ELU 501
Data science, graph theory and social network studies

Yannis Haralambous (IMT Atlantique)

November 30, 2018

Part V

Lecture 5
Models of network formation

Generative network models

- *Generative network models* ⟨*modèles de réseau génératifs*⟩ model the mechanisms by which networks are created.

Power law distribution

- Let p_k be the fraction of vertices with degree k. (p_k) represents the *degree distribution* of the network.
- A network follows a *power law degree distribution* if
$$p_k \simeq Ck^{-\alpha}.$$
We call α the *exponent* of the power law. Values $2 \leq \alpha \leq 3$ are typical for networks such as the Internet.
- Power law degree distribution networks are also called *scale-free*.
- We can calculate mean and standard deviation of α by:
$$\overline{\alpha} = 1 + N\left[\sum_{i \geq i_{\min}} \log \frac{k_i}{k_{\min} - \frac{1}{2}}\right]^{-1} \qquad \sigma = \frac{\alpha - 1}{\sqrt{N}},$$
where $k_{\min}$ is the min degree for which the power law holds and N the number of vertices of degree $\geq k_{\min}$.

Preferential attachment by Price

- In the 1970s, Price has studied the network of bibliographical citations.
- His assumption was that *a newly appearing paper will cite previous papers with probability proportional to the number of citations these papers already have* (if a paper A has been cited 10 times more than paper B then the probability that the new paper C cites A is ten times higher than the probability it cites B).
- If we want to generate a network using this principle then we have a problem: how do we start? If there is no citation yet, the probability for every paper to be cited is 0.
- To solve this problem, Price introduces a factor a of probability attached to every paper, independently of the number of citations it may have.

Preferential attachment by Price

- More formally: to obtain a network implementing Price's paradigm, we add vertices with edges to (in average) c other vertices chosen at random with probability proportional to the indegrees of the destination vertices plus a constant a.
- Let q_i be the indegree of a directed graph vertex i, k_i the degree of an undirected graph vertex, $p_q(n)$ the fraction of vertices having indegree q in a directed graph of order n.
- We have

$$P(j \to i) = \frac{q_i + a}{\sum_i (q_i + a)} = \frac{q_i + a}{n(c + a)}.$$

- This is the probability that i is attained by an edge of a newly created vertex.

Preferential attachment by Price

- When a new vertex is created it has on average c new edges, and hence the probability that i is attained when creating a new vertex is
$$\frac{c(q_i + a)}{n(c + a)}.$$
- If we consider only vertices of indegree q then the expected number of citations to such vertices when creating a new vertex is
$$np_q(n)\frac{c(q + a)}{n(c + a)} = \frac{c(q + a)}{c + a}p_q(n).$$

Preferential attachment by Price

- Let us study the evolution of the network of order n when we add a vertex.
- The number of vertices of indegree q will be $(n+1)p_q(n+1)$ which is equal to
  1. those who were of indegree $q-1$ and are now of indegree q: there are $\frac{c(q-1+a)}{c+a}p_{q-1}(n)$ of them;
  2. minus those who were of indegree q and are now of indegree $q+1$: there are $\frac{c(q+a)}{c+a}p_q(n)$ of them;
  3. plus those whose indegree hasn't changed, there are $np_q(n)$ of them.
- That gives us the "master" formula, for $q \geq 1$:

$$(n+1)p_q(n+1) = \frac{c(q-1+a)}{c+a}p_{q-1}(n) - \frac{c(q+a)}{c+a}p_q(n) + np_q(n).$$

- For $q = 0$ the formula becomes:

$$(n+1)p_0(n+1) = 1 - \frac{ca}{c+a}p_0(n) + np_0(n),$$

since there is one new vertex of indegree 0 that is created.

