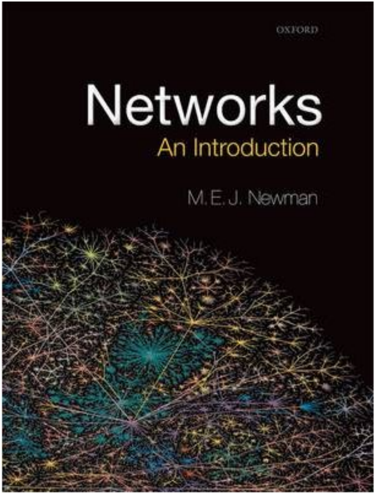
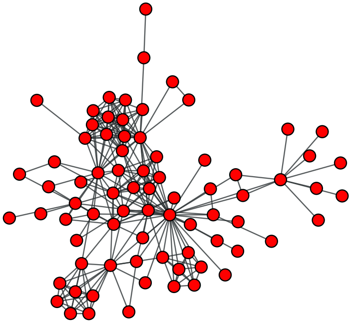
Large-scale structure of complex networks (Part 1)

Snehal M. Shekatkar

Centre for modeling and simulation,
S.P. Pune University, Pune


OXFORD
Networks
An Introduction
M. E. J. Newman

Networks/Graphs



- Points connected by lines
- Points: nodes/vertices/actors
- Lines: links/edges/ties

Real-world networks

- **Social**: Facebook, Friendships, Scientific collaborations
- **Biological**: Human brain, Metabolic reactions
- **Technological**: Internet, World-Wide-Web
- **Transport**: Airports-Air routes, Cities-Highways

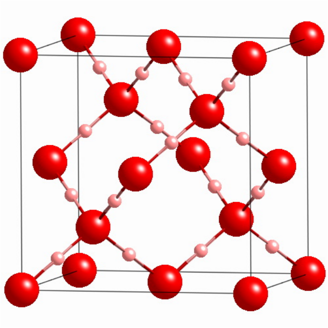
Two-questions

What are the nodes?

What are the links?

Is this a network?

Is this a network?



Is this a network?

When is a network description useful?

When is a network description useful?

- Sparse data

When is a network description useful?

- Sparse data
- Lack of regularity

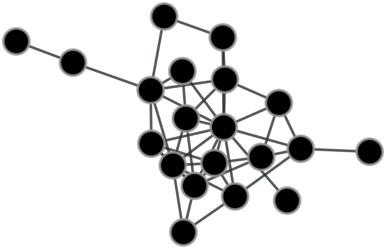
When is a network description useful?

- Sparse data
- Lack of regularity
- Lack of a better model

Adjacency matrix

Adjacency matrix: Binary matrix of size $N \times N$

$$A_{ij} = \begin{cases} 1 & i \text{ and } j \text{ connected} \\ 0 & i \text{ and } j \text{ not connected} \end{cases}$$



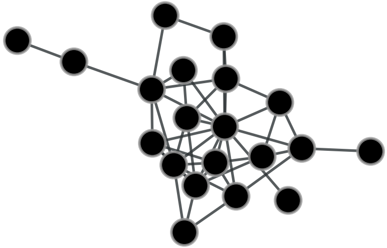
Degree

Degree of a vertex:

$$k_i = \sum_j A_{ij}$$

Average degree of the network:

$$< k >= \frac{1}{N} \sum_i k_i$$



Complex networks

Complex networks

- **Complex**: Edge of order and randomness

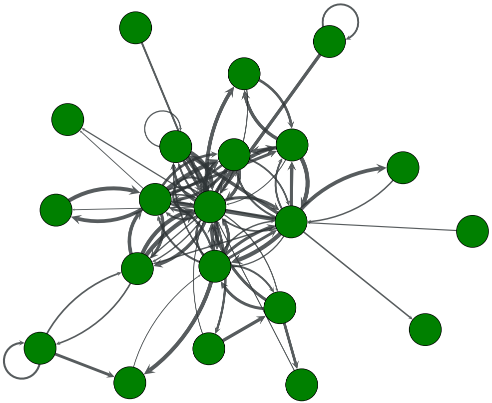
Complex networks

- **Complex**: Edge of order and randomness
- **Structure vs Processes**
  - Spreading of epidemics, rumors, ideas
  - Traffic
  - Neuronal dynamics

