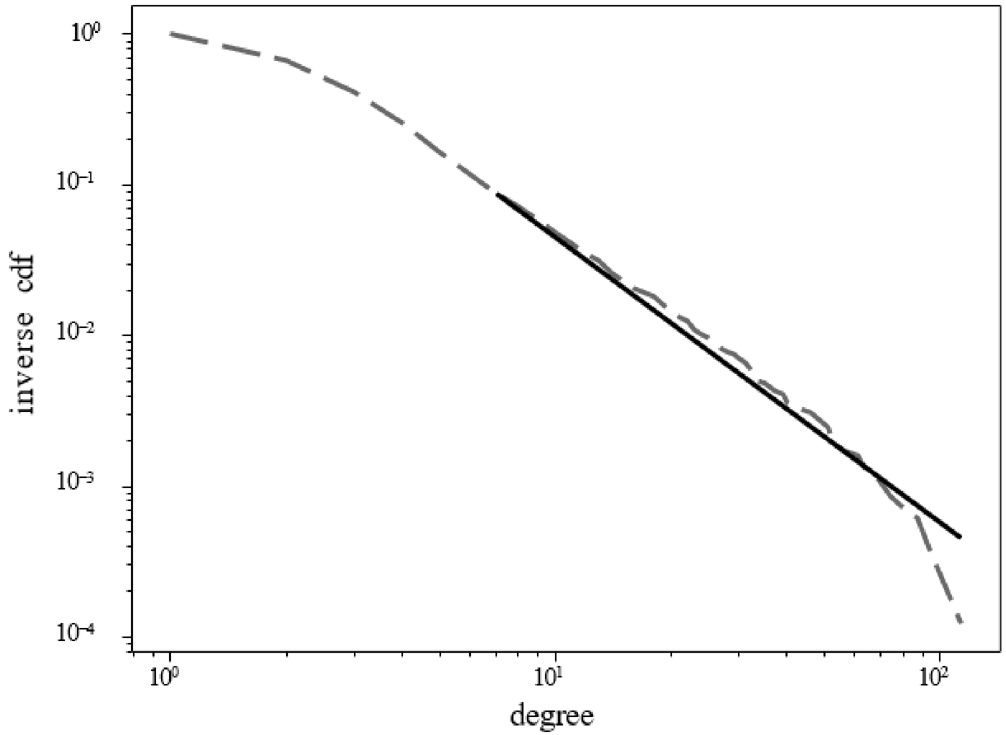
$c = c(n)$ depends on the minimum (or average) degree $\delta \geq 1$ and $i_0 = i_0(n)$ depends on the maximum degree Δ.

It follows that $c = \delta n^{1/(\gamma-1)}$ and $i_0 = n/\left(\frac{\Delta}{\delta}\right)^{\gamma-1}$ so

$$w_i = \delta\left(\frac{n}{i - 1 + n/(\Delta/\delta)^{\gamma-1}}\right)^{1/(\gamma-1)}$$

- It is possible to show that the expected number of nodes of degree k is proportional to $\Gamma(k - \gamma + 1)/\Gamma(k + 1) \approx k^{-\gamma}$, where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$ is **the gamma function**.

Chung-Lu Model



$\mathcal{G}(w)$ generated on $n = 10{,}000$ nodes using the set of weights with $\gamma = 2.5, \delta = 1$, and $\Delta = \sqrt{n} = 100$.

We got $\delta' = 0$ and $\Delta' = 113$, respectively.