Preferential attachment by Price

- When $n \to \infty$, these formulas become:
$$p_q = \frac{c}{c+a}[(q-1+a)p_{q-1} - (q+a)p_q,$$
$$p_0 = 1 - \frac{ca}{c+a}p_0,$$
where p_q is the probability of a vertex having indegree q in an infinite network.
- I.e.,
$$p_q = \frac{q+a-1}{q+a+1+a/c}p_{q-1},$$
$$p_0 = \frac{1+a/c}{a+1+a/c}.$$
- I.e.,
$$p_q = \frac{(q+a-1)(q+a-2)\cdots a}{(q+a+1+a/c)\cdots(a+2+a/c)}\frac{(1+a/c)}{(a+1+a/c)}.$$

Preferential attachment by Price

- Using the gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ with the property that $\Gamma(x+1) = x\Gamma(x)$ for all $x > 0$, we can write
$$p_q = (1 + a/c)\frac{\Gamma(q+a)\Gamma(a+1+a/c)}{\Gamma(a)\Gamma(q+a+2+a/c)}.$$
- Using the beta function $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ we get
$$p_q = \frac{B(q+a, 2+a/c)}{B(a, 1+a/c)}.$$
- By Stirling's approximation of the gamma function we have
$$\Gamma(x) \simeq \sqrt{2\pi}e^{-x}x^{x-\frac{1}{2}}$$
and therefore
$$B(x,y) \simeq x^{-y}\Gamma(y).$$
- Applying this to our network gives
$$p_q \sim q^{-\alpha} \text{ where } \alpha = 2 + \frac{a}{c}.$$
- Conclusion: the Price network has a power law degree distribution of exponent $2 + \frac{a}{c}$.

Preferential attachment by Price

- Note that the Price approach generates a dag (why?)

Preferential attachment by Barabási and Albert

- In the Barabási and Albert model we require an undirected graph where (1) every new vertex is connected to exactly c vertices and (2) the probability of an edge $j - i$ for a new vertex j is precisely proportional to degree k_i.
- This is a special case of the Price network: if we direct edges in order of creation of vertices, then every vertex has degree k_i equal to indegree q_i plus outdegree c.
- Therefore it is exactly as taking a Price network with $a = c$, and therefore its power law exponent is exactly 3.

Temporal degree distribution

- We can study the evolution of a network by calculating the *evolution of the indegree for a given vertex, as a function of time* (of creation of new vertices).
- Let us consider that at every time unit one vertex is created.
- Let $p_q(t, n)$ be *the average fraction of vertices that were created at time t and have degree q at the time when the network has reached n vertices*.
- We can take over previous formulas and replace $p_q(\cdot)$ by $p_q(t, \cdot)$ and $p_{q-1}(\cdot)$ by $p_{q-1}(t, \cdot)$. But if $n \to \infty$, these values tend to zero.
- Let us take a finite global time equal to 1 and a new variable $\tau = t/n$.
- We define $\pi_q(\tau, n)d\tau$ as the fraction of vertices created between τ and $\tau + d\tau$ and having degree q when the network has size n.

Temporal degree distribution

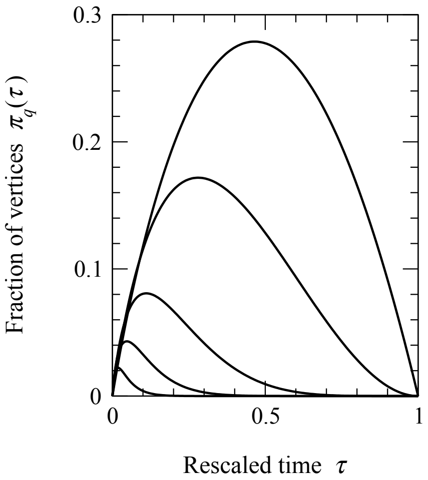
- We have
$$\pi_q(\tau, n) = np_q(t, n).$$
PROOF. In the interval $[\tau_0, \tau_0 + d\tau]$ we have $nd\tau$ new vertices. At time $t_0 = n\tau_0$, $np_q(t_0, n)$ vertices are created with final degree q. Therefore in the interval $[\tau_0, \tau_0 + d\tau]$ we have $np_q(t_0, n)d\tau$ new vertices with final degree q, and this is exactly the definition of $\pi_q(\tau_0, n)$ for all τ_0 and t_0, giving the above formula. QED

- Now we can rewrite the master equation as
$$\pi_q(\frac{n}{n+1}\tau, n+1) = \pi_q(\tau, n) + \frac{c}{c+a}[(q-1+a)\frac{\pi_{q-1}(\tau, n)}{n} - (q+a)\frac{\pi_a(\tau, n)}{n}],$$
where the $\frac{n}{n+1}$ comes from the fact that if $\pi_q(\tau, n) = np_q(t, n)$ with $\tau = t/n$, then $\pi_q(\tau', n+1) = (n+1)p_q(t, n+1)$ with $\tau' = t/(n+1)$, i.e., $\tau' = \frac{n}{n+1}\tau$.