Complex networks

- **Complex**: Edge of order and randomness
- **Structure vs Processes**
  - Spreading of epidemics, rumors, ideas
  - Traffic
  - Neuronal dynamics
- **Structure is intersting on its own!**

Simplifications

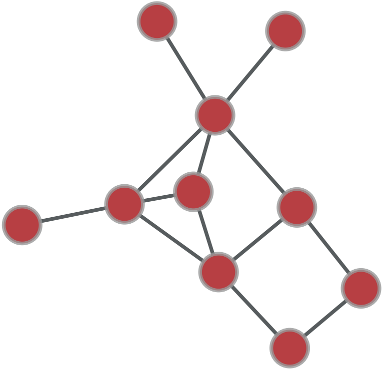


- Simple
- Undirected
- Unweighted
- Static

Large-scale structure of complex networks

- **Small-scale structures**:
  - degree
  - local clustering
  - centrality scores
- **Meso-scale structures**:
  - motifs
  - vertex similarity
  - rich-club effect
- **Large-scale structures**:
  - components and percolation
  - small-world effect
  - ranking
  - degree distribution
  - assortative mixing
  - community structure

Degree-distrbution



Total 10 vertices

$$p_1 = \frac{3}{10}$$

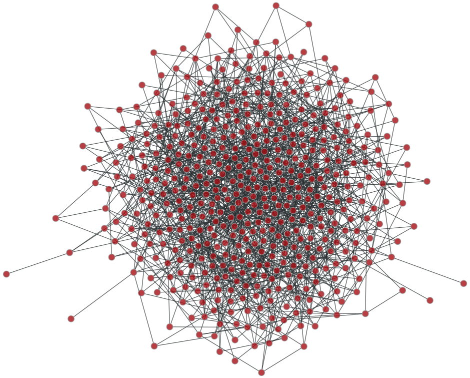
$$p_2 = \frac{2}{10}$$

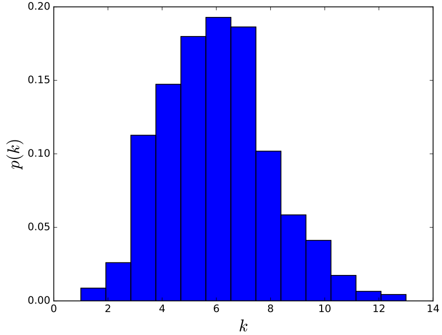
