Threshold functions for global graph properties

Also known as the "phase transition".

The "epochs of evolution": Consider the structure of random graphs $G \in \mathcal{G}(n,p)$, as $p = p(n)$ increases. The following results can be shown (note that np = average node degree):

0. If $p \prec n^{-2}$, then a.e. G is empty.

1. If $n^{-2} \prec p \prec n^{-1}$, then a.e. G is a forest (a collection of trees).

   - The threshold for the apperarance of any k-node tree structure is $p = n^{-k/(k-1)}$.

   - The threshold for the appearance of cycles (of all constant sizes) is $p = n^{-1}$.

2. If $p \sim cn^{-1}$ for any $c < 1$ (i.e. $np \to c < 1$ as $n \to \infty$), then a.e. G consists of components with at most one cycle and $\Theta(\log n)$ nodes.

3. "Phase transition" or "emergence of the giant component" at $p \sim n^{-1}$ (i.e. $np \to 1$).

4. If $p \sim cn^{-1}$ for any $c > 1$ (i.e. $np \to c > 1$), then a.e. G consists of a unique "giant" component with $\Theta(n)$ nodes and small components with at most one cycle.

5. If $n^{-1} \prec p \prec \frac{\ln n}{n}$, then a.e. G is disconnected, consisting of one giant component and trees.

6. If $p \succ \frac{\ln n}{n}$, then a.e. G is connected (in fact Hamiltonian).

Theorem 7.15 *Let* $p_l(n) = \frac{\ln n - \omega(n)}{n}$, $p_u(n) = \frac{\ln n + \omega(n)}{n}$ *where* $\omega(n) \to \infty$. *Then*

(i) *a.e.* $G \in \mathcal{G}(n, p_l)$ *is disconnected;*

(ii) *a.e.* $G \in \mathcal{G}(n, p_u)$ *is connected.*

Proof. We shall use the second moment method on random variables $X_k = X_k(G)$ = number of components on G with exactly k nodes.

Assume without loss of generality that $\omega(n) \leq \ln \ln n$ and $\omega(n) \geq 10$.

(i) Set $p = p_l$ and compute $\mu = E(X_1), \sigma^2 = \mathrm{Var}(X_1)$. By linearity of expectation,

$$\begin{aligned}\mu = E(X_1) &= n(1-p)^{n-1} = ne^{(n-1)\ln(1-p)}\\ &\le ne^{-np} = ne^{-\ln n+\omega(n)} = e^{\omega(n)} \xrightarrow[n\to\infty]{} \infty.\end{aligned}$$

Furthermore, the expected number of ordered pairs of isolated nodes is

$$E(X_1(X_1-1)) = n(n-1)(1-p)^{2n-3}.$$

Hence,

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X_1) = E(X_1^2) - \mu^2\\ &= E(X_1(X_1-1)) + \mu - \mu^2\\ &= n(n-1)(1-p)^{2n-3} + n(1-p)^{n-1} - n^2(1-p)^{2n-2}\\ &\le n(1-p)^{n-1} + pn^2(1-p)^{2n-3}\\ &\le \mu + (\ln n - \omega(n))ne^{-2\ln n+2\omega(n)}\underbrace{(1-p)^{-3}}_{\le 2}\\ &\le \mu + \frac{2\ln n}{n}e^{2\omega(n)} \le \mu + 1 \qquad \text{for large } n.\end{aligned}$$

Thus, $\frac{\sigma^2}{\mu^2} \le \frac{\mu+1}{\mu^2} \to 0$ as $n \to \infty$, and by lemma 7.10,

$$\Pr(G \text{ is disconnected}) \ge \Pr(X_1(G) > 0) \to 1 \text{ as } n \to \infty.$$

(ii) (Here basic expectation estimation, or "1$^{\text{st}}$ moment method" suffices.)
Set $p = p_u = \frac{\ln n+\omega(n)}{n}$ and compute

$$\begin{aligned}\Pr(G \text{ is disconnected}) &= \Pr\left(\sum_{k=1}^{\lfloor n/2\rfloor} X_k \ge 1\right)\\ &\le E\left(\sum_{k=1}^{\lfloor n/2\rfloor} X_k\right) = \sum_{k=1}^{\lfloor n/2\rfloor} E(X_k)\\ &\le \sum_{k=1}^{\lfloor n/2\rfloor} \binom{n}{k}(1-p)^{k(n-k)} \qquad (5)\end{aligned}$$

