School of Information Sciences
University of Pittsburgh

TELCOM2125: Network Science and Analysis

Konstantinos Pelechrinis

Spring 2015

Figures are taken from:
M.E.J. Newman, “Networks: An Introduction”

Part 5: Random Graphs

Graph models

- **We want to have formal processes which can give rise to networks with specific properties**
  - E.g., degree distribution, transitivity, diameters etc.
- **These models and their features can help us understand how the properties of a network (network structure) arise**
- **By growing networks according to a variety of different rules/models and comparing the results with real networks, we can get a feel for which growth processes are plausible and which can be ruled out**
  - Random graphs represent the “simplest” model

Random graph

- **A random graph has some specific parameters fixed**
  - In all other aspects the network are random
  - Simplest example, we fix only the number of nodes and edges
    - ✓ We choose m pairs of the n vertices uniformly at random to connect them
      - Typically we stipulate simple graph

- **We refer to this random graph model as G(n,m)**

- **We can also define the model by saying that the network is created by choosing uniformly at random a network among all sets of simple graphs with exactly n nodes and m edges**

Random graph

- **Strictly the model of random graph is not defined by one specific network instance**
  - It is an *ensemble* of networks
    - ✓A probability distribution over possible networks

- **Mathematically, G(m,n) is defined as a probability distribution P(G) over all graphs G**
  - P(G)=1/Ω, where Ω is the number of simple graphs with exactly m edges and n nodes
  - In other words, we do not want to see what happens in a given network metric at a specific instance, but rather what is its probability distribution over the ensemble G(m,n)

Random graph

- **For instance, the diameter of G(n,m), would be the diameter of a graph G, averaged over the ensemble:**

$$\langle \ell \rangle = \sum_G P(G)\ell(G) = \frac{1}{\Omega}\sum_G \ell(G)$$

- **This approach is in general convenient**
  - Analytical calculation
  - We can see the typical properties of the network model we consider
  - The distribution of many network metrics, at the limit of large n, is sharply peaked around the mean value
    - ✓ Hence in the limit of large n we expect to see behaviors very close to the mean of the ensemble

Random graph

- **For the G(n,m) model**
  - Average number of edges: m
  - Mean node degree: <k>=2m/n
  - However, other properties are hard to analytically obtain

- **We will use a slightly different model, G(n,p)**
  - The number of nodes is fixed
  - Furthermore, we fix the probability p, that every possible edge between the n nodes appears in the graph
  - Note: the number of edges in this network is **not** fixed

Random graph

- **G(n,p) is the ensemble of all networks with n vertices in which a simple network G having m edges appears with probability**

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

  - m is the number of edges in G
  - For non simple graphs G, P(G)=0

- **G(n,p) is also referred in the literature as "Erdos-Renyi random graph", "Poisson random graph", "Bernoulli random graph" etc.**

Mean number of edges

- **The number of graphs with exactly m edges is given from the number of possible ways to pick m pairs of vertices among all possible pairs**
  - That is: $\left(\begin{pmatrix} n \\ 2 \end{pmatrix} \right)$
                m
- **Then the probability of drawing at random a graph with m edges from the G(n,p) is:**

$$P(m) = \left(\begin{pmatrix} n \\ 2 \end{pmatrix} \right) p^m (1-p)^{\binom{n}{2}-m}$$
                m

Mean number of edges

- **Then the mean number of edges is:**

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

Why?

- **In a more intuitive thinking:**
  - Every edge has a probability p of being part of the network
  - There are in total $\binom{n}{2}$ possible edges
  - Edges appear in the graph independently
  - Hence mean value for m is $\binom{n}{2}$ p

Mean degree

- **The mean degree in a graph with m edges is 2m/n**

- **Hence the mean degree is given by:**

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

  - In the random graph model, the mean degree is denoted with c

- **Intuitively:**
  - Every node can connect to other n-1 nodes
  - Each edge exists with probability p
  - Mean degree of a node is c=p(n-1)

Degree distribution

- **We want to find the probability p_k that a node is connected to exactly k other vertices**
  - There are n-1 possible connections for a vertex
    - ✓ The probability of being connected to a specific set of k vertices is: $p^k(1-p)^{n-1-k}$
    - ✓ Accounting for all the possible sets of these k vertex among the n-1 possible neighbors we get the total probability of being connected to exactly k vertices:

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

    - ✓ G(n,p) has a binomial degree distribution

Degree distribution

- **If we let n become very large, we have seen that many networks retain a constant average degree**
  - Then c=(n-1)p→p=c/(n-1), and as n increases p becomes very small

- **By expanding the Taylor series of ln[(1-p)$^{n-1-k}$], taking the exponents and taking the limit of $\binom{n-1}{k}$ we get:**

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

  - At the limit of n→∞, the degree distribution follows a Poisson distribution
    - ✓Poisson random graph

Clustering coefficient

- **Clustering coefficient is defined as the probability that two vertices with a common neighbor are connected themselves**

- **In a random graph the probability that *any* two vertices are connected is equal to p=c/(n-1)**
  - Hence the clustering coefficient is also:

$$C = \frac{c}{n-1}$$

- **Given that for large n, c is constant, it follows that the clustering coefficient goes to 0**
  - This is a sharp difference between the G(n,p) model and real networks

Giant component

- **How many components exist in G(n,p) model**
  - p=0 → Every node is isolated → Component size = 1 (independent of n)
  - p=1 → All nodes connected with each other → Component size = n (proportional to n)

- **It is interesting to examine what happens for values of p in-between**
  - In particular, what happens to the largest component in the network as p increases?
    - ✓ **The size of the largest component undergoes a sudden change, or *phase transition*, from constant size to extensive size at one particular special value of p**

Giant component

- **A network component whose size grows in proportion to n is called giant component**

- **Let u be the fraction of nodes that do not belong to the giant component. Hence,**
  - If there is no giant component → u=1
  - If there is giant component → u<1

- **In order for a node i *not* to connect to the giant component:**
  - i needs not connect to any other node j → 1-p
  - i is connected to j, but j itself is not connected to the giant component → pu

Giant component

- **Hence, the total probability of i not being connected to giant component via vertex j is: 1-p+pu**
  - Considering all n-1 vertices through which i can connect:

$$u = (1 - p + up)^{n-1} = \left[1 - \frac{c}{n-1}(1-u)\right]^{n-1}$$

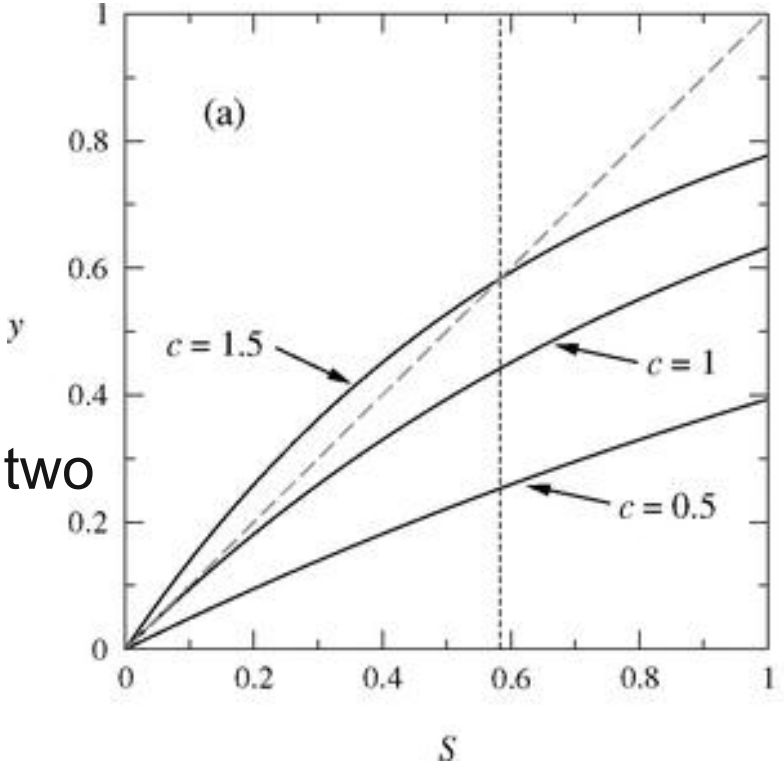
  - Taking the logarithms at both sides and the Taylor approximation for large n:

$$u = e^{-c(1-u)} \xrightarrow{S=1-u} S = 1 - e^{-cS}$$

  - This equation cannot be solved in closed form

Giant component

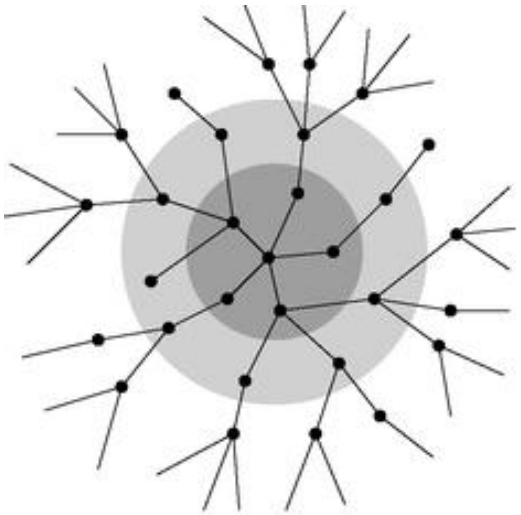
- **We plot $y=1-e^{-cS}$ with S between 0 and 1 (since it represents fraction of nodes)**
- **We also plot y=S**
- **The point where the two curves intersect is the solution**
- **For small c only one solution**
  - S=0
- **For greater c there might be two solutions**
  - The point where two solutions start appearing is when the gradients of the two curves are equal at S=0
    - ✓This happens for c=1

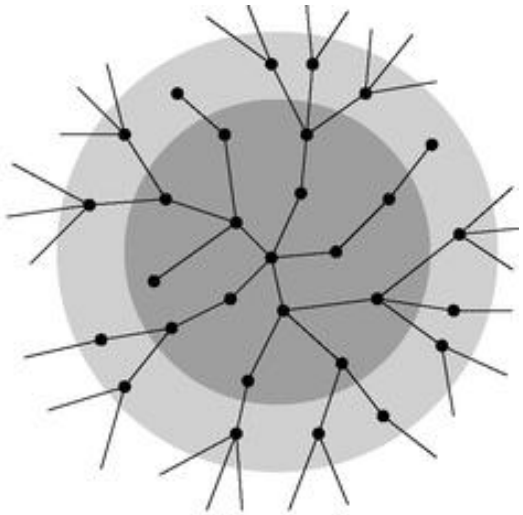


Giant component

- **Until now we have proved that if c≤1 there cannot be any giant component**
  - However, we have not proved what happens if c>1
    - ✓How can we be sure that there is a giant component?
    - ✓After all there are two solutions to the equation; one for S=0 and one for larger S

