Random graphs II

13.12.2018
Recap: Random graphs

- Consider ensembles of networks: *Representation matters*
  - (Unlabeled) graphs
  - Adjacency matrix (labeled graphs)

- Random graph models:
  - Basic structure of graphs, e.g. average number of edges
  - Graphs otherwise created or grown at random

- Erdős-Rényi model $G(n, p)$ where every simple graph $G$ with $m$ edges has probability

$$P(G) = p^m(1 - p)^{\frac{1}{2}n(n-1)-m}$$

Basic properties:
- Average number of edges $\langle m \rangle = \frac{1}{2}n(n-1)p$
- Mean degree $\langle k \rangle = (n-1)p$
- Binomial degree distribution
Giant component of $G(n, p)$

*Giant component*: A connected component which grows in proportion to the number of vertices $n$, i.e. size $O(n)$.

- There is a giant component if and only if $\langle k \rangle > 1$
- Large $n \to \infty$ limit: The fraction $S$ of vertices in the giant component is then given by the non-zero solution of

$$S = 1 - e^{-\langle k \rangle S}$$
Diameter

What is the average diameter of random graphs from $G(n, p)$?

Here, we will show that it grows logarithmically with $n$, i.e. $O(\ln n)$.
Diameter

Consider two vertices \( i \) and \( j \)

- Average number of vertices with distance at most \( s \) and \( t \) from \( i \) and \( j \) respectively
  
  \[
  \frac{1 - \langle k \rangle^{s+1}}{1 - \langle k \rangle} \in O(\langle k \rangle^{s+1}) \quad \text{and} \quad \in O(\langle k \rangle^{t+1})
  \]

- Distance \( d_{ij} \) between \( i \) and \( j \) is greater than \( l = s + t + 1 \) if and only if there is no edge between neighbourhood sets around \( i \) and \( j \)

- Thus,

  \[
  P(d_{ij} > l) \approx (1 - p)^{\langle k \rangle^{s+t+2}} = (1 - p)^{\langle k \rangle^{l+1}}
  \]

  \[
  = (1 - \frac{\langle k \rangle}{n-1})^{\langle k \rangle^{l+1}}
  \]

  \[
  \ln P(d_{ij} > l) = \langle k \rangle^{l+1} \ln(1 - \frac{\langle k \rangle}{n-1})
  \]

  \[
  \approx -\frac{\langle k \rangle^{l+2}}{n}
  \]
Diameter

\[ P(d_{ij} > l) = e^{-\frac{\langle k \rangle^{l+2}}{n}} \]

Diameter is the smallest distance \( l \) such that \( P(d_{ij} > l) = 0 \).

- Requires that \( \langle k \rangle^{l+2} \) grows faster than \( n \), i.e.
  \[ \langle k \rangle^{l+2} = an^{1+\epsilon} \]
  where \( \epsilon \to 0 \) as we look for the smallest \( l \)

- Taking logarithms, we find
  
  \[ l = \frac{\ln a}{\ln \langle k \rangle} + \lim_{\epsilon \to 0^+} \frac{(1 + \epsilon) \ln n}{\ln \langle k \rangle} - 2 = \text{const} + \frac{\ln n}{\ln \langle k \rangle} \]

Small world effect: Random graphs tend to have small diameters!
Problems of $G(n, p)$

$G(n, p)$ does not capture important properties of real-world networks

- Degree distribution
  $G(n, p)$ generates a binomial degree distribution, but often power-laws are found

- Ex.: Centrality measures are highly correlated
The *configuration model* is an extension of $G(n, m)$ which starts from a fixed degree sequence $\{k_i\}$:

- From vertex $i$ we have $k_i$ edge "stubs" originating.
  With $m$ edges there are $\sum_i k_i = 2m$ edge stubs in total.
- Choose two stubs uniformly at random and connect them by an edge.
- Continue matching of remaining stubs until all of them are connected.
Configuration model

Catch with configuration model:

- Created network can contain self-loops and multi-edges
- Distribution is uniform on all stub matchings:
  - Many matchings give the same graph:

\[
N(\{k_i\}) = \prod_i k_i!
\]

permutations of stub matchings give rise to same network (without self-loops and multi-edges)

- Additional symmetries reduce factor for self-loops \(A_{ii} > 0\) and multi-edges \(A_{ij} > 1\):

\[
N(A_{ij}) = \frac{\prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!!}
\]

where \(n!! = n(n-2)\cdots2\) for \(n\) even
Configuration model

Catch with configuration model:

- Created network can contain self-loops and multi-edges
- Distribution is uniform on all stub matchings:
  Fortunately
  - fraction of self-loops and multi-edges (often) vanishes for large networks
  - each network is overcounted by a constant factor $N \approx \prod_i k_i!$
Configuration model

Properties of the configuration model:

- Probability of an edge between $i$ and $j$:
  \[ p_{ij} = \frac{k_i k_j}{2m - 1} \]

- Average number of multi-edges:
  \[ \frac{1}{2} \sum_{ij} \frac{k_i k_j (k_i - 1)(k_j - 1)}{(2m)^2} = \frac{1}{2} \left[ \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \right]^2 \]

where $\langle k \rangle = \frac{1}{n} \sum_i k_i$ and $\langle k^2 \rangle = \frac{1}{n} \sum_i k_i^2$

- Average number of self-loops:
  \[ \sum_i \frac{k_i (k_i - 1)}{4m} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} \]
Why not extend $G(n, p)$?