$$p_3 = \frac{2}{10}$$

$$p_4 = \frac{2}{10}$$

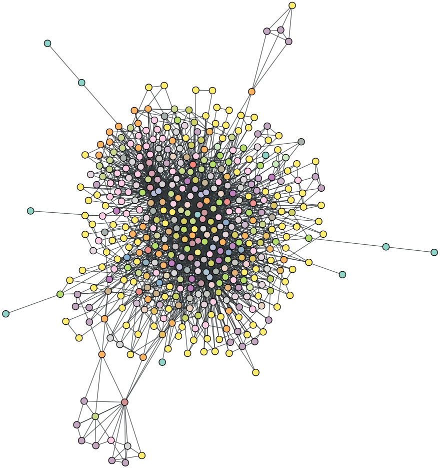
$$p_5 = \frac{1}{10}$$

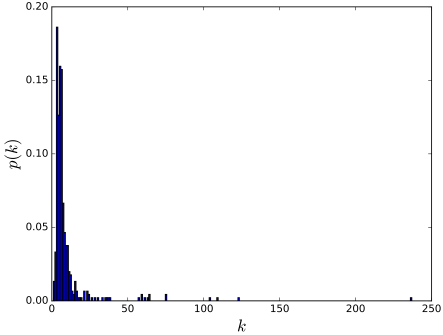
Degree distribution



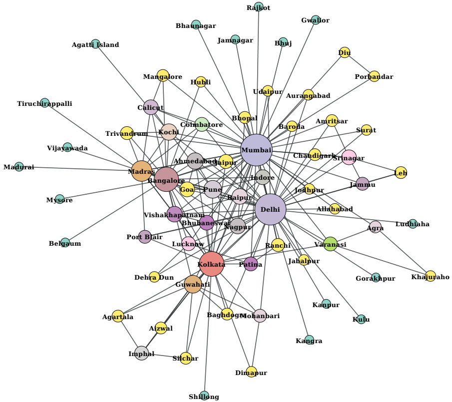


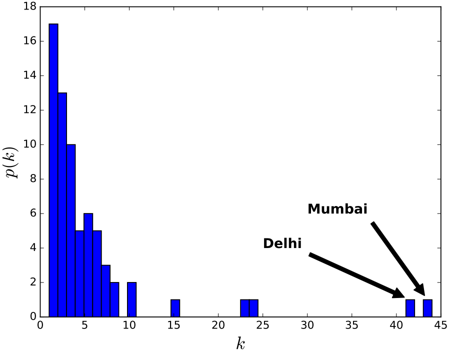
Metabolic network of the worm C-elegans



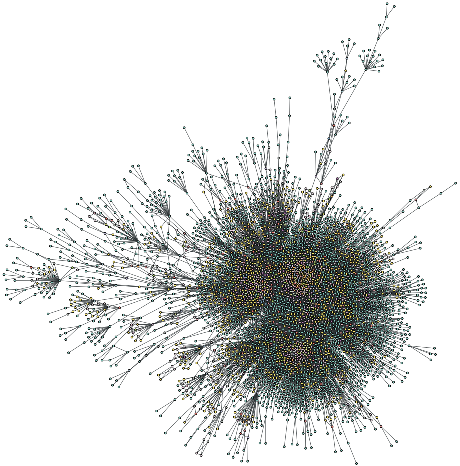


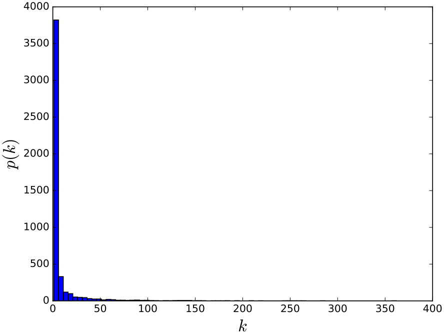
Degree distribution of the real world networks



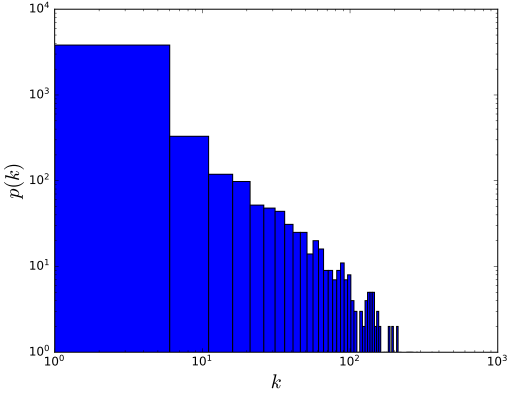


Degree distribution of the real-world networks





Power-laws and scale-free networks



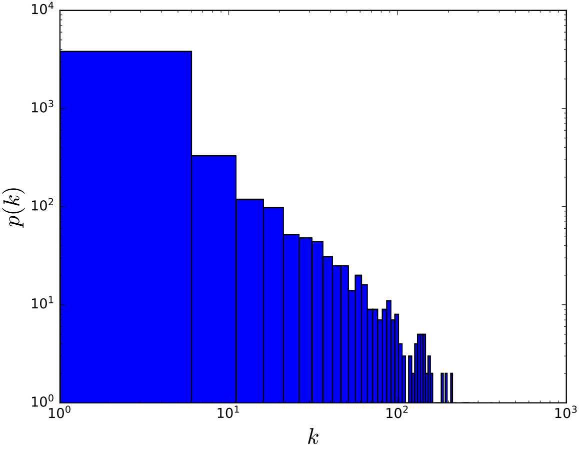
$$\ln p(k) = -\alpha \ln k + c$$

$$p(k) = Ck^{-\alpha} \quad \forall\, k > k_{\min}$$

Detecting and visualizing power-laws

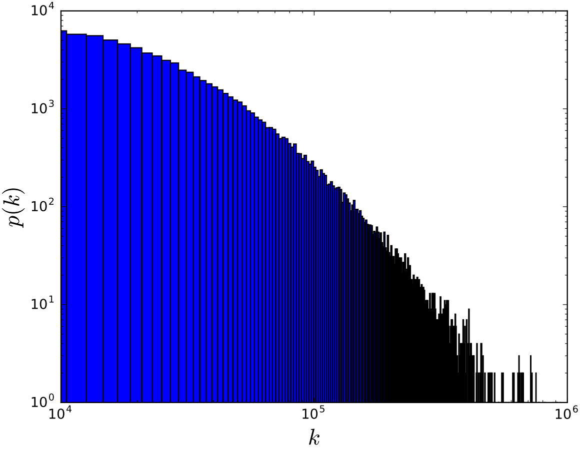
How do we know that a given distribution is a power-law?
(For $k > k_{\min}$)