Split the sum (5) in two parts:

(a)
$$\begin{aligned}
&\sum_{1\le k\le n^{3/4}} \binom{n}{k}(1-p)^{k(n-k)}\\
&\le \sum_{1\le k\le n^{3/4}} \left(\frac{en}{k}\right)^k e^{k(n-k)(-p)}\\
&= \sum_{1\le k\le n^{3/4}} \left(\frac{en}{k}\right)^k e^{-knp}e^{k^2p}\\
&\le \sum_{1\le k\le n^{3/4}} k^{-k}n^ke^ke^{-k(\ln n+\omega(n))}e^{k^2\cdot 2\ln n/n}\\
&= \sum_{1\le k\le n^{3/4}} k^{-k}e^{(1-\omega(n))k}e^{2k^2\ln n/n}\\
&\le e^{-\omega(n)}\cdot \underbrace{\sum_{1\le k\le n^{3/4}} \exp\left(-k\ln k+k+2k^2\frac{\ln n}{n}\right)}_{\le 3}\\
&\le 3e^{-\omega(n)}.
\end{aligned}$$

(b)
$$\begin{aligned}
&\sum_{n^{3/4}\le k\le n/2} \binom{n}{k}(1-p)^{k(n-k)}\\
&\le \sum_{n^{3/4}\le k\le n/2} \left(\frac{en}{k}\right)^k e^{k(n-k)(-p)}\\
&\le \sum_{n^{3/4}\le k\le n/2} \left(en^{1/4}\right)^k n^{-n/4}\\
&\le \frac{n}{2}e^{n/2}n^{-\frac{1}{4}n^{3/4}}\\
&\le n^{-n^{3/4}/5}\\
&= \exp\left(-\frac{n^{3/4}}{5}\ln n\right)\\
&\le e^{-\omega(n)} \text{ for large } n.
\end{aligned}$$

Thus, altogether

$$\Pr(G \text{ is disconnected}) \le 4e^{-\omega(n)} \xrightarrow[n\to\infty]{} 0. \square$$

What happens at the “phase transition” $p \sim n^{-1}$? For fixed values of n and $N = \binom{n}{2}$, consider the space of “graph processes” $\widetilde{G} = (G_t)_{t=0}^{N}$, where at each “time instant” t a new edge is selected uniformly at random for insertion into an n-node graph. (Thus, picking graph G_t from a randomly chosen process $\widetilde{G} \in \mathcal{G}(n,M)$, where $M = t$.)

Theorem 7.16 *Let $c > 0$ be a constant and $\omega(n) \to \infty$. Denote $\beta = (c - 1 - \ln c)^{-1}$ and $t = t(n) = \lfloor cn/2 \rfloor$. Then*

(i) *At $c < 1$, every component C of a.e. G_t satisfies*

$$\left| |C| - \beta \left(\ln n - \frac{5}{2} \ln\ln n \right) \right| \leq \omega(n).$$

(ii) *At $c = 1$, for any fixed $h \geq 1$ the h largest components C of a.e. G_t satisfy*

$$|C| = \Theta(n^{2/3}).$$

(iii) *At $c > 1$, the largest component C_0 of a.e. G_t satisfies*

$$||C_0| - \gamma n| \leq \omega(n) \cdot n^{1/2},$$

where $0 < \gamma = \gamma(c) < 1$ is the unique root of

$$e^{-c\gamma} = 1 - \gamma.$$

The other components C of a.e. G_t satisfy also in this case

$$\left| |C| - \beta \left(\ln n - \frac{5}{2} \ln\ln n \right) \right| \leq \omega(n).$$

Thus, the fraction of nodes in the “giant” component of a.e. G_t for $t = cn/2$ behaves as illustrated in Figure 8.

Let us prove one part of this result, the emergence of a gap in the component sizes of $G \in \mathcal{G}(n,p)$ at $p \sim n^{-1}$. (This corresponds to $t \sim N_p \sim n/2$.)

