Lecture 2: Classical random graphs

Prof. Sotiris Nikoletseas

University of Patras
and CTI

ΥΔΑ ΜΔΕ, Patras
2020 - 2021

In this lecture

- We give an insight into the simplest, most studied random networks: the classical random graphs
- Two basic models:
  - $G_{N,p}$: a probability space (statistical ensemble) of networks with N nodes and probability p that any two nodes are linked, independently for the various links.
  - $G_{N,L}$: a probability space whose points are all possible labelled graphs of N nodes and L links (all such graphs having equal probability).

An example of a $G_{N,p}$ graph

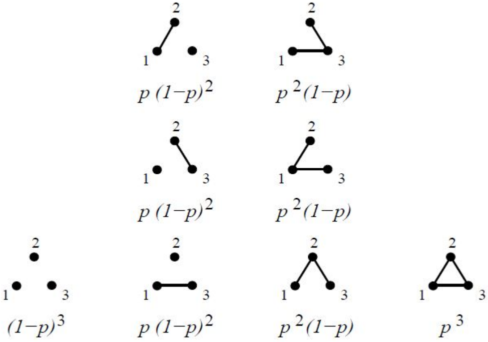


Figure: *The $G_{N,p}$ space, for N=3. All graphs in each column are isomorphic, that is they can be transformed into each other by simply relabelling their nodes.*

An example of a $G_{N,L}$ graph

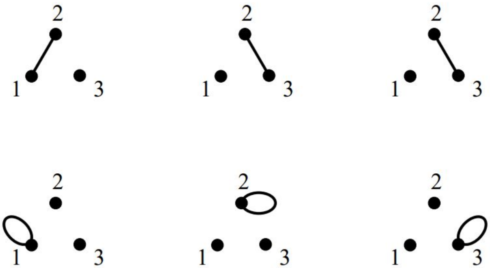


Figure: *The $G_{N,L}$ space, for N=3 and L=1.*

On the equivalence of the two models

- When $N \to \infty$ and the network is sparse, the two models are equivalent[1], taking

$$p = \frac{L}{\binom{N}{2}}$$

- Indeed, note that the number of links in $G_{N,p}$ follows the binomial distribution $B(\binom{N}{2}, p)$, so the average number of links is $\binom{N}{2} \cdot p$
- The degree distribution of $G_{N,p}$ is clearly:

$$P(q) = \binom{N-1}{q} \cdot p^q (1-p)^{N-1-q}$$

(the probability that a random node has degree q)

- The mean degree of a node is $\langle q \rangle = p(N-1)$

[1]Note: the multiple connections and loops in large $G_{N,L}$ do not harm the equivalence, since in large, sparse graphs there are only very few of them.

The notion of uncorrelated networks

- When $N \to \infty$ and the mean degree $\langle q \rangle$ is finite ($i.e.$, when $p \to \frac{\textit{constant}}{N}$) then the binomial distribution converges to the Poisson and we get:

$$P(q) = e^{-\langle q \rangle} \cdot \frac{\langle q \rangle^q}{q!}$$

- Because of the factorial in the denominator, the degrees decay very fast (in contrast to real networks where degrees decay much slower).
- Most importantly, the degrees of various nodes are statistically independent of each other; this applies even to connected nodes! (the only restriction is the fixed mean degree of each node).
- Such networks are called uncorrelated networks, and we will address this notion in the following lectures in detail.

Loops (cycles) in classical random graphs (I)

- We will see that large, sparse random graphs have few loops.
- Indeed, recall that the clustering coefficient of a node is the probability that two neighbors of the node are themselves neighbors. In the $G_{N,p}$ case this is:

$$C = p = \frac{\langle q \rangle}{N-1} \simeq \frac{\langle q \rangle}{N},$$

  where $\langle q \rangle$ the mean degree.
- So, in infinite, sparse $G_{N,p}$ the clustering coefficient approaches zero, and clustering has only a finite effect.
- As an example, imagine a random network with 10^5 nodes where the mean number of neighbors of a node is 10, so the clustering coefficient would be $c \simeq 10^{-4}$, which is much smaller than in real networks (such as in the Internet).