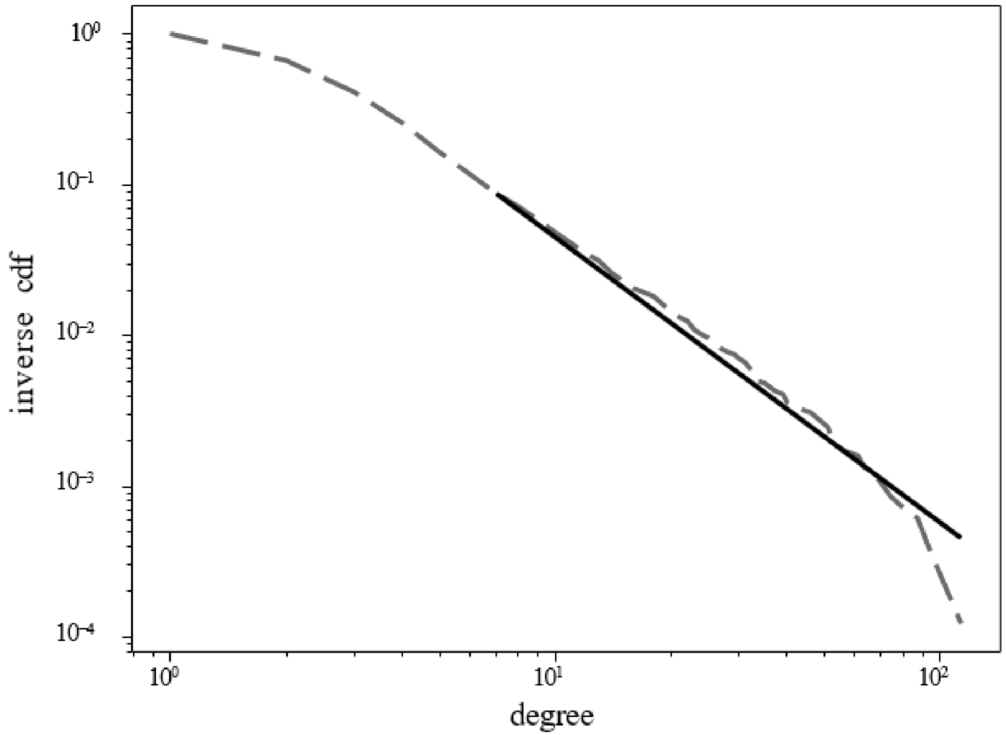
But did we preserve the degree exponent?

Chung-Lu Model

- **Kolmogorov–Smirnov statistic**: focus only on the large degrees. For a given cutoff for small degrees $\ell \in [\max\{\delta, 1\}, \Delta]$,
$$\gamma^{\ell} = 1 + \frac{|\{j: \deg(j) \geq \ell|}{\sum_{j:\deg(j)\geq\ell} \ln\left(\frac{\deg(j)}{\ell - 1/2}\right)}$$

- The **divergence** of the experimental distribution from the theoretical one is defined as:
$$D_{\ell} = \max_{k\in[\ell,\Delta]} \left| \frac{|\{j: \deg(j) \geq k|}{|\{j: \deg(j) \geq \ell|} - \frac{\int_k^{\infty} x^{-\gamma^{\ell}} dx}{\int_{\ell}^{\infty} x^{-\gamma^{\ell}} dx} \right|$$
- The value of γ_{ℓ} that minimizes D_{ℓ} (over all $\ell \in [\max\{\delta, 1\}, \Delta]$) is used as an estimate γ' of the power-law exponent.

Chung-Lu Model



$\mathcal{G}(w)$ generated on $n = 10{,}000$ nodes using the set of weights with $\gamma = 2.5, \delta = 1$, and $\Delta = \sqrt{n} = 100$.

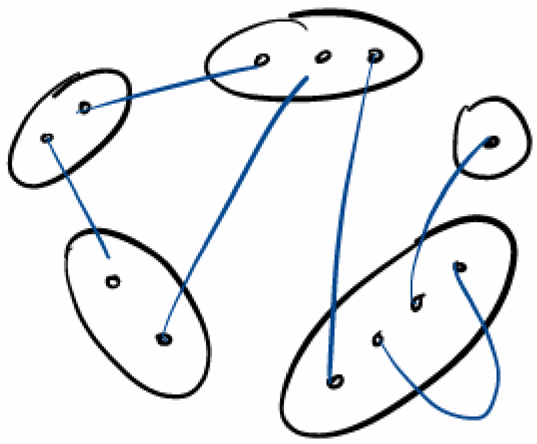
We got $\delta' = 0$ and $\Delta' = 113$, respectively.

The fitted line has slope of -1.89 ($\gamma' = 2.89$), which was obtained with the procedure we described, with $\ell' = 7$.

Configuration Model

Configuration Model

- **Chung-Lu model** returns a Graph with a degree sequence approximately close to the sequence $\boldsymbol{w}$.
- **Configuration model** $\mathcal{P}_{n,d}$ generates a random graph that strictly follows a given, graphic degree sequence $\boldsymbol{d} = (\deg(1), \deg(2), \ldots, \deg(n))$



Random d-regular graphs

Fix $d \in \mathbb{N} \cup \{0\}$. Let Ω be the family of all labelled graphs on the set of nodes $[n]$ that are d-regular. The random d-regular graph, denoted by $\mathcal{G}_{n,d}$, assigns to every graph $G \in \Omega$ the same probability, that is,

$$\mathbb{P}(G) = \frac{1}{|\Omega|}$$

- Since the total volume of any graph is even, n has to be even if d is odd.
- Different mathematical tools are required when $d = d(n)$ grows together with n.
- Generating all d-regular graphs on n nodes is impossible.