Theorem 7.17 *Let $a \geq 2$ be fixed. Then for large n, $\varepsilon = \varepsilon(n) < 1/3$ and $p = p(n) = (1+\varepsilon)n^{-1}$, with probability at least $1 - n^{-a}$, a random $G \in \mathcal{G}(n,p)$ has no component C that satisfies*

$$\frac{8a}{\varepsilon^2} \ln n \leq |C| \leq \frac{\varepsilon^2}{12} n.$$

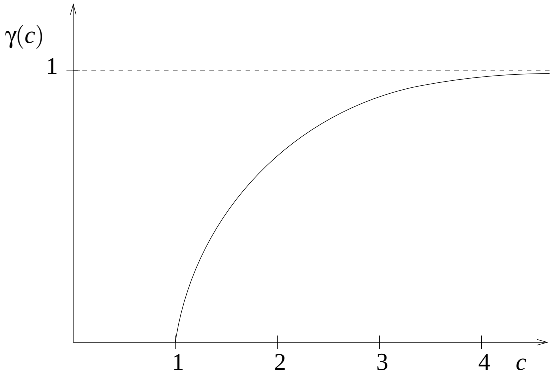

Figure 8: Fraction of nodes in the giant component.

Proof. Let us consider "growing" the component $C(u)$ of an arbitrary node u in G incrementally as follows:

1. (Stage 0:) Set $A_0 = \varnothing, B_0 = \{u\}$.
2. (Stage $i+1$:) If $B_i = A_i$, then stop with $C(u) = B_i$. Otherwise pick an arbitrary $v \in B_i \setminus A_i$; set $A_i = A_i \cup \{v\}$, $B_{i+1} = B_i \cup \{\text{neighbours of } v \text{ in } G\}$.

Now what is the probability distribution of $|B_i|$ (=size of set B_i)?

Consider any node $v \in G \setminus \{u\}$. It participates in i independent Bernoulli trials for being included in B_i, each with success probability equal to p. Thus the inclusion probability for any fixed $v \neq u$ is $1-(1-p)^i$, independently of each other.

Consequently, the size of each B_i obeys a simple binomial distribution

$$\Pr(|B_i| = k) = \binom{n-1}{k}(1-(1-p)^i)^k(1-p)^{i(n-k-1)}.$$

This gives also for each k an upper bound on the probability

$$\Pr(|C(u)| = k) = \Pr(|B_i| = k \wedge \text{ process stops at stage } i).$$

Denoting $p_k = \Pr(|C(u)| = k)$ for any fixed $u \in G$, it is clear that

$$\Pr(G \text{ contains a component of size } k) \leq np_k,$$

and to prove the theorem it suffices to show that

$$\sum_{k=k_0}^{k_1} p_k \leq n^{-a-1},$$

where $k_0 = \lceil 8a\varepsilon^{-2}\ln n\rceil$, $k_1 = \lceil \varepsilon^2 n/12\rceil$.
Since presumably $k_0 \le k_1$, we may assume $\varepsilon^4 \ge \frac{96a\ln n}{n} \ge \frac{1}{n}$.
We may now estimate

$$p_k \le \Pr(|B_i| = k) \le \frac{n^k}{k!}e^{-\frac{k^2}{2n}}(kp)^k(1-p)^{k(n-k-1)}, \tag{6}$$

because

$$\binom{n-1}{k} = \frac{n^k}{k!}\prod_{j=1}^{k}\left(1-\frac{j}{n}\right) \le \frac{n^k}{k!}e^{-\frac{k^2}{2n}}, \text{ and}$$

$$(1-p)^k \ge 1-kp.$$

Applying Stirling's formula

$$\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \le k! \le e^{\frac{1}{12k}}\sqrt{2\pi k}\left(\frac{k}{e}\right)^k$$

and the bounds $k_0 \le k \le k_1$ to (6) we obtain

$$\begin{aligned} p_k &\le \exp\left(\frac{-k^2}{2n} - \frac{\varepsilon^3 k}{3} + \frac{k^2(1+\varepsilon)}{n}\right) \\ &\le \exp\left(\frac{-\varepsilon^2 k}{3} + \frac{k^2}{n}\right) \\ &\le \exp\left(\frac{-\varepsilon^2 k}{4}\right), \end{aligned}$$