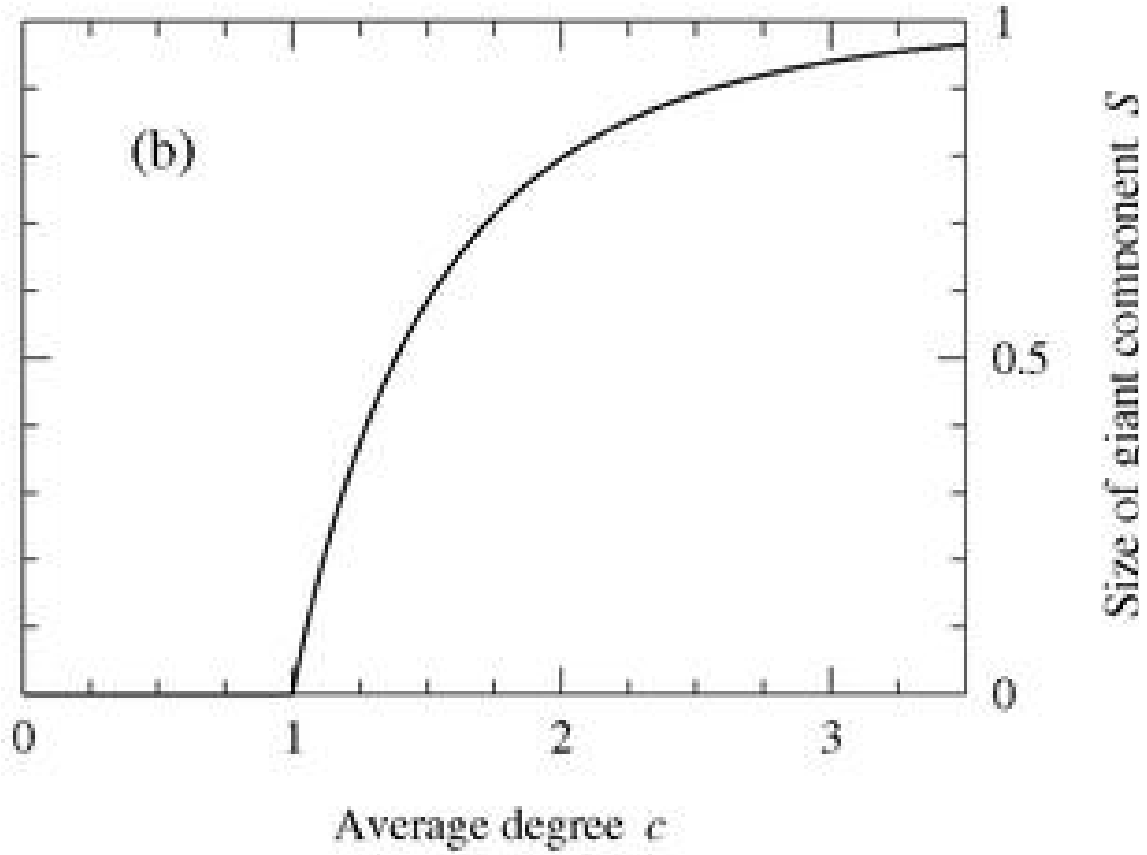
The periphery of this initially small component can only increase if c>1





Giant component

- **We expect a giant component iff c > 1**



Small components

- **The giant component of a random graph typically fills up a very large portion of the graph vertices but not 100% of them**

- **How are the rest of the nodes connected?**
  - Many small components whose average size is constant with the size of the network

- **Let us first show that there can be only one giant component in a G(n,p) graph → all other components are non-giant (i.e., small)**

Small components

- **Let us assume that there are two or more giant components**
  - Consider two of them, each with S_1n and S_2n nodes
  - There are $S_1nS_2n=S_1S_2n^2$ pairs (i,j) where i belongs to the first giant component and j to the second

- **In order for the two components to be separate there should not be an edge between them. This happens with probability:**

$$q = (1-p)^{S_1S_2n^2} = \left(1 - \frac{c}{n-1}\right)^{S_1S_2n^2}$$

Small components

- **Again, taking the logarithm in both sides and letting n become large we get:**

$$q = e^{c((c/2)-1)S_1S_2} \cdot e^{-cS_1S_2n}$$

  - The first term is constant if c is constant as n increases
  - Hence, the probability of the two components being separated decreases exponentially with n

- **From this it follows that there is only one giant component in a random graph and since this does not contain all the vertices of the graph, there should be small components as well**

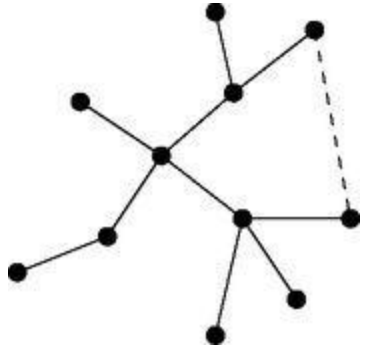
Small components

- **π_s is the probability that a randomly chosen vertex belongs to a small component of size s in total:**

$$\sum_{s=0}^{\infty} \pi_s = 1 - S$$ Why not 1?