Configuration/pairing model

Consider dn points partitioned into n labelled buckets $v_1, v_2, \ldots, v_n$ of d points each.

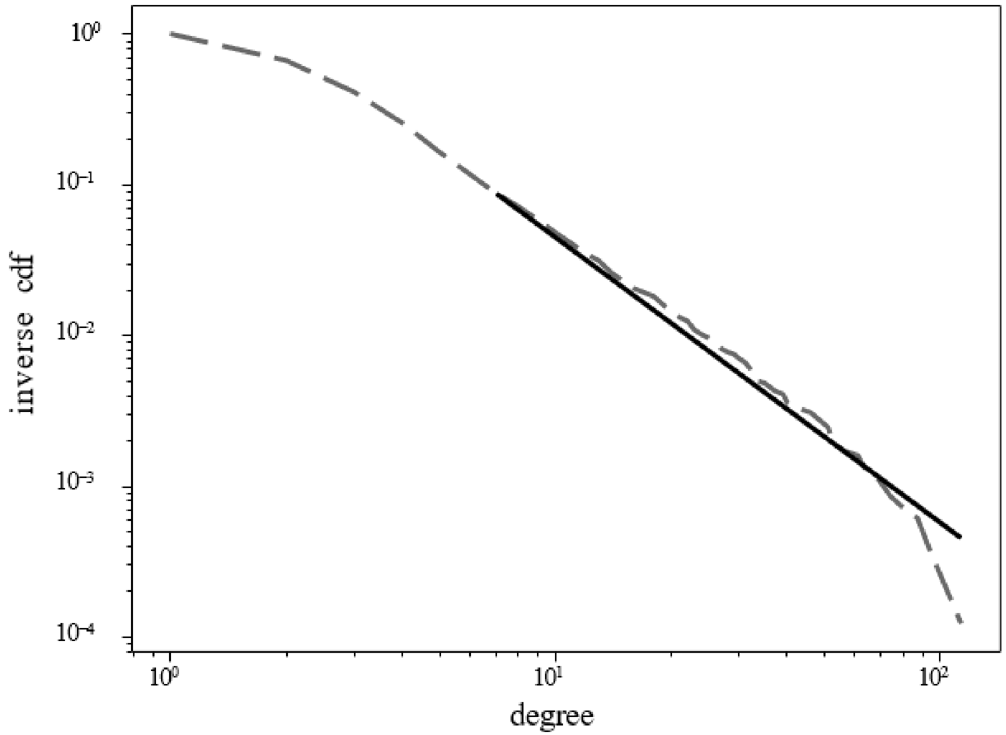
A **pairing** of these points is a perfect matching into $dn/2$ pairs.

Given a random pairing P, we may construct a multigraph $\mathcal{P}_{\mathrm{n,d}} = \mathcal{P}(P)$, with loops and parallel edges allowed, after contracting buckets into nodes.

Configuration Model

- The restriction of $\mathcal{P}_{n,d}$ to simple graphs is precisely $\mathcal{G}_{n,d}$.
- Practical implication: keep generating $\mathcal{P}_{n,d}$ (independently) and stop when you get a simple graph; you get $\mathcal{G}_{n,d}$.
- Random pairing generates a simple graph with probability asymptotic to $e^{-(d^2-1)/4}$ depending on d but not on n.
- Practical implication: the expected number of attempts is $e^{(d^2-1)/4}$ (large for large d but reasonable for small values).
- Theoretical implication: any event holding a.a.s. in $\mathcal{P}_{n,d}$ also holds a.a.s. in $\mathcal{G}_{n,d}$.

Configuration Model



$\mathcal{P}_{n,d}$ is simple with probability $\sim e^{-(d^2-1)/4}$

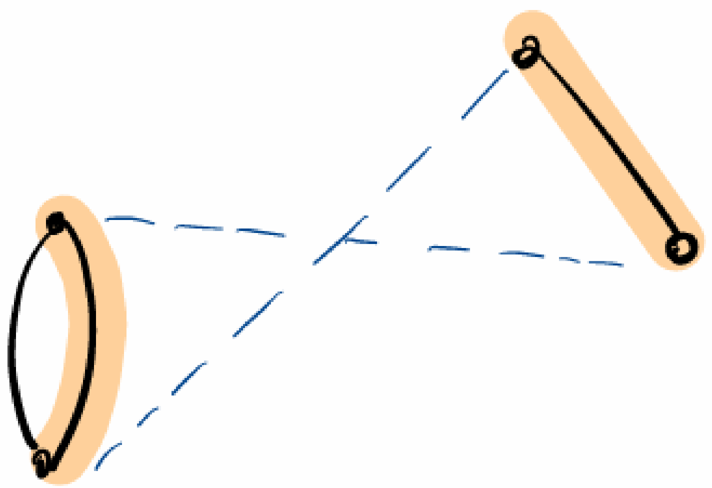
Theoretical predictions and empirical results for $n = 100$ and $n = 10{,}000$.