- Creating a log-log plot



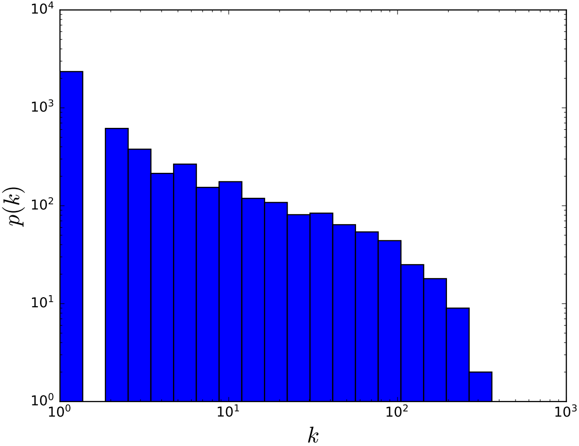
Detecting and visualizing power-laws

Power-law is tricky!



Detecting and visualizing power-laws

- Logarthmic binning: next bin is fixed multiple wider than the previous one
- Better but still noisy



Detecting and visualizing power-laws

Cumulative distribution

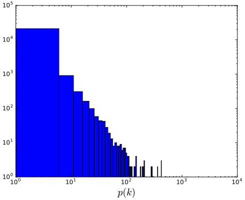
$$P(k) = \sum_{k'=k}^{\infty} p(k')$$

$$P(k) = C \sum_{k'=k}^{\infty} k'^{-\alpha} \approx C \int_{k}^{\infty} k'^{-\alpha} dk' = \frac{C}{\alpha - 1} k^{-(\alpha-1)}$$

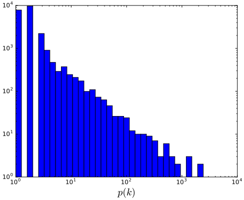
Detecting and visualizing power-laws

A portion of the internet [1]

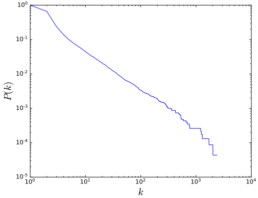
Log-log scale



Logarthmic bins



Cumulative



[1] Taken from the website of Mark Newman

Calculation of the scaling exponent

$$\alpha = 1 + \frac{N}{\left[\sum_i \frac{k_i}{k_{\min} - \frac{1}{2}}\right]}$$

Statistical error on α

$$\sigma = \frac{\alpha - 1}{\sqrt{N}}$$

Validating power-laws

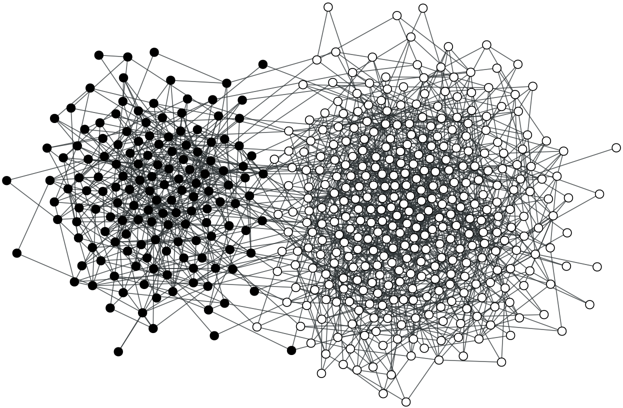
“Power-law distributions in empirical data.”

Clauset, Shalizi, and Newman. SIAM review 51.4 (2009): 661-703.