In the configuration model, the number of edges is fixed:

$$m = \frac{1}{2} \sum_i k_i$$

Choose an edge between $i$ and $j$ with probability

$$p_{ij} = \begin{cases} 
\frac{c_i c_j}{2m} & \text{for } i \neq j \\
\frac{c_i^2}{4m} & \text{for } i = j 
\end{cases}$$

with parameters $(c_i)_{i=1}^n$ and $m = \frac{1}{2} \sum_i c_i$
Configuration model

Choose an edge between \( i \) and \( j \) with probability

\[
p_{ij} = \begin{cases} 
\frac{c_i c_j}{2m} & \text{for } i \neq j \\
\frac{c_i^2}{4m} & \text{for } i = j 
\end{cases}
\]

Then,

- the expected number of edges is given as

\[
\sum_{i \leq j} p_{ij} = \sum_{i < j} \frac{c_i c_j}{2m} + \sum_{i} \frac{c_i^2}{4m} = \sum_{ij} \frac{c_i c_j}{4m} = m
\]

- the expected degree of vertex \( i \) is given as

\[
\langle k_i \rangle = 2p_{ii} + \sum_{j \neq i} p_{ij} = \sum_{j} \frac{c_i c_j}{2m} = c_i
\]
Configuration model

Choose an edge between \( i \) and \( j \) with probability

\[
p_{ij} = \begin{cases} 
\frac{c_i c_j}{2m} & \text{for } i \neq j \\
\frac{c_i^2}{4m} & \text{for } i = j
\end{cases}
\]

Then,

- the expected number of edges is given as

\[
\sum_{i \leq j} p_{ij} = \sum_{i < j} \frac{c_i c_j}{2m} + \sum_i \frac{c_i^2}{4m} = \sum_{ij} \frac{c_i c_j}{4m} = m
\]

- the expected degree of vertex \( i \) is given as

\[
\langle k_i \rangle = 2p_{ii} + \sum_{j \neq i} p_{ij} = \sum_j \frac{c_i c_j}{2m} = c_i
\]

Main shortcoming:

Degree distribution of actual degrees \( k_i \) does not equal the distribution of average degrees \( c_i \)!
As in the Erdős-Rényi model, we want to compute the size of the giant component for the configuration model.

Required tools:

- Excess degree distribution
- Generating functions
Excess degree distribution

Consider a network with degree distribution \((p_k)_{k=0}^{\infty}\), i.e. a randomly picked vertex has degree \(k\) with probability \(p_k\).

Now, pick a vertex at random and follow one of its edges. What is the probability that the vertex reached has degree \(k\)?
Consider a network with degree distribution \((p_k)_{k=0}^{\infty}\), i.e. a randomly picked vertex has degree \(k\) with probability \(p_k\).

Now, pick a vertex at random and follow one of its edges. What is the probability that the vertex reached has degree \(k\)?

In general, this will be different from \(p_k\):
Consider the case \(k = 0\) and \(p_0 > 0\), but we can never reach a vertex of degree 0. Thus, the above probability is \(0 \neq p_0\).
Consider an edge in the configuration model starting from some vertex $i$

It has an equal probability to end at any of the $2m - 1$ other edge “stubs”

Thus, the probability that it connects to a vertex of degree $k$ is found to be

$$\frac{k}{2m - 1} \cdot np_k \approx \frac{k}{\langle k \rangle}$$

since $2m = n\langle k \rangle$.
Excess degree distribution

\[ \frac{k}{2m-1} \cdot n p_k \approx \frac{k p_k}{\langle k \rangle} \]

From this, we find the average degree of a neighbour:

\[ \sum_k k \frac{k p_k}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle} \]

Usually this is larger than \( \langle k \rangle \):

\[ \frac{\langle k^2 \rangle}{\langle k \rangle} - \langle k \rangle = \frac{1}{\langle k \rangle} (\langle k^2 \rangle - \langle k \rangle^2) = \frac{\sigma_k^2}{\langle k \rangle} \geq 0 \]

*Your friends have more friends than you!*
Excess degree distribution

Formally, we define the *excess degree* of a vertex (reached from some neighbour) as its degree minus one (accounting for the edge reaching it).