Temporal degree distribution

- Writing $\varepsilon = 1/n$ and removing $o(\varepsilon^2)$ terms, we get
$$\frac{\pi_q(\tau) - \pi_q(\tau - \varepsilon\tau)}{\varepsilon} + \frac{c}{c+a}[(q-1+a)\pi_{q-1}(\tau) - (q+a)\pi_q(\tau)] = 0.$$
- When $\varepsilon \to 0$, we get
$$\tau\frac{d\pi_q}{d\tau} + \frac{c}{c+a}[(q-1+a)\pi_{q-1}(\tau) - (q+a)\pi_q(\tau)] = 0.$$
- Using tedious calculations of differential equations, we find that
$$\pi_q(\tau) = \frac{\Gamma(q+a)}{\Gamma(q+1)\Gamma(a)}\tau^{ca/(c+a)}(1-\tau^{c/(c+a)})^q.$$

Temporal degree distribution



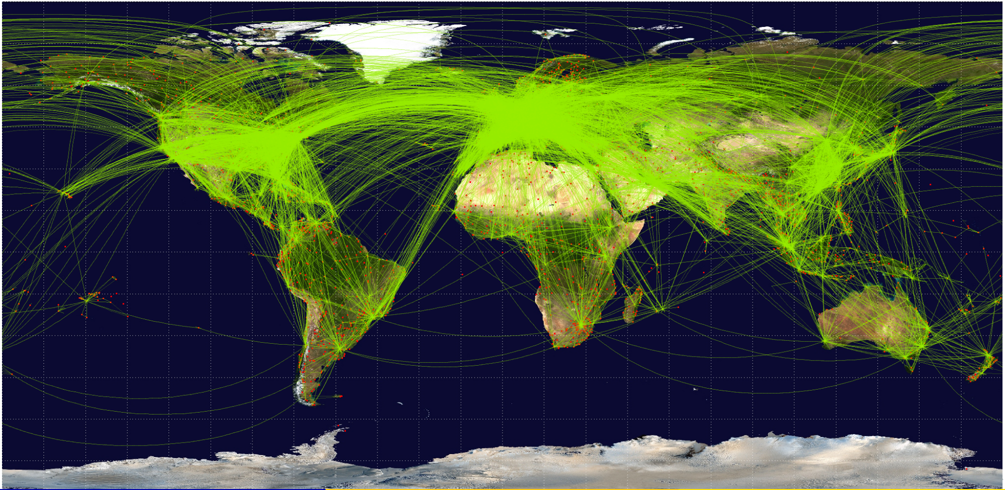
$c = 3$, $a = 1.5$, $q = 1, 2, 5, 10, 20$ (from top to bottom). [Newman, 2010, Fig. 14.3(a)] We see that ultimately high-indegree vertices are created early!

Temporal degree distribution

- [Salganik et al., 2006] have experimented fake download figures for songs and have discovered that they are more important than song quality.
- To be successful in some area you should better enter early: *first movers have a large advantage over others* [Newman, 2010, p. 508].

Network optimization

- A different formation mechanism for networks is *structural optimization*.
- A typical example is the optimization of air traffic network into a *hub-and-spoke* ⟨*réseau étoilé*⟩ network.



Network optimization

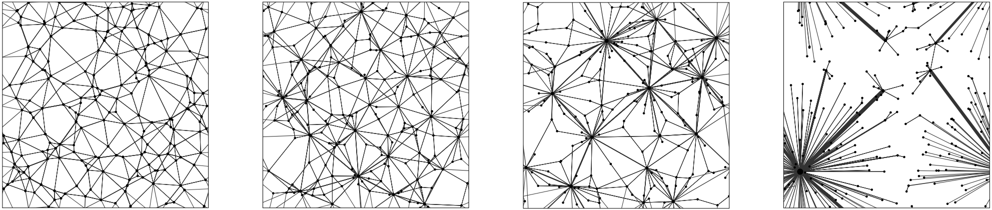
- In the case of air traffic we have a maintenance and operation cost measure m over edges (we simplify) and a dissatisfaction measure of the client ℓ which is the mean geodesic distance between vertex pairs.
- To get a small ℓ we have to increase the number of edges (between small airports), and that increases m. To decrease cost m we have to have a minimal number of edges (while keeping the graph connected), but that increases ℓ.
- Ferrer i Cancho and Solé studied the quality function
$$E(m,\ell) = \lambda m + (1-\lambda)\ell.$$
- One way of finding an approximation to $\operatorname{argmin} E$ is to minimize m by taking a tree and then search all possible trees to minimize ℓ. The solution is known: it is the *star graph*.