and consequently

$$\begin{aligned} \sum_{k=k_0}^{k_1} p_k &\le \sum_{k=k_0}^{k_1} e^{-\varepsilon^2 k/4} \le e^{-\varepsilon^2 k_0/4}\cdot(1-e^{-\varepsilon^2/4})^{-1} \\ &\le \frac{5}{\varepsilon^2}\cdot e^{-\varepsilon^2 k_0/4} \le 5\sqrt{n}\cdot n^{-2a} \\ &= 5n^{-2a+1/2} < n^{-a-1}. \end{aligned}$$

for large n. □

7.2 Nonuniform Models

Introduction

Obviously (in hindsight), most large "real-world" networks do not conform to the Erdös-Rényi random graph model. Consider e.g. the Internet, the WWW, traffic networks (airline connections, roads), collaboration networks (scientists, artistic, business), etc. All these exhibit strong nonuniformities: clustering, nodes with exceptionally high degree, ("hubs") etc.

This was noted (vaguely) in the social sciences at least in the 1960's (Milgram, "six degrees of separation") and also in popular culture ("small worlds", "the Kevin Bacon game").

Curiously, the first serious mathematical (physical) investigation of the phenomenon seems to have been Duncan Watts' Ph.D. thesis (under Steven Strogatz) in 1998 (?), and the "letter" to Nature by Watts and Strogatz in June 1998.

The Watts & Strogatz paper set off a veritable avalanche of work in the area – fueled in no small part by the current interest in modeling the Internet and the WWW.

"Small World" Networks

Watts & Strogatz 1998 etc.

Empirical measurements of real networks vs. predictions of the ER random graph model showed that the ER model is not an adequate model of practical networks.

Statistical measures on a graph $G = (V, E)$, $|V| = n$:

- **Characteristic path length** = average distance between nodes:

  $$\mathcal{L}(G) = \binom{n}{2}^{-1} \sum_{u \neq v} \mathrm{dist}(u, v),$$

  where $\mathrm{dist}(u, v)$ is the length of the shortest path between u and v.

- **Clustering coefficient**

  $$\mathcal{C}(G) = n^{-1} \sum_{v} \rho(\Gamma_v),$$

  where Γ_v is the subgraph of G induced by the neighbours of node v in G,

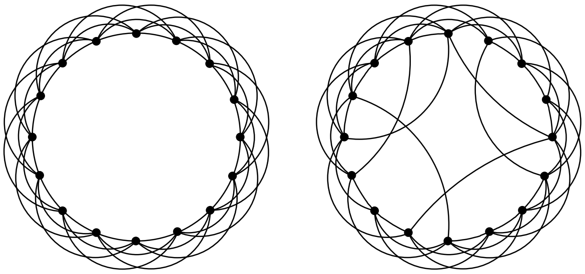

Figure 9: The SW random graph model: circulant graph and rewired graph.

and for a graph Γ with k nodes and l edges, the *density* of Γ is[5]

$$\rho(\Gamma) = l/\binom{k}{2}.$$

Watts and Strogatz considered the following three empirical graphs (n = number of nodes, δ = average node degree; only the largest component of each graph was chosen):

- Hollywood film actors collaboration network: $n = 225226$, $\delta = 61$
- Power grid of the western US: $n = 4941$, $\delta = 2.67$
- Neural network of nematode *Caenorhabditis elegans*: $n = 282$, $\delta = 14$

Watts and Strogatz obtained the following comparisons ($\mathcal{L}_{ER}$ and $\mathcal{C}_{ER}$ denote the corresponding values for ER random graphs of comparable size and density):

	$\mathcal{L}$	$\mathcal{L}_{ER}$	$\mathcal{C}$	$\mathcal{C}_{ER}$
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.08	0.0005
C. elegans	2.65	2.25	0.28	0.05

The empirical conclusion is thus that "real networks" have path length comparable to ER random graphs (= short) but considerably higher clustering. To model such observations, Watts and Strogatz introduced a specific "small world" (SW) random graph model, whereby one starts with a "circulant graph" $C_{n,k}$, and then randomly "rewires" some small fraction p of the edges. (Cf. Figure 9.)

[5] To be precise, the definition requires that $k \geq 2$. For nodes v with 0 or 1 neighbours, it is most convenient to stipulate that the neighbourhood density corresponds to the global density, i.e. that $\rho(\Gamma_v) = |E|/|V|$.