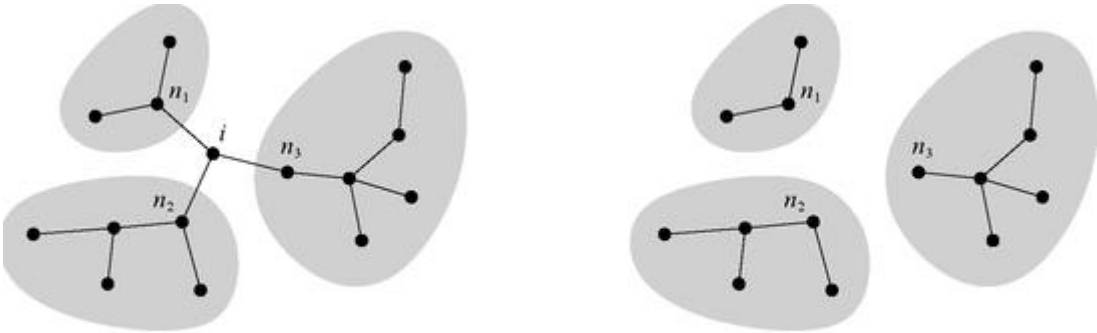
- **In order to calculate the above probability we will make use of the fact that <u>small components are trees</u>**
  - Consider a small component of size s that forms a tree
    - ✓ There are s-1 edges in total
  - If we add one more edge there will be a loop
    - ✓ Number of new edges possible: $\binom{s}{2} - (s-1) = \frac{1}{2}(s-1)(s-2)$
    - ✓ Considering the probability of existence for these edges
      - The total number of extra edges in the component is

$$\frac{1}{2}(s-1)(s-2)\frac{c}{n-1} \xrightarrow{n\to\infty} 0$$



Small components

- **Since a small component is tree any vertex i (with degree k) in the component is the "connection" of *k* sub-graphs in the same component**
  - The component size is equal to the sum of the subgraphs i's vertices lead to (plus 1 – node i)
- **Assume now we remove vertex i from the network**
  - For large network this does not really change any of the statistical properties of the network (*cavity* method)
  - In this graph the previous sub-graphs are the components themselves



Small components

- Therefore, the probability that neighbor n_1 belongs to a component of size s_1 is itself π_{s1}

- The probability P(s|k) that vertex i belongs to a small component of size s, given that its degree is k, is given by the probability that its k neighbors belong to components with sizes that sum up to s-1 (in the new cavity network):

$$P(s|k) = \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta(s-1, \textstyle\sum_j s_j)$$

- We remove the condition using the degree distribution of the random graph:

$$\pi_s = \sum_{k=0}^{\infty} p_k P(s|k) = \sum_{k=0}^{\infty} p_k \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta(s-1, \textstyle\sum_j s_j)$$

$$= e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta(s-1, \textstyle\sum_j s_j),$$

Small components

- **The previous equation is difficult to compute due to the Kronecker delta**

- **A trick that is often used in such situations is that of <u>generating function or z-transform</u>**

$$h(z) = \pi_1 z + \pi_2 z^2 + \ldots = \sum_{s=1}^{\infty} \pi_s z^s$$

  - Polynomial function or series in z whose coefficients are the probabilities π_s
  - Given function h we can then recover the probabilities as:

$$\pi_s = \frac{1}{s!} \frac{d^s h}{dz^s}\bigg|_{z=0}$$

Small components

- **In our case, substituting the probabilities at the z-transform we get:**

$$
\begin{aligned}
h(z) &= \sum_{s=1}^{\infty} z^s \mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] \delta\left(s - 1, \textstyle\sum_j s_j\right) \\
&= \mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} \right] z^{1+\sum_j s_j} \\
&= z\mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \sum_{s_1=1}^{\infty} \cdots \sum_{s_k=1}^{\infty} \left[\prod_{j=1}^{k} \pi_{s_j} z^{s_j} \right] \\
&= z\mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \left[\sum_{s=1}^{\infty} \pi_s z^s \right]^k = z\mathrm{e}^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} \left[h(z) \right]^k \\
&= z \exp\left[c\left(h(z) - 1 \right) \right].
\end{aligned}
$$

Cauchy product of z-transform

Taylor series for exponential

Small components

- The previous equation still does not have a closed form solution for h(z), however we can use it to calculate some interesting quantities

- The mean size of the component a randomly chosen vertex belongs is given by:

$$\langle s \rangle = \frac{\sum_s s\pi_s}{\sum_s \pi_s} = \frac{h'(1)}{1-S}$$

- Computing the derivative of h(z) from the previous equation we finally get:

$$\langle s \rangle = \frac{1}{1-c+cS}$$

Small components

- **When c < 1, there is no giant component:**

$$\langle s \rangle = \frac{1}{1-c}$$

- **When c > 1, we first need to solve for S and then substitute to find <s>**

- **Note that for c = 1, we have a divergence**
  - Small components get larger as we increase c from 0 to 1
  - They diverge when the giant component appears (c=1)
  - They become smaller as the giant component gets larger (c>1)