Network optimization

- In real life operation is *not* proportional to the number of edges only but also to their lengths, so the pure star graph is not applicable. But this explains nevertheless the hub-and-spoke phenomenon.
- But Ferrer i Cancho and Solé considered local minima by a greedy algorithm.
- They started with a random network and for every random pair of vertices added or deleted an edge at random comparing the value of E before and after.
- They did this until convergence. This algorithm provides a local minimum.
- Interesting behavior: for $\lambda \gg 0$ the algorithm rapidly gets in trouble and cannot find a hub-and-spoke arrangement. For $\lambda \ll 1$ it typically manages to find the star graph solution.

Network optimization

- Gastner and Newman considered also the geographic distance traveled in the calculation of the optimized network. They find a range of solutions between *road-like* and *airline-like* networks.



From [Newman, 2010, Fig. 14.11].

Random graphs

- We define a *random graph* $G(n, p)$ as a graph with n vertices and an independent probability p of having an edge between two arbitrary (distinct) vertices.
- The probability of having a given $G(n, p)$ graph with m edges is $P(G(n, p)) = p^m(1-p)^{\binom{n}{2}-m}$.
- A $G(n, p)$ graph is called a *Erdős-Rényi graph*.
- The probability of drawing a graph with n vertices and m edges is $P(m) = \binom{\binom{n}{2}}{m} p^m(1-p)^{\binom{n}{2}-m}$, which is just the standard binomial distribution. Therefore the mean $\overline{m}$ is

$$\overline{m} = \sum_{m=0}^{\binom{n}{2}} mP(m) = \binom{n}{2} p.$$

Random graphs

- The mean degree of a graph with n vertices and m edges is $2m/n$ (why?).
- Therefore in $G(n,p)$ we have

$$\bar{k} = \sum_{m=0}^{\binom{n}{2}} \frac{2m}{n} P(m) = \frac{2}{n}\binom{n}{2} p = (n-1)p.$$

We call this value c.

Random graphs

- The degree distribution of $G(n, p)$ is binomial: the probability for a given vertex of being connected to specific k other vertices and not to any of the others is $p^k(1-p)^{n-1-k}$. There are $\binom{n-1}{k}$ ways of doing this, therefore the total probability of being connected to any k others is

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

  which is binomial.
- When $n \to \infty$, we have $\log((1-p)^{n-1-k}) \simeq -(n-1-k)\frac{c}{n-1} \simeq -c$, and hence $(1-p)^{n-1-k} \simeq e^{-c}$. Also for large n we have $\binom{n-1}{k} \simeq \frac{(n-1)^k}{k!}$ and therefore we have

$$p_k \simeq \frac{(n-1)^k}{k!} p^k e^{-c} \simeq e^{-c} \frac{c^k}{k!},$$

  which is Poisson. Therefore $G(n, p)$ is sometimes called *Poisson random graph*.

Random graphs

- The *transitivity coefficient* of $G(n,p)$ is very easy to calculate: it is the probability that two neighbors of a vertex are also neighbors of each other. Here the probability of *any* two vertices being neighbors is always the same, namely $p = c/(n-1)$, therefore:
$$C = \frac{c}{n-1}.$$
- We see that for $n \to \infty$, the transitivity coefficient tends to 0.

Random graphs

- We will now study *giant components* of $G(n, p)$, i.e., largest components in the network increasing proportionally to n (= *extensive*).
- When $p = 0$ a $G(n, p)$ graph is discrete. When $p = 1$ it is a complete graph, i.e., a single component, its size is n.

> What happens when p goes from 0 to 1? When does the largest component becomes extensive?

- This happens at a specific value of p and is called *phase transition*.