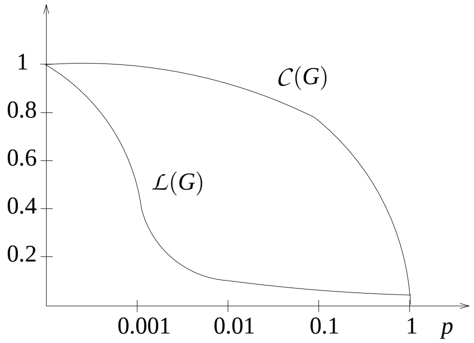

Figure 10: Path length and clustering coefficient in SW random graphs.

Watts & Strogatz experimented on the effect of the rewiring probability p on $\mathcal{L}(G)$ and $\mathcal{C}(G)$ in this model and obtained results as indicated in Figure 10 (curves normalised by $\mathcal{C}(C_{n,k})$ and $\mathcal{L}(C_{n,k})$; $n = 1000$, $k = 5$). Thus, the "small world" phenomenon of small $\mathcal{L}$ and large $\mathcal{C}$ seems to occur for p in the range $0.0005\ldots0.05$.

Watts and Strogatz call all graph families with this qualitative property "small world graphs". The notion has also been quantified by Walsh (1999) in terms of the *proximity ratio*

$$\mu = \frac{\mathcal{C}/\mathcal{L}}{\mathcal{C}_{ER}/\mathcal{L}_{ER}}.$$

Thus, presumably $\mu \gg 1$ for small world graphs. However, this quantity does not seem to be very invariant over various SW graph families. E.g. for *C. elegans*, $\mu \approx 4.8$ and for the power grid graph $\mu \approx 106$, but for the actors' network $\mu \approx 2400$.

For analytical simplicity, Newman et al. (1999, 2000) modified the Watts-Strogatz SW model to simply adding a fraction p of random cross edges, rather than rewiring. This variant of the model is called the "solvable SW", or SSW model.

Other Small World Models

- **Kleinberg's (2000) lattice model:** Basis is an $s \times s$ square lattice, with Manhattan (L_1) metric:

  $$\mathrm{d}(u,v) = \mathrm{d}\left((i,j),(k,l)\right) = |k-i| + |l-j|.$$

  Each node u has local connections to all nodes within distance $d \leq p$, and in addition $q \geq 0$ directed "long distance" connections. The probability of creating a long distance connection between u and v is proportional to thei distance, $\Pr\left((u,v)\right) \propto \mathrm{d}(u,v)^{-r}, r \geq 0$.

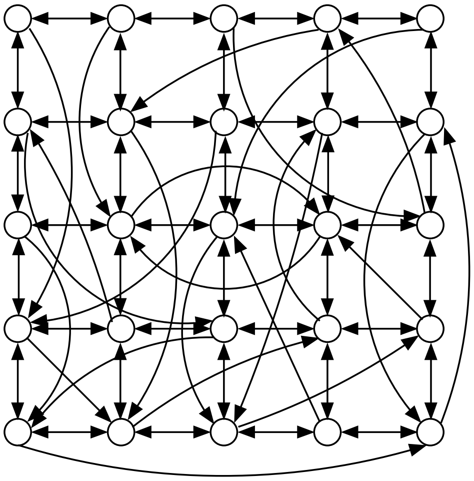

Figure 11: A Kleinberg lattice.

- **“Caveman graphs”:** (Watts 1999; old idea?) Deterministic SW graph model. Connect a collection of *r* “*k*-man caves” (*k*-cliques) together in a systematic manner.

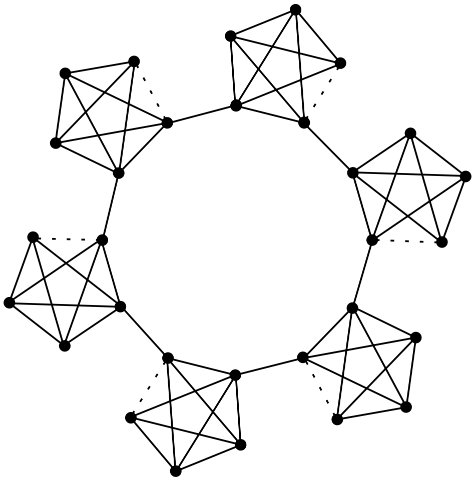

Figure 12: A collection of six 5-caves connected together in a 6-cycle.