Assortative mixing

Assortative mixing

Social network of school-children with two races: Black and White



Assortative mixing

- **Social networks:** race, age, physical locations, language, income, educational level
- **Citation networks:** topics of the study
- **World Wide Web:** contents of the webpages
- **Internet:** physical locations

Assortative mixing by enumerative characteristics

- Characteristics with a finite set of values
- No ordering
- Nationality, Gender, Race

Assortative mixing by scalar characteristics

- Characteristics with a finite or infinite set of values
- Ordering
- Age, income, degree

Assortative mixing by enumerative characteristics

- Characteristics with a finite set of values
- No ordering
- Nationality, Gender, Race

Assortative mixing by scalar characteristics

- Characteristics with a finite or infinite set of values
- Ordering
- Age, income, degree

The network is **assortative** if a large fraction of the edges fall between vertices of the same type

If the opposite is true, the network is called **dissortative**

Quantification for enumerative characteristics

Quantification for enumerative characteristics

- Fraction of edges connecting vertices of the same type?

Quantification for enumerative characteristics

- Fraction of edges connecting vertices of the same type?
- Maximize: Actual number of edges between the same type minus the number of expected edges between the same types

Quantification for enumerative characteristics

- Fraction of edges connecting vertices of the same type?
- Maximize: Actual number of edges between the same type minus the number of expected edges between the same types
- Has value 0 in trivial cases

Number of edges between the same types

c_i : the class or type of vertex i

n_c : total number of types

Total number of edges between the vertices of the same type:

$$\sum_{\text{edges}(i,j)} \delta(i,j) = \frac{1}{2} \sum_{i,j} A_{ij} \delta(c_i, c_j)$$

Expected number of edges between the same types

- Half-edges or stubs, degrees preserved
- For a given stub at vertex i, there are $2m - 1$ stubs to which it can connect to
- Probability of connecting vertex j is $\frac{k_j}{2m}$
- Expected number of edges between i and j is $\frac{k_i k_j}{2m-1}$
- Expected number of edges between all the pairs of the same type:

$$\frac{1}{2}\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j)$$

$$\frac{1}{2}\sum_{i,j} A_{ij}\delta(c_i, c_j) - \frac{1}{2}\sum_{ij} \frac{k_i k_j}{2m}\delta(c_i, c_j) = \frac{1}{2}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$$

$$\frac{1}{2}\sum_{i,j} A_{ij}\delta(c_i, c_j) - \frac{1}{2}\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j) = \frac{1}{2}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$$

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)$$

Q is called the **modularity** of the network w.r.t. to c

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

Normalized modularity

Modularity is not 1 even for a perfectly mixed network.

Normalized modularity

Modularity is not 1 even for a perfectly mixed network.

$$Q_{\max} = \frac{1}{2m}\left(2m - \sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j)\right)$$

Normalized modularity

Modularity is not 1 even for a perfectly mixed network.

$$Q_{\mathrm{max}} = \frac{1}{2m}\left(2m - \sum_{ij}\frac{k_i k_j}{2m}\delta(c_i, c_j)\right)$$

$$Q_{\mathrm{norm}} = \frac{Q}{Q_{\mathrm{max}}}$$

Quantification for scalar characteristics

x_i : a scalar value for vertex i

$$r = \frac{\sum_{ij} (A_{ij} - k_i k_j / 2m) x_i x_j}{\sum_{ij} (k_i \delta_{ij} - k_i k_j / 2m) x_i x_j}$$

Degree-correlations/Degree-assortativity

- Using degree itself as a scalar property of the nodes
- Degree is the property of the network structure
- One property (degree) dictating the others (position of the edges)

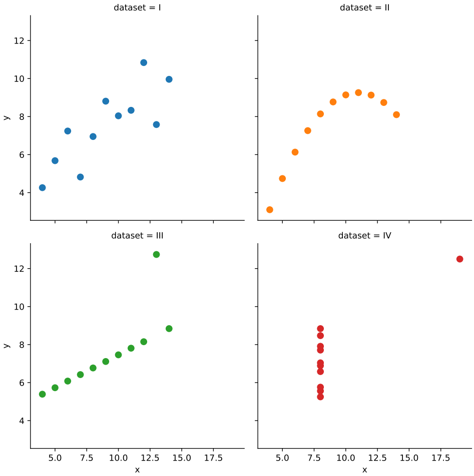
Network	n	r
Physics coauthorship (a)	52 909	0.363
Biology coauthorship (a)	1 520 251	0.127
Mathematics coauthorship (b)	253 339	0.120
Film actor collaborations (c)	449 913	0.208
Company directors (d)	7 673	0.276
Internet (e)	10 697	−0.189
World-Wide Web (f)	269 504	−0.065
Protein interactions (g)	2 115	−0.156
Neural network (h)	307	−0.163
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276