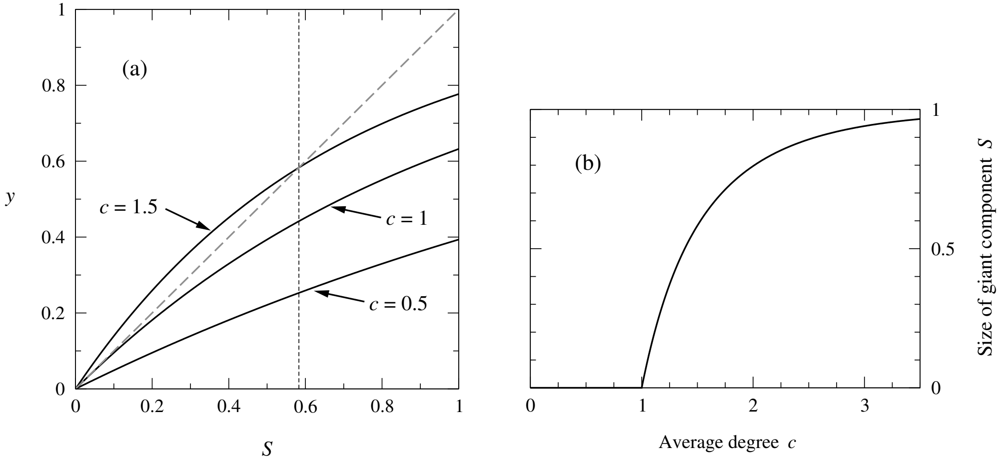
Random graphs

- Let u be the fraction of $G(n, p)$ which does not belong to the giant component.
- For i not to belong to the GC, we need two things for every other vertex j: either i is not connected to j (probability $1 - p$), or i is connected to j but j is itself not a member of the GC (probability pu).
- Therefore $u = (1 - p + pu)^{n-1} \simeq e^{-c(1-u)}$. And if S is the fraction of vertices in the GC, then

$$S = 1 - e^{-cS}.$$

- One proves that this equation has a nonzero solution only if $c > 1$.

Random graphs



On the left: curves $y = S$ and $y = 1 - e^{-cS}$ for different values of c.
On the right: the size of the GC depending on c.
From [Newman, 2010, Fig. 12.1].

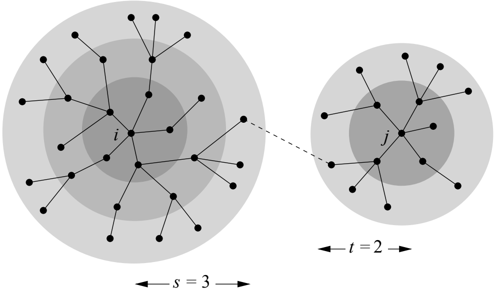
Random graphs

- The *small world effect* ⟨*effet du petit monde*⟩ is the hypothesis that in a very large network, paths between arbitrary vertices can be always very short.
- It is the case of Facebook: in 2011 there were *721M users, and the average shortest path length was 4.74!*
- What about $\mathrm{G}(n,p)$? Can we expect a similar behavior?
- We will study the diameter of $\mathrm{G}(n,p)$. Let c be the mean degree. Obviously the average number of vertices s steps away from a vertex is c^s. Roughly, when the whole graph is attained we will have $c^s \simeq n$ and hence $s \simeq \frac{\log n}{\log c}$.
- If there are 7B humans and every human has $\simeq$ 1k acquaintances, then $s \simeq 3.3$, which is even smaller than prophecied by Milgram.

Random graphs

The diameter of $G(n,p)$ is $s \simeq \frac{\log n}{\log c}$.

- Take surfaces at distance s and t from vertices i and j.