Configuration Model

- Generating $\mathcal{P}_{n,d}$ is fast but the graph might not be simple!
- Depending on our application, we may:
  - stay with multigraphs;
  - remove all potential loops and parallel edges (**erased configuration model**); the graph might not match exactly the given sequence $\boldsymbol{d}$.
  - resample until we get $\mathcal{G}_{n,d}$ or avoid creating loops/parallel edges and re-start if we get stuck; slow.
  - do **switching**; preserves distribution, typically fast, asymptotically equivalent to $\mathcal{G}_{n,d}$.

Configuration Model

- do **switching**; preserves distribution, typically fast, asymptotically equivalent to $\mathcal{G}_{n,d}$.



Random Geometric Graphs

Random Geometric Graphs

- Often nodes are described by some properties – they might impact the existence/strength of an edge.
- While generating a random graph that tries to mimic the real world network, we want to utilize this fact.
- **Random Geometric Graphs** are a category of models that uses the properties (*metadata*) to generate a network.

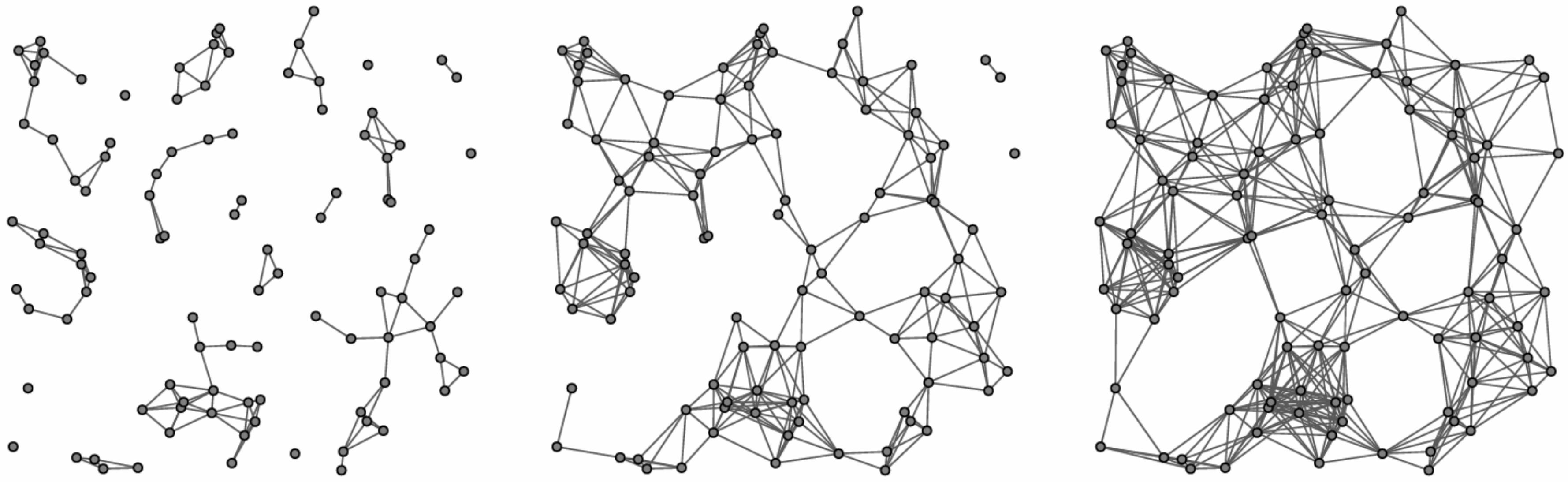
Random Geometric Graphs

Let $r \in \mathbb{R} + \cup \{0\}$. The random geometric graph $\mathcal{RGG}(n, r)$ can be generated by starting with the empty graph on n nodes, $v_1, v_2, \dots, v_n$, that are randomly sampled from the uniform distribution of the underlying space $[0,1]^d$. Each pair of nodes v_i, v_j such that $1 \leq i < j \leq n$ is connected by an edge if and only if $d(v_i, v_j) \leq r$.

- The choice of metric space is important (especially, if d is large)
- For continuous (especially spatial) data one might use the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

Random Geometric Graphs



Instances of $RGG(100, r)$ for $r \in \{0.1, 0.15, 0.2\}$ on the unit square ($d = 2$).
Threshold for connectivity is well understood but, for example, the appearance of the giant component is not!