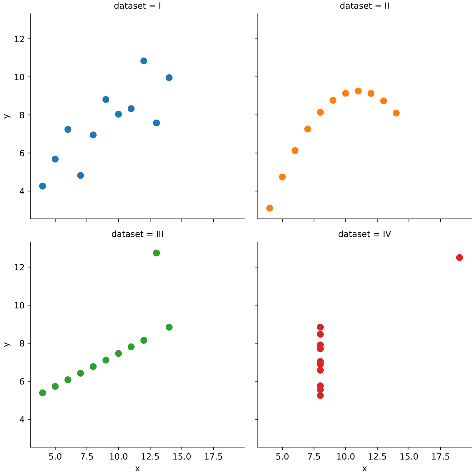
[2]

[2] Newman, M.E.J., Assortative Mixing in Networks, PRL, 89, 20.

Random graph models

- Better way to describe the structure of the networks
- Generative models of networks
- Probabilistic models capable of generating an observed data
- Certain properties can be fixed for generative models


dataset = I
dataset = II
dataset = III
dataset = IV
y
y
x
x
4
6
8
10
12
5.0
7.5
10.0
12.5
15.0
17.5



- Mean of $x = 9$
- Variance of $x = 11$
- Mean of $y = 7.5$
- Variance of $y = 4.125$
- Correlation $= 0.816$

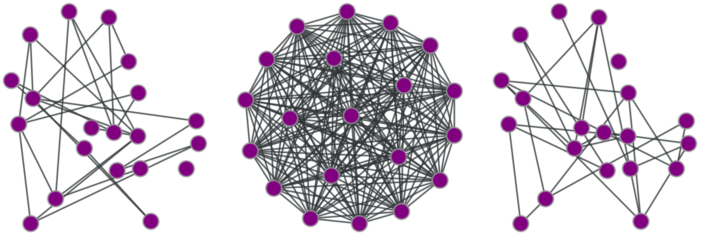
Anscombe's quartet!

Random graph models (RGMs)

- Erdős-Rényi model
- Configuration model
- Stochastic-block model
- Degree-corrected SBM
- Equitable random graphs
- Hierarchical block models
- Random graphs with expected degrees
- Microcanonical SBM
- Poisson SBM

Erdős-Rényi model (ER model/$G(n,p)$)

- Fix n and the average degree c
- Connect every pair of nodes with a probability $p = \frac{c}{n-1}$
- Number of graphs in the ensemble: $2^{\binom{n}{2}}$
- ER model: Every member of the ensemble is equally likely



Properties of the ER model

Average properties of the ensemble

$$< m >= \sum_G P(G)m(G) = \frac{1}{\Omega}\sum_G m(G)$$

Properties of the ER model

Degree distribution

- Given vertex can connect to remaining $n-1$ vertices
- Probability of connecting to particular k vertices:

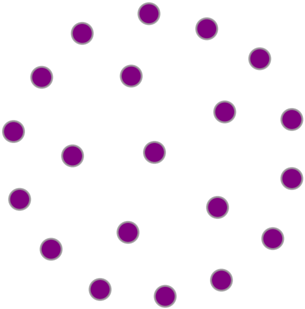
$$p^k(1-p)^{n-1-k}$$

- There are $\binom{n-1}{k}$ ways to choose k vertices

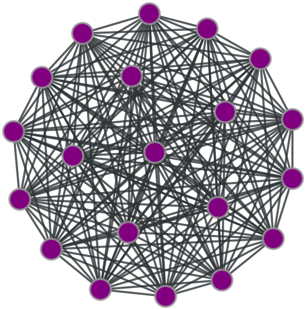
$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} = e^{-c}\frac{c^k}{k!}$$

The giant component

$p = 0$



$p = 1$



What is the size of the largest component?

The giant component

- Giant component: size is an extensive function of the network size
- Transition between the two extremes with p: gradual or sudden?
- Size of the giant component as a function of p can be calculated exactly in the limit

The giant component

u : Fraction of vertices not in the giant component

When is a vertex i not in the giant component?

For any other vertex j:

1. i is not connected to j
2. i is connected to j but j is not in the giant component

The giant component

u : Fraction of vertices not in the giant component

When is a vertex i not in the giant component?