Small components revisited

- **One interesting property of the mean value calculated before, is that the mean size of a small component does not change with an increase in the number of vertices**

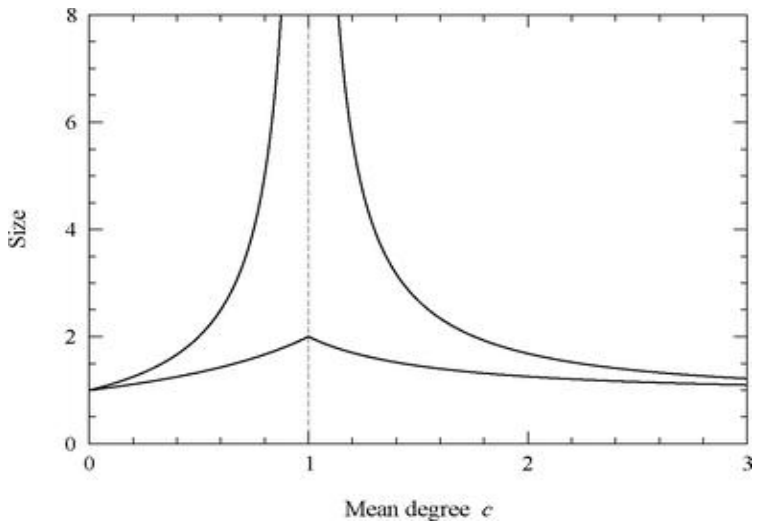
- **Is the equation computed giving the mean size of a small component ?**
  - Not exactly!
    - ✓ Recall that π_s is the probability that a randomly selected vertex belongs to a component of size s
      - It is not the probability of the size of a small component being s
    - ✓ Given that larger components have more vertices → the probability of randomly choosing a node in a larger component is higher
      - Hence, the previous computations are biased

Small components revisited

- **Let n_s be the number of components of size s in the network**
  - The probability of a randomly chosen vertex belonging to such a component is:
    $$\pi_s = \frac{s n_s}{n}$$
  - The average size of the component is:
    $$R = \frac{\sum_s s n_s}{\sum_s n_s} = \frac{n \sum_s \pi_s}{n \sum_s \frac{\pi_s}{s}} = \frac{1 - S}{\sum_s \frac{\pi_s}{s}}$$
  - Note that:
    $$\int_0^1 \frac{h(z)}{z} dz = \sum_{s=1}^{\infty} \pi_s \int_0^1 z^{s-1} dz = \sum_{s=1}^{\infty} \frac{\pi_s}{s}$$

Small components revisited

- **Finally we get:** $$R = \frac{2}{2 - c + cS}$$
  - Note that this unbiased calculation of the average size of a small component still does not depend on n
- **R does not diverge for c = 1**
  - While the largest component in the network becomes infinite, so does the number of the components