Scale Free Networks

So are small world graphs a good model of real world networks? Not always. (Usually not?)

One aspect of real networks that SW graphs often do not model well is the degree

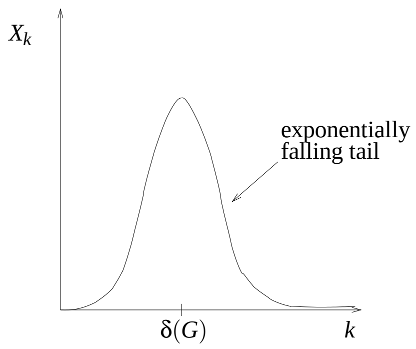

Figure 13: Degree distribution of an ER random graph.

distribution. In an ER random graph $G \in \mathcal{G}(n,p)$, the degree distribution is almost binomial with parameters $n-1, p$. For large n and small p, the distribution approaches Poisson(λ), where $\lambda = np$.

More precisely, if $X_k = X_k(G)$ = number of nodes in G with deg = k, then

$$P(k) = \frac{E(X_k)}{n} = \binom{n-1}{k} p^k (1-p)^{n-1-k} \approx e^{-np}\frac{(np)^k}{k!} \approx e^{-\delta}\frac{\delta^k}{k!},$$

where δ = average degree of graph G. Thus, the degree distribution of a typical ER graph G looks as illustrated in Figure 13.

The degree distributions of SW graphs are typically even more peaked around $\delta(G)$. E.g. in WS graphs based on the circulant $C_{n,t}$, approximately fraction $1-2tp$ of the nodes has degree equal to $2t$ (recall that $p \ll 1$ is the rewiring probability).

However, many real world networks seem to have very heavy tailed degree distributions, well matched by "power laws"

$$P(k) \propto k^{-\gamma},$$

where $\gamma = 2 \ldots 4$. This indicates that there are some nodes with unreasonably large (in the ER or SW models) degrees. Also, such networks are called "scale free", because there is no characteristic "scale" or node degree value at which large networks would concentrate.

On a log-log plot, the degree distributions of such networks look somewhat as in Figure 14

For instance, the following values for γ have been estimated for real world networks (Barabási & Albert 1999)

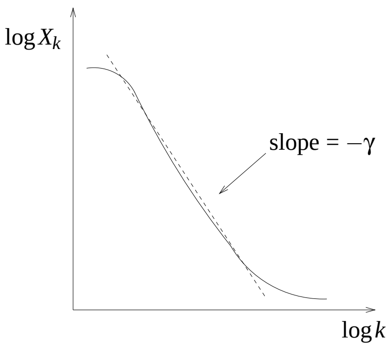

Figure 14: Degree distribution of a "scale-free" random graph.

	n	δ	γ
Film actors	212250	28.8	2.3 ± 0.1
WWW (local)	325729	5.46	2.1 ± 0.1
Power grid	4941	2.67	4

Barabási & Albert (1999) proposed the following attractive "growth and preferential attachment" model (BA model) to explain the emergence of such power law degree distributions in networks:

- The network is initialised at time $t = 0$ with some small set of nodes and edges, $G_0 = (V_0, E_0)$
- At time $t+1$, a new node u is introduced to the network, with d_0 edges that are preferentially attached to the existing nodes $v \in V_t$ so that

$$\Pr\left((u,v) \in E_{t+1}\right) \propto \deg_t(v).$$

Barabási and Albert argue heuristically and experimentally that this growth process yields networks with power law degree distributions

$$P(k) \propto k^{-3}.$$

They also claim that with nonlinear preferences the exponent γ can be adjusted also to values different than 3.

These arguments have been made rigorous by Eriksen & Hörnquist (2002) and by Krapivsky (2000). (However some problems still remain with nonlinear preferences?)

Finally, note that the popular experimental graphs (Internet, actors, power grid, etc.) have both small world and scale free properties, so neither the SW nor the BA model (which are mutually contradictory) provides a fully satisfactory explanation for them.