Loops (cycles) in classical random graphs (II)

Recall that

$$C = 3 \cdot \frac{\#\textit{loops of length 3 in the network}}{\#\textit{connected triples of nodes}} = 3 \cdot \frac{N_3}{T},$$

where the denominator is clearly

$$T = \sum_{i=1}^{N} \binom{q_i}{2} = \sum_{i=1}^{N} \frac{q_i(q_i - 1)}{2} = \sum_{i=1}^{N} \frac{q_i^2}{2} - \sum_{i=1}^{N} \frac{q_i}{2},$$

where q_i the degree of node i. If $\langle\ \rangle$ represents average, then we easily get that:

$$T = \frac{N(\langle q^2 \rangle - \langle q \rangle)}{2}$$

Loops (cycles) in classical random graphs (III)

But for Poisson distributions it is $\langle q^2 \rangle = \langle q \rangle^2 + \langle q \rangle$, because

$$\begin{aligned} Var(q) &= E(q^2) - E^2(q) \\ &= \langle q^2 \rangle - \langle q \rangle^2 \\ &= E(q) = \langle q \rangle \end{aligned}$$

so:

$$T = N \cdot \frac{\langle q \rangle^2}{2}$$

and finally

$$\begin{aligned} N_3 &= \frac{C \cdot T}{3} \\ &= \frac{\langle q \rangle \cdot N \frac{\langle q \rangle^2}{2}}{N \cdot 3} \\ &= \frac{\langle q \rangle^3}{6} \end{aligned}$$

Loops (cycles) in classical random graphs (IV)

- This shows that in sparse random graphs the number of triangles does not depend on its size; this number is finite even if these graphs are infinite.
- Similarly, the number of loops of length L is

$$N_L \simeq \frac{\langle q \rangle^L}{2L},$$

provided L is smaller than $\ln N$ (the network diameter).

- In other words, any finite neighborhood almost certainly does not contain any loops; such networks are **locally tree-like**.
- However, there are plenty of long loops of length exceeding $\ln N_L \sim N$ if $L >> \ln N$. Obviously, such long loops do not spoil the local tree-like character.

Cliques in random graphs

- Cliques are fully connected subgraphs e.g. a triangle is a 3-clique.
- Since there are so few loops in such networks, the 3-cliques are the maximum possible cliques and the bigger cliques in sparse classical random graphs are almost entirely absent.

Random Regular Graphs

- A similar random network is the random regular graph: all vertices of this graph have equal degrees.
- It is the probability space of all possible graphs with N vertices of degree q all, each such graph realized with equal probability.
- The number of loops of length L is, similar to the $G_{N,p}$ case:

$$N_L \simeq \frac{(q-1)^L}{2L},$$

  so these networks also have a locally tree-like structure.
- An infinite random regular graph approaches the Bethe lattice with the same degree.

The Diameter of random graphs (I)

- Diameter: maximum shortest path length
- We will exploit the local tree-like character of random networks (we start with a random tree).
- Let $\bar{b}$ the mean (expected) branching of a node ($\bar{b} = \bar{q} - 1$, where $\bar{q}$ the expected degree of the node).
- Then, by similar arguments as in the Bethe lattice/Cayley tree case, we have that the number z_n of the $n{-}th$ nearest neighbors of a node grows as $\bar{b}^n$.
- So the number of network nodes S_n which are not further than distance n from a given node is $\bar{b}^n$.
- Taking, roughly, $\bar{b}^{\bar{\ell}} \sim N$, where $\bar{\ell}$ the mean internode distance, yields:

  $$\bar{\ell} \simeq \frac{\ln N}{\ln \bar{b}},$$

  for large N.
- This result is actually valid for all uncorrelated networks.

The Diameter of random graphs (II)

- In random q-regular graph $b = q - 1$ so we get

$$\bar{\ell} \simeq \frac{\ln N}{\ln(q-1)},$$

- To obtain the diameter of the $G_{N,p}$ random graph, we need to evaluate its **average branching**.
- Let the node degrees be $q = 0, 1, 2, ...$ Let $N(q)$ the number of nodes of degree q.
- For a random node, the degree distribution is:

$$P(q) = \frac{N(q)}{N}$$

The Diameter of random graphs (III)

- Now let us focus on the degree distribution of nodes, who are end nodes of a randomly chosen link.

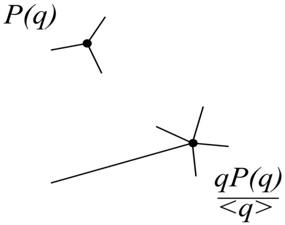


Figure: *End nodes of a randomly chosen link in a network have different statistics of connections from the degree distribution of this network.*

- Interestingly, we will show that the degree distribution of such end nodes is different to the degree distribution of a random node (which is not necessarily an end node)!