For any other vertex j:

1. i is not connected to j $\Rightarrow$ **probability** $= 1 - p$
2. i is connected to j but j is not in the giant component $\Rightarrow$ **probability** $= pu$

Thus, the probability of not being connected to the giant component through vertex j is $(1 - p + pu)$

$$u = (1 - p + pu)^{n-1}$$

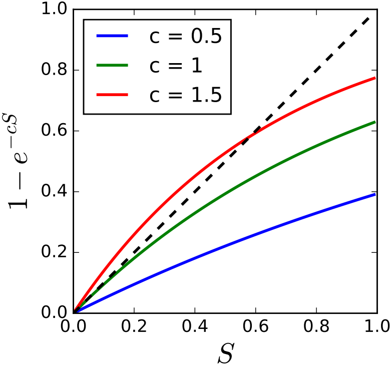
$$u = \left(1 - \frac{c}{n-1} + p\frac{c}{n-1}\right)^{n-1} = \left[1 - \frac{c}{n-1}(1-u)\right]^{n-1}$$

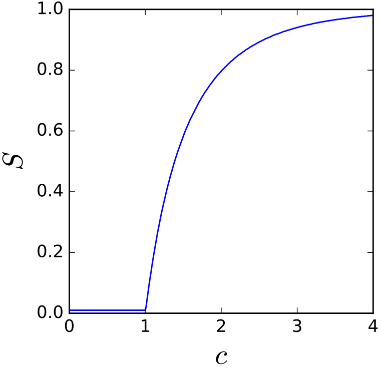
$$u = e^{-c(1-u)}$$

Let S be the size of the giant component

$$\therefore 1 - S = e^{-cS}$$

Graphical solution for the giant component



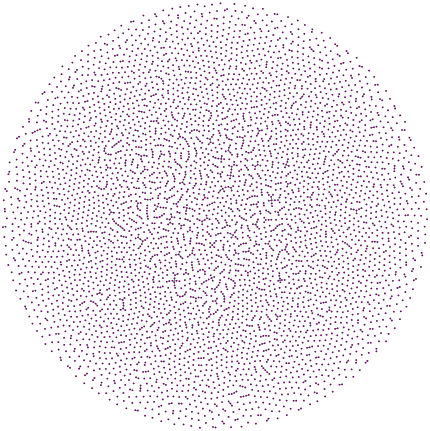


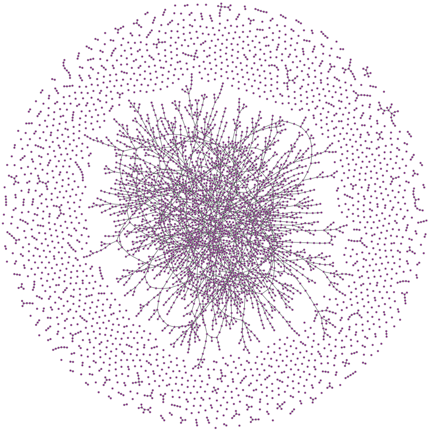
$$\frac{d}{dS}(1 - e^{-cS})\Big|_{S=0} = 1$$