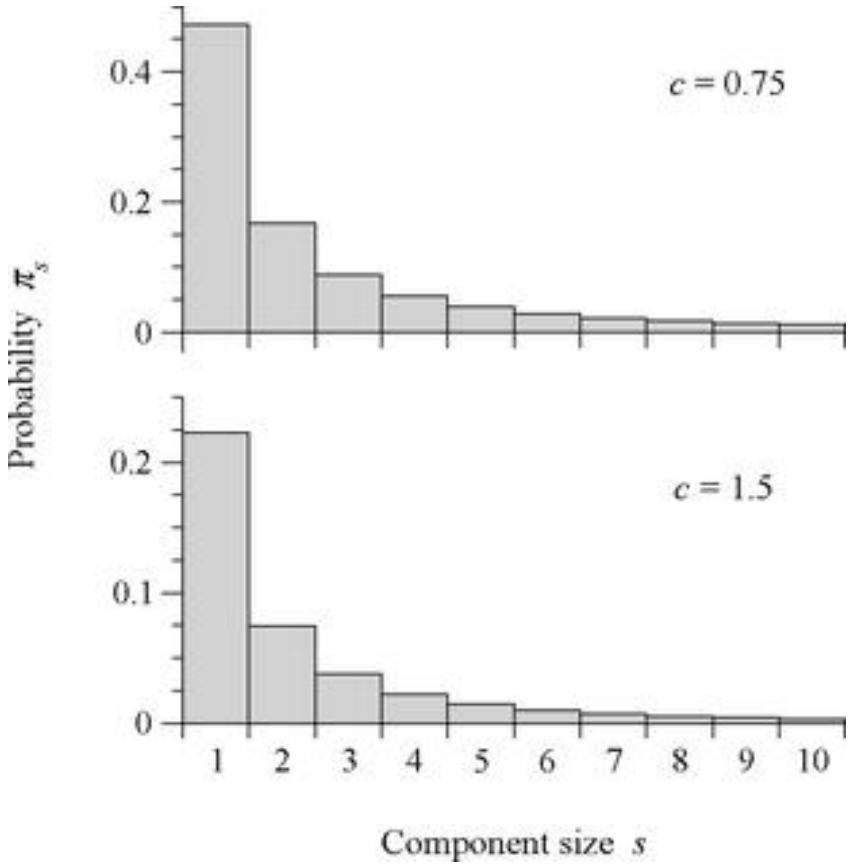
Distribution of component sizes

- **In order to compute the individual probabilities π_s we cannot rely on the z-transform and its derivatives, since we cannot solve directly for h(z)**

- **We can use though the Lagrange inversion formula**
  - Allows to solve explicitly equations of the form: $f(z)=z\varphi(f(z))$
  - In our case we have:
    - ✓ $f(z) \rightarrow h(z)$
    - ✓ $\Phi(f) \rightarrow e^{c(h-1)}$

Distribution of component size

$$\pi_s = \frac{1}{s!}\left[\frac{d^{s-1}}{dh^{s-1}} e^{sc(h-1)}\right]_{h=0} = \frac{e^{-sc}(sc)^{s-1}}{s!}$$



Threshold functions

- **This behavior – sharp dichotomy between two states – is true for a variety of network properties for random networks**
  - Phase transitions

- **Typically we examine whether the probability of the property under consideration tends to 1 or 0 as the size of the network *n* approaches infinity**
  - Most of the cases it is hard to calculate the exact probability for a fixed *n*
  - We also index the link formation probability with the population size, p(n)

Threshold functions

- **A property is generally specified as a set of networks for each *n* that satisfy the property**

- **Then a property is a list A(V) of networks that have the property when the set of nodes is V**
  - For example, the property that a network has no isolated nodes is: $A(V)=\{g|k_i \neq 0, \text{ for all } i \text{ in } V\}$

- **Most properties that are studied are *monotone***
  - If a given network satisfies them, then so is any supernetwork (in the sense of set inclusion)
    - ✓A is monotone if $g \in A(V) \;\; and \;\; g \subset g' \Rightarrow g' \in A(V)$

Threshold functions

- **For a given property A(V), t(n) is a *threshold function* iff**

$$\Pr[A(V) \mid p(n)] \to 1 \quad if \quad \frac{p(n)}{t(n)} \to \infty$$

$$\Pr[A(V) \mid p(n)] \to 0 \quad if \quad \frac{p(n)}{t(n)} \to 0$$

- **When such a threshold function exists, it is said that a phase transition occurs at that threshold**

Threshold functions

- $t(n)=1/n^2$ is a threshold function for the network to have at least one link
- $t(n)=n^{-3/2}$ is a threshold function for the network to have at least one component with at least 3 nodes
- $t(n) = 1/n$ is a threshold function for the network to have a giant component
- $t(n) = \log(n)/n$ is a threshold function for the network to be connected