The excess degree distribution $q_k$ is then given as

$$q_k = \frac{(k + 1)p_{k+1}}{\langle k \rangle}$$
The *probability generating function* for the probability distribution $p_k$ is defined as the polynomial

$$g(z) = p_0 + p_1 z + p_2 z^2 + \ldots = \sum_{k=0}^{\infty} p_k z^k$$

It contains the same information as the probability distribution:

$$p_k = \frac{1}{k!} \frac{d^k g}{dz^k} \bigg|_{z=0}$$
Generating function

Properties:
  ▶ Normalization:

\[ g(1) = \sum_{k=0}^{\infty} p_k = 1 \]

▶ Moments:

▶ Summation:
Generating function

Properties:

- **Normalization:**
  \[ g(1) = \sum_{k=0}^{\infty} p_k = 1 \]

- **Moments:**
  \[ g'(z) = \sum_{k=0}^{\infty} p_k k z^{k-1} \]
  and therefore
  \[ g'(1) = \sum_{k=0}^{\infty} k p_k = \langle k \rangle \]

  Similarly,
  \[ \langle k^2 \rangle = (z \frac{d}{dz})^2 g(z)|_{z=1} \]

- **Summation:**
Generating function

Properties:

- **Normalization:**
  \[
g(1) = \sum_{k=0}^{\infty} p_k = 1
\]

- **Moments:**
  \[
g'(1) = \sum_{k=0}^{\infty} kp_k = \langle k \rangle
\]

- **Summation:**
  If \( s = \sum_{i=1}^{m} k_i \) is the sum of \( m \) random numbers \( k_i \) drawn independently from \( p_k \) with generating function \( g(z) \), then the generating function of the probability distribution of \( s \) is
  \[
h(z) = [g(z)]^m
\]
Generating function

Show that $s = \sum_{i=1}^{m} k_i$ has generating function $h(z) = [g(z)]^m$

Given that all $k_i$’s are drawn independently, the probability $q_s$ of $s$ can be computed as the sum of all sets $\{k_i\}$ with $\sum_i k_i = s$:

$$q_s = \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \prod_{s=\sum_i k_i}^{m} p_{k_i}$$

where any set $\{k_i\}$ has probability $\prod_i p_{k_i}$. 
Generating function

The generating function of $q_s$ is then found to be

$$h(z) = \sum_{s=0}^{\infty} q_s z^s$$

$$= \sum_{s=0}^{\infty} z^s \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \prod_{i=1}^{m} p_{k_i}$$

$$= \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} z^{\sum_i k_i} \prod_{i=1}^{m} p_{k_i}$$

$$= \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \prod_{i=1}^{m} p_{k_i} z^{k_i}$$

$$= \left[ \sum_{k=0}^{\infty} p_k z_k \right]^m = [g(z)]^m$$
Power-law distribution

A probability distribution \( p_k \) on the natural numbers \( k \in \mathbb{N} \) is called *power-law distribution* if

\[
p_k \propto k^{-\alpha}
\]

with (tail) exponent \( \alpha > 0 \).
A probability distribution $p_k$ on the natural numbers $k \in \mathbb{N}$ is called *power-law distribution* if

$$p_k \propto k^{-\alpha}$$

with (tail) exponent $\alpha > 0$

Normalization requires that $\sum_{k=0}^{\infty} p_k = 1$ and thus

$$p_k = \begin{cases} 
0 & \text{for } k = 0 \\
\frac{k^{-\alpha}}{\zeta(\alpha)} & \text{for } k > 0
\end{cases}$$

where $\zeta(\alpha)$ is the Riemann zeta function defined as

$$\zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}$$

Note: A power-law distribution has *heavy tails* as the probability decays much slower than exponential for $k \to \infty$. 
Generating function

Examples:

- Binomial distribution \( p_k = \binom{n}{k} p^k (1 - p)^{n-k} \)

\[
g(z) = (1 - p + pz)^n
\]

- Power-law distribution \( p_k = \frac{k^{-\alpha}}{\zeta(\alpha)} \)

\[
g(z) = \sum_{k=0}^{\infty} p_k z^k = \frac{\text{Li}_\alpha}{\zeta(\alpha)}
\]

where \( \text{Li}_\alpha = \sum_{k=1}^{\infty} k^{-\alpha} z^k \) is the polylogarithm of \( z \).

\[
\frac{d}{dz} \text{Li}_\alpha = \frac{d}{dz} \sum_{k=1}^{\infty} k^{-\alpha} z^k = \sum_{k=1}^{\infty} k^{-(\alpha-1)} z^{k-1} = \frac{\text{Li}_{\alpha-1}(z)}{z}
\]
Giant component