$$\therefore ce^{-cS}\Big|_{S=0} = 1$$

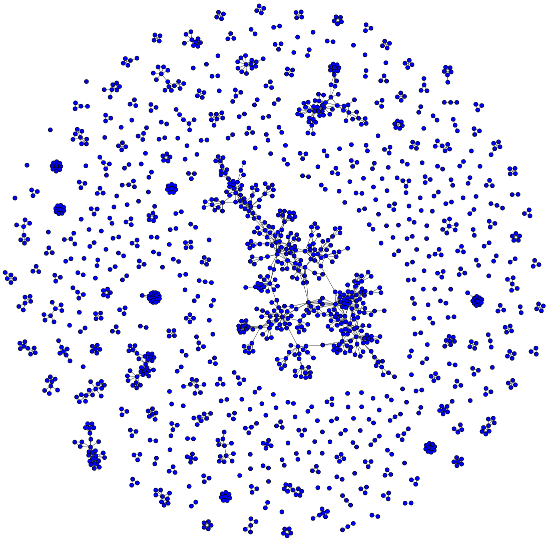
$$\therefore c = 1$$

$c < 1$ and $c > 1$

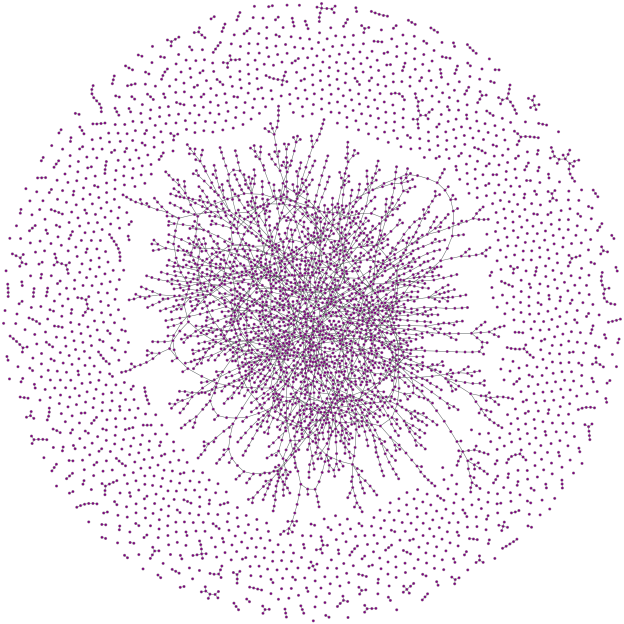




Network science researchers



Small components



Small components

Two giant components?

Distinct pairs (i, j) with i in S_1 and j in S_2 : $S_1 n \times S_2 n = S_1 S_2 n^2$

Probability that there is no edge between the two:

$$q = (1-p)^{S_1 S_2 n^2} = \left(1 - \frac{c}{n-1}\right)^{S_1 S_2 n^2} \sim e^{-c S_1 S_2 n}$$

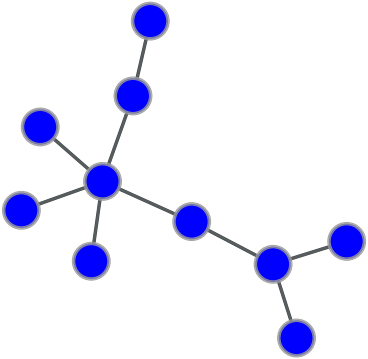
Sizes of the small components

π_s : probability that a randomly chosen vertex belongs to component of size s

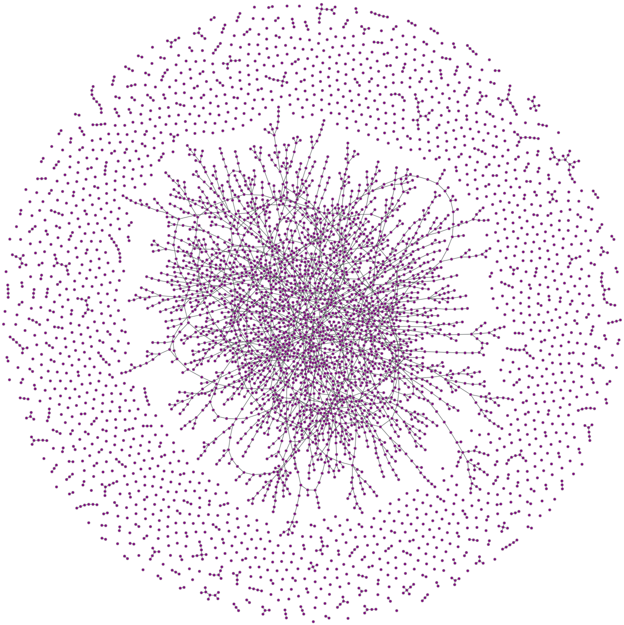
$$\sum_{s=1}^{\infty} = 1 - S$$

Small-components are trees!

Tree graph



- A graph without loops
- n vertices and $n - 1$ edges
- Removal of any vertex or edge witll disconnect the graph



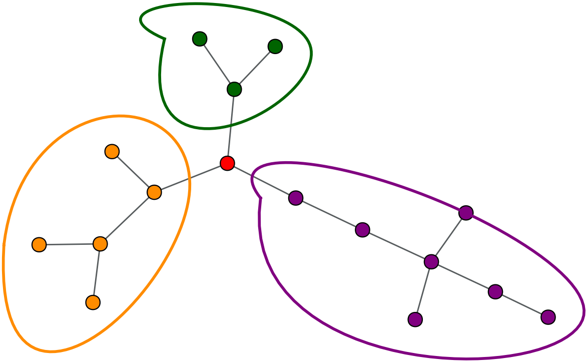
Small components