The Diameter of random graphs (IV)

- Clearly $\sum_q N(q) = N$. Also, $\sum_q q \cdot N(q) = N\langle q \rangle$, where $\langle q \rangle$ the mean degree.
- Let us randomly choose a link and then randomly one of its end nodes. The probability of this end node having degree q is

$$\frac{q \cdot N(q)}{N\langle q \rangle},$$

  since the number of all ("directed") links in the network is $N\langle q \rangle$ and the "directed" links adjacent to q-degree node is clearly $N(q) \cdot q$
- Thus, the degree distribution of a q-degree end node is

$$\frac{q}{\langle q \rangle} \cdot \frac{N(q)}{N} = \frac{q \cdot P(q)}{\langle q \rangle}$$

The Diameter of random graphs (V)

- So we have proven that: in a random network with degree distribution $P(q)$, the degree distribution of an end node of a randomly chosen link, is not $P(q)$ but
$$\frac{q \cdot P(q)}{\langle q \rangle}$$
- In other words, the connections of end nodes of links are organized in a different way from those of randomly chosen nodes!

The Diameter of random graphs (VI)

- Now, the average degree of an end node of a randomly chosen link is:

$$\sum_q q \cdot Pr\{\textit{degree} = q\} = \sum_q q \cdot \frac{q \cdot P(q)}{\langle q \rangle} = \frac{1}{\langle q \rangle} \sum_q q^2 \cdot P(q) = \frac{\langle q^2 \rangle}{\langle q \rangle},$$

  which is greater than the mean degree $\langle q \rangle$ of random nodes.

- So, the mean branching is:

$$\bar{b} = \frac{\langle q^2 \rangle}{\langle q \rangle} - 1$$

  But for the Poisson distribution it is $\langle q^2 \rangle = \langle q \rangle^2 + \langle q \rangle$, so:

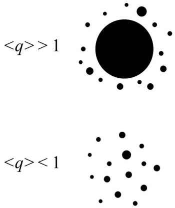
$$\bar{b} = \langle q \rangle$$

  And the final famous diameter formula is:

$$\ell \sim \frac{\ln N}{\ln \langle q \rangle}$$

The birth of a giant component (I)

- In the above derivation of the diameter we assumed the graph is connected.
- However, when the mean degree is low (e.g. $\langle q \rangle$ close to 0), the graph is actually disconnected, consisting of several, different connected components.
- Interestingly, when $\langle q \rangle$ exceeds 1, the graph includes a single "giant" component: a large connected component with $\epsilon \cdot N$ nodes ($\epsilon > 0$ constant, N the total number of nodes). Also, numerous much smaller components are included.
- On the other hand, if $\langle q \rangle < 1$ a giant component is absent and there are only plenty of small components.



The birth of a giant component (II)

- The emergence of the giant component (when the mean degree $\langle q \rangle$ surpasses 1) happens without a jump; its birth is a continuous phase transition where $\langle q \rangle = 1$ is the critical point.
- Note that this transition happens when the network is still quite sparse ($\langle q \rangle << N$ and the number of links is linear in the number of network nodes). Actually, the giant component relative size becomes almost 99% already when $\langle q \rangle = 5$.
- Near the birth point, the relative size is $s \simeq 2(\langle q \rangle - 1)$, e.g. when $\langle q \rangle = 1.01$ then $S = 0.02 = 2\%$ (in other words $0.02 \cdot N$ nodes belong to the giant component).

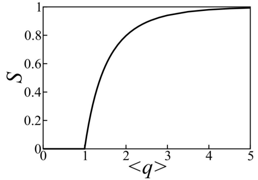


Figure: *The relative size of a giant connected component in a classical random graph versus the mean degree of its nodes. Near the birth point,* $s \simeq 2(\langle q \rangle - 1)$.