Threshold functions

- **What is the threshold function for the degree of a node to be at least 1?**
  - The probability that the node has no links is $(1-p(n))^{n-1}$
  - Hence, the probability that the property holds is $1-(1-p(n))^{n-1}$
  - In order to derive the threshold function we need to find for which p(n) the above expression tends to 0 and for which it tends to 1
  - Consider t(n)=r/(n-1)

$$\frac{p(n)}{t(n)} \to 0 \Rightarrow p(n) \geq \frac{r}{n-1} \Rightarrow \lim_n (1-p(n))^{n-1} \leq \lim_n (1-t(n))^{n-1} = e^{-r}, \forall r \Rightarrow \lim_n (1-p(n))^{n-1} = 0$$

$$\frac{p(n)}{t(n)} \to 0 \Rightarrow p(n) \leq \frac{r}{n-1} \Rightarrow \lim_n (1-p(n))^{n-1} \geq \lim_n (1-t(n))^{n-1} = e^{-r}, \forall r \Rightarrow \lim_n (1-p(n))^{n-1} = 1$$

Path lengths

- **We will show that the diameter increases logarithmically with the number of nodes in the network**

- **Let us see first the intuition behind this result**
  - The number of nodes s steps away from a randomly chosen node is c^s (why?)
  - The above quantity increases exponentially and hence, it does not take many steps to reach the whole network

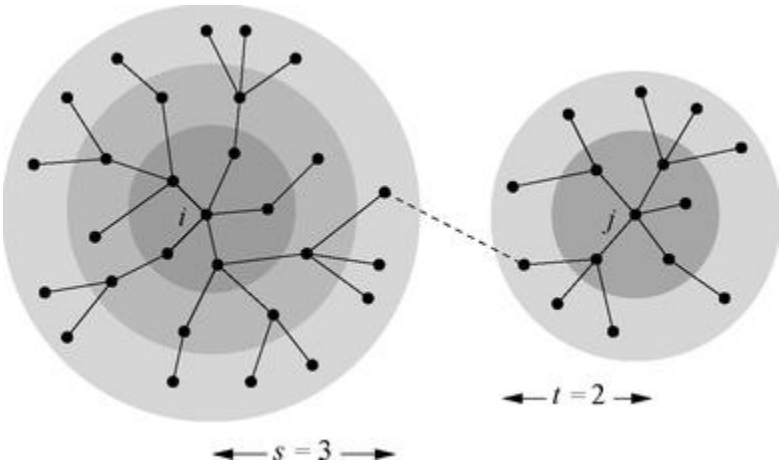
$$c^s \cong n \Rightarrow s \cong \frac{\ln n}{\ln c}$$

  - This intuition holds for every network – not only random

Path lengths

- **The above argument, while intuitively correct, it does not formally apply**
  - When s is taking large values such as c^s is comparable to n the argument essentially breaks
    - ✓A small increase in the step size s will drastically increase c^s and cause it become greater than n (which cannot happen)

- **To avoid this approximation let us consider two random nodes i and j and two regions around them that the above holds**



Path lengths

- **The absence of an edge between the surfaces of the neighborhoods of nodes i and j is a necessary and sufficient condition for the distance d_{ij} between i and j to be greater than s+t+1**
  - Hence, $P(d_{ij}>s+t+1)$ is equal to the probability that there is no edge between the two surfaces

$$P(d_{ij} > s+t+1) = (1-p)^{c^{s+t}} \xrightarrow{\ell=s+t+1} P(d_{ij} > \ell) = (1-p)^{\ell-1} = (1-\frac{c}{n-1})^{c^{\ell-1}} \cong (1-\frac{c}{n})^{c^{\ell-1}}$$

  - Taking the logarithm and Taylor approximation for large n we get:

$$P(d_{ij} > \ell) = e^{-\frac{c^{\ell}}{n}}$$

Path lengths

- **The diameter of the graph is the smallest l for which $P(d_{ij}>l)$ is zero**

- **In order for this to happen, c^l needs to grow faster than n, i.e., $c^l=an^{1+\varepsilon}$, $\varepsilon \rightarrow 0^+$**
  - We can achieve this by keeping c^s and c^t separately, smaller than O(n)

$$\ell = A + \frac{\ln n}{\ln c}$$

More detailed realistic derivations can be found at:
D. Fernholz and V. Ramachandran, "The diameter of sparse random graphs", Random Struct. Alg. 31, 482-516 (2007)

Problems with random Poisson graphs

- **Random Poisson graphs have been extensively studied**
  - Easily tractable

- **Major shortcomings make them inappropriate as a realistic network model**
  - No transitivity
  - No correlation between the degrees of adjacent vertices
  - No community structure
  - Degree distribution differs from real networks