To compute the size of the giant component in the configuration model, we first need two definitions of generating function

▶ for the degree distribution

\[ g_0(z) = \sum_{k=0}^{\infty} p_k z^k \]

▶ and the excess degree distribution

\[ g_1(z) = \sum_{k=0}^{\infty} q_k z^k \]

Note:

\[ g_1(z) = \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} (k + 1)p_{k+1} z^k = \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} kp_k z^{k-1} = \frac{g'_0(z)}{g'_0(1)} \]
Giant component

First, we assume that no giant component forms and compute the distribution $\rho_s$ of component sizes that vertices reached via an edge belong to:

\[
\begin{align*}
\text{[Diagram: Tree structures]} & = \text{[Diagram: Tree structures]} + \text{[Diagram: Tree structures]} + \text{[Diagram: Tree structures]} + \text{[Diagram: Tree structures]} + \cdots \\
\end{align*}
\]

**Note:** We can assume that clusters are trees and therefore the sizes of unconnected clusters are independent.
Giant component

First, we assume that no giant component forms and compute the distribution $\rho_s$ of component sizes that vertices reached via an edge belong to:

\[ \begin{array}{c}
\boxed{\text{square}} = \boxed{\text{circle}} + \boxed{\text{circle}} + \boxed{\text{circle}} + \boxed{\text{circle}} + \cdots
\end{array} \]

Its generating function

\[ h_1(z) = \sum_{s=1}^{\infty} \rho_s z^s \]

solves

\[ h_1(z) = z \sum_{k=0}^{\infty} q_k [h_1(z)]^k = zg_1(h_1(z)) \]
Giant component

The generating function for the size of a small component that a randomly chosen vertex belongs to is then computed as

\[ h_0(z) = z \sum_{k=0}^{\infty} p_k[h_1(z)]^k = zg_0(h_1(z)) \]

From this, we can compute the mean component size

\[ \langle s \rangle = h'_0(1) = [g_0(h_1(z)) + zg_0'(h_1(z))h'_1(z)]_{z=1} = 1 + g'_0(1)h'_1(z) = 1 + \frac{g'_0(1)}{1 - g'_1(1)} \]

*Phase transition:* The component size diverges at \( g'_1(1) = 1 \)
The giant component is not tree-like, so the approximation used to compute the component size distribution breaks down. Instead, we note that now, all non-giant components are tree-like:

\[ h_0(1) = \text{Fraction of vertices not in giant component} \]

Thus, the size of the giant component \( S \) solves

\[ S = 1 - g_0(u) \text{ and } u = g_1(u) \]
Example: Consider a network with degree at most 3. Then,

\[
g_0(z) = p_0 + p_1 z + p_2 z^2 + p_3 z^3
\]

\[
g_1(z) = \frac{g_0'(z)}{g_0'(1)} = \frac{p_1 + 2p_2 z + 3p_3 z^2}{p_1 + 2p_2 + 3p_3}
\]

\[
= q_0 + q_1 z + q_2 z^2
\]

Thus, \( u = q_0 + q_1 u + q_2 u^2 \) solves a quadratic equation:

\[
u = \frac{1 - q_1 \pm \sqrt{(1 - q_1)^2 - 4q_0 q_2}}{2q_2}
\]

Using \( 1 - q_1 = q_0 + q_2 \), we obtain

\[
u = 1 \text{ or } \frac{q_0}{q_2}
\]
**Giant component**

\[ u = 1 \]  In this case, the giant component vanishes:

\[ S = 1 - g_0(1) = 0 \]

\[ u = \frac{q_0}{q_2} \]  This solution exists if

\[ \frac{q_0}{q_2} = \frac{p_1}{3p_3} < 1 \iff p_3 > \frac{1}{3} p_1 \]

independent of \( p_0 \) and \( p_2 \)!

Its size is given as

\[ S = 1 - g_0(u) = 1 - p_0 - \frac{p_1^2}{3p_3} - \frac{p_1^2 p_2}{3p_3^2} - \frac{p_1^3}{27p_3^2} \]
Giant component

Graphical solution of \( u = g_1(u) \):

- \( u = 1 \) is always a solution
- \( g_1 \) is monotone and convex on \([0, 1]\)
- Thus, second solution exists if

\[
g'_1(1) > 1
\]

\[
\langle k^2 \rangle > 2 \langle k \rangle^2
\]
**Giant component**

Graphical solution of $u = g_1(u)$:

- $u = 1$ is always a solution
- $g_1$ is monotone and convex on $[0, 1]$
- Thus, second solution exists if

$$g'_1(1) > 1$$

From

$$g'_1(1) = \sum_{k=0}^{\infty} kq_k = \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} k(k + 1)p_{k+1} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle}$$

we conclude that the giant component exists when

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