Smaller components

- What about the connected components, beyond the giant one?
- Actually, away from the birth point, the biggest non-giant component, the second biggest, the third etc., all have sizes of the order of $\ln N$ (much smaller than the giant one) and their number grows with N.
- Let us now move to the critical point, where a giant component is still absent. At this point, the biggest connected component, the second/third biggest and so on, all of these components are of the order of $N^{2/3}$, a size much smaller than N (the network size) but much bigger than $\ln N$. This is due to the fact that, away from the critical point, the distribution of connected component size has a rapid exponential decay; in contrast, exactly at the critical point, the size distribution of component decays slowly as a power law:

$$P(s) \sim s^{-5/2}$$

The transition regime

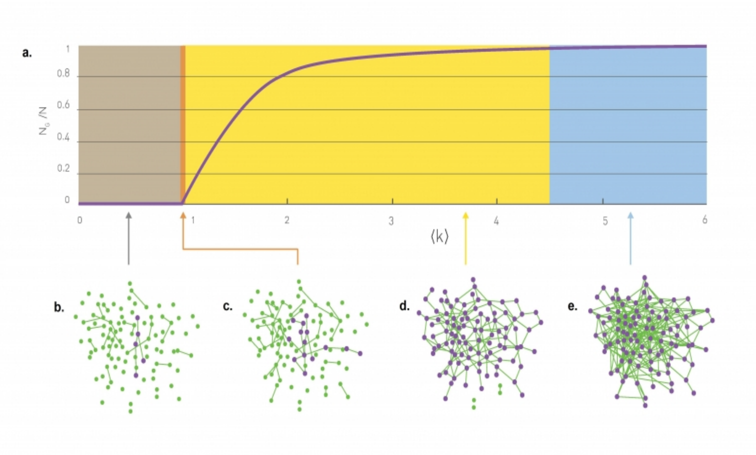


Figure: *The evolution of connectivity*

The supercritical regime

- It has the most relevance to real systems. It takes place when

$$p > 1/N, \ (\langle k \rangle > 1)$$

  where $\langle k \rangle$ the average degree.
- It contains numerous isolated components coexisting with the giant component. These smaller components are trees, while the giant component contains loops and cycles.

The connected regime

- For sufficient large p ($p > \frac{\ln N}{N}$) we have $\langle k \rangle > \ln N$ and the giant component "absorbs" all nodes and components, and the network becomes connected. Note that the network is still sparse ($> N \ln N \quad links$).

Rough estimation of the giant component birth

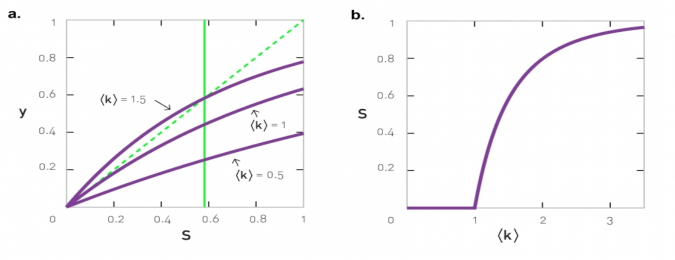
- Let $u = 1 - \frac{N_G}{N}$ the fraction of nodes not in the giant component G_C (N_G : the giant component size)
- Let i a node not in G_C and j another node. Then, either a) node i is not connected to node j (the probability for this is $1 - p$) or b) i is connected to j but $j \notin G_C$; this happens with probability $p \cdot u$. The total probability that i is not connected to G_C is

$$(1 - p + p \cdot u)^{N-1}$$

- As u is the fraction of nodes not in the G_C, taking $p = \frac{\langle k \rangle}{N-1}$ then solving $u = (1 - p + pu)^{N-1}$ gives $\ln u \simeq -\langle k \rangle (1 - u)$

Rough estimation of the giant component birth

- Taking exponential of both sides leads to $u = e^{-\langle k \rangle (1-u)}$
- Taking S the fraction of nodes in the G, it is $S = \frac{N_G}{N}$, so $S = 1 - u$ and $S = 1 - e^{-\langle k \rangle \cdot S}$
- This formula provides the size S of the G_C as a function of $\langle k \rangle$. Although looking simple, it does not have a closed solution, so we can “solve” it graphically



Rough estimation of the fully connected regime

- the probability that a randomly selected node does not have a link to the giant component is:

$$(1-p)^{N_G} \simeq (1-p)^N,$$

where N_G the giant component size (in this regime $N_G \simeq N$). The expected number of such isolated nodes is:

$$I_N = N(1-p)^N \simeq N \cdot e^{-Np}$$

- Let us examine when only one (1) node is disconnected from giant component:

$$I_N = 1 \Rightarrow N \cdot e^{-Np} = 1 \Rightarrow p = \frac{\ln N}{N},$$

which yields $\langle k \rangle = \ln N$

Why real networks are not Poisson?