From [Newman, 2010, Fig. 12.6]

- Let d_{ij} be the distance between i and j. The probability that $d_{ij} > s+t+1$ is equal to the probability that there is no edge between the two surfaces:

$$P(d_{ij} > s+t+1) = (1-p)^{c^{s+t}}.$$

Random graphs

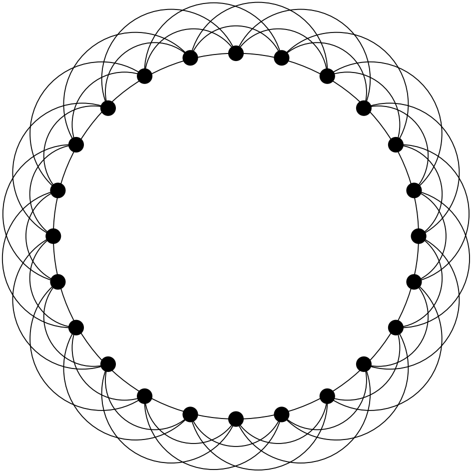
- Let $\ell := s + t + 1$. We have $\mathrm{P}(d_{ij} > \ell) = (1 - \frac{c}{n})^{c^{\ell-1}}$ and therefore $\log \mathrm{P}(d_{ij} > \ell) \simeq -\frac{c^\ell}{n} \Rightarrow \mathrm{P}(d_{ij} > \ell) \simeq e^{-\frac{c^\ell}{n}}$.
- The diameter of the network is the smallest value such that $\mathrm{P}(d_{ij} > \ell) = 0$. This can happen only if c^ℓ grows faster than n, i.e., $c^\ell = an^{1+\varepsilon}$.
- This means that $\ell = A + \frac{\log n}{\log c}$. QED

The small-world model

- Let us look into models of network formation other than preferential attachment and the random graph.
- Even though the random graph has a small-world property it is not a very good approximation of real networks (like Facebook) because of its low *transitivity*.
- The global transitivity coefficient of a social network like Facebook is around 0.4 while the transitivity coefficient of $G(n, p)$ is $\frac{c}{n-1}$: if we consider that the median friends count is 99 [Ugander et al., 2011, p. 3], for 721M users that gives 0.00000014, which is much less than 0.4.
- So we may ask: *how can we find a network formation model with better transitivity properties?*

The small-world model

- Take the following network (a *circle model* from [Newman, 2010, Fig. 15.2b]) where every vertex is connected to its c nearest neighbors (c even). Here $c = 6$:



The small-world model

- To find the clustering coefficient of the circle model, observe that a triangle is going twice right and then coming back. The way back can be at most $c/2$ units apart, therefore to advance we have $\binom{c/2}{2}$ choices, i.e., we have a total of $n\binom{c/2}{2}$ triangles.
- The number of pairs of edges adjacent at a given vertex is $\binom{c}{2}$.
- Therefore

$$C = \frac{3n\binom{c/2}{2}}{n\binom{c/2}{2}} = \frac{3(c-2)}{4(c-1)}.$$

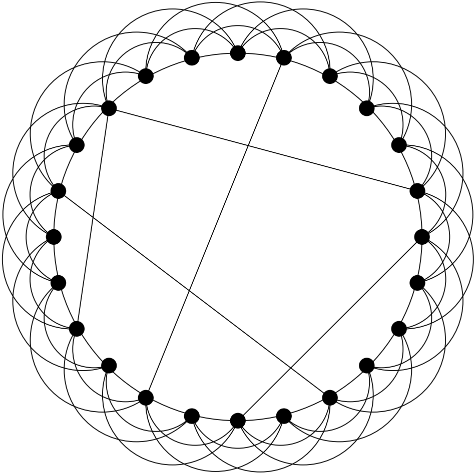
- This means that, independently of n, we can vary C from 0 to .75 when changing c.

The small-world model

- The circle model has nice transitivity characteristics but is a "large world": two vertices m units apart are connected by a shortest path of $\lceil 2m/c \rceil$ steps. Averaging over the complete range $m \in \{0, \ldots, \lceil \frac{n}{2} \rceil\}$ gives a mean shortest path of $n/2c$. For Facebook this would be 3.6M, that is too much.
- The circle model captures transitivity but is not a small world. $G(n, p)$ is a small world but does not capture transitivity.
- How can we have our cake and eat it too? ⟨*avoir le beurre et l'argent du beurre*⟩

The small-world model

- The *small-world model* ⟨🇫🇷 *modèle de petit monde*⟩ is defined as a circle model where for each edge, a new edge (called *shortcut* ⟨🇫🇷 *raccourci*⟩) is possibly added (with probability p):



The small-world model

- Degree distribution: at start there are $\frac{1}{2}nc$ edges, we add $\frac{1}{2}ncp$ new ones, that makes ncp edge ends, in average cp per vertex. The specific number s of shortcuts attached to any vertex is given by

$$p_s = e^{-cp}\frac{(cp)^s}{s!}.$$

- If we are interested in $k = s + c$, this gives

$$p_k = e^{-cp}\frac{(cp)^{k-c}}{(k-c)!}$$

for $k \geq c$ and $p_k = 0$ for $k < c$.