- How big are the differences between node degrees? Can high-degree nodes coexist with small-degree nodes?
- Recall that the degree distribution in random networks is approximately Poisson:

$$P_k = e^{-\langle k \rangle} \cdot \frac{\langle k \rangle^k}{k!}$$

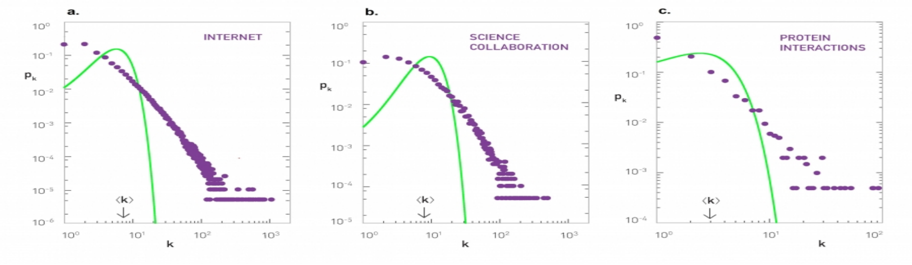
- Form Stirling's approximation: $k! \sim [\sqrt{2\pi k}](\frac{k}{e})^k$, we get:

$$P_k = \frac{e^{-\langle k \rangle}}{\sqrt{2\pi k}} \cdot \left(\frac{e \cdot \langle k \rangle}{k}\right)^k$$

- For degree $k > e \cdot \langle k \rangle$, the parenthesis term is smaller than 1, and both this term amd $1/\sqrt{k}$ decrease rapidly with k increasing. Thus, the chance of hubs (nodes of high degree) decreases very fast (faster than exponentially).

Why real networks are not Poisson?

- The figure below shows the degree distribution of three real networks, together with the corresponding Poisson fit:



- The figure shows the significant deviations, since the Poisson model underestimates both the number of high-degree nodes (hubs) as well as the number of low degree nodes.

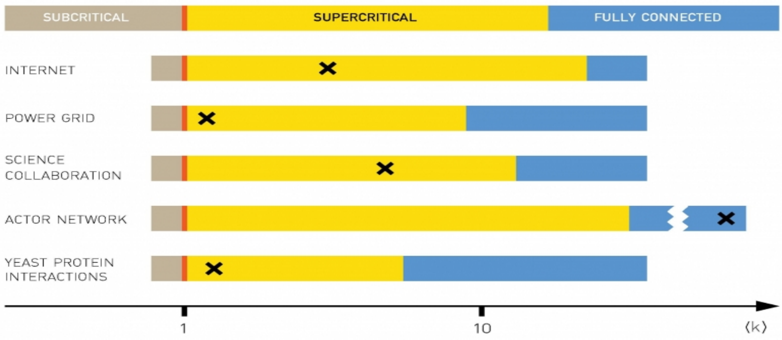
Most real networks are supercritical

- As the figure below shows, most real networks are not connected:

Network	N	L	‹K›	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,437	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61

Most real networks are supercritical

- As a matter of a fact, most of them are at the supercritical regime:



Random graph evolution

- In Erdös - Renyi $G_{n,p}$ random graphs, certain properties exhibit a threshold behavior, in the sense that they appear quite suddenly, for a small change of independent parameter (the link probability p) around a critical value p_c.
- Actually, when $p < p_c$ then the probability of $G_{n,p}$ having this property tends to 0 (as $N \to \infty$), while $p > p_c$ implies that the probability of the property tends to 1 as $N \to \infty$ (in other words, either no graph or all graphs in $G_{n,p}$ probability space have the property)!

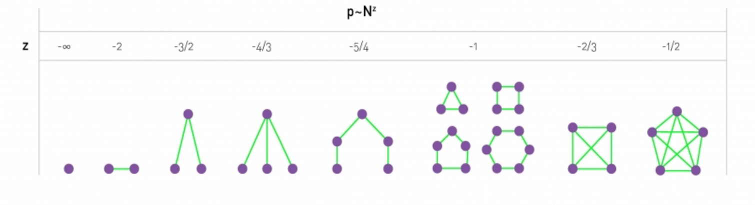


Figure: *Evolution of a Random Graph*