The small-world model

- Clustering coefficient: tedious calculations provide that
$$C = \frac{3(c-2)}{4(c-1) + 8cp + 4cp^2},$$
which is equal to the clustering coefficient of the cycle model when $p = 0$, and smaller otherwise.
- Average path lengths: an analytic expression of the path length is still an open problem.
- What can be done?
- Consider $c = 2$, then we have only a circle of length n and s shortcuts. The average distance between ends of shortcuts around the circle is $\xi = n/2s$.
- For c fixed, n and ξ specify entirely the model.

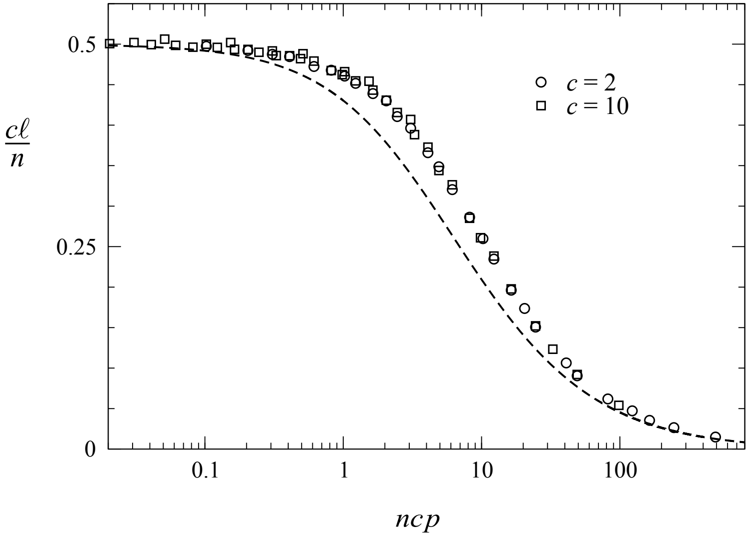
The small-world model

- But n and ξ both measure "length". If we take ℓ the length of the average shortest path and n, they are also "lengths". If we are interested in the ratio ℓ/n it is "dimensionless" and therefore *can only be a function of n/ξ*.
- We don't know that function, let us denote it by F: $\ell/n = F(n/\xi)$, and hence $\ell = nF(2s)$.
- For $c = 4$ we will roughly have halved ℓ (we halve the part of the path running on the circle, we don't touch the part in shortcuts, but this should be small).
- So we can assume that we have a formula of the kind $\ell/n = 2/cF(2s)$, or (with $f(x) = 2F(x)$):

$$\frac{c\ell}{n} = f(ncp).$$

The small-world model

- Arrived at the point $\frac{c\ell}{n} = f(ncp)$ we can numerical calculations:



From [Newman, 2010, Fig. 15.6]

The small-world model

- The curve found experimentaly can be described as
$$f(x) = \frac{1}{\sqrt{x^2+4x}} \log \frac{\sqrt{1+4/x}+1}{\sqrt{1+4/x}-1} \simeq \frac{\log(x)}{x} \text{ for } x \gg 1$$
and therefore
$$\ell \simeq \frac{\log(ncp)}{c^2 p} \text{ for } ncp \gg 1.$$
and so the increase of ℓ is logarithmic with respect to n, and this is precisely a small-world effect.

Bibliography

Barber, D. (2012).
Bayesian Reasoning and Machine Learning.
Cambridge University Press.

House, T. (2011).
Modelling epidemics on networks.
preprint `http://arxiv.org/abs/1111.4875`.

Koller, D. and Friedman, N. (2009).
Probabilistic Graphical Models, Principles and Techniques.
The MIT Press.

Newman, M. (2010).
Networks, an introduction.
Oxford University Press.

Salganik, M. J., Dodds, P., and Watts, D. (2006).
Experimental study of inequality and unpredictability in an artificial cultural market.
Science, 311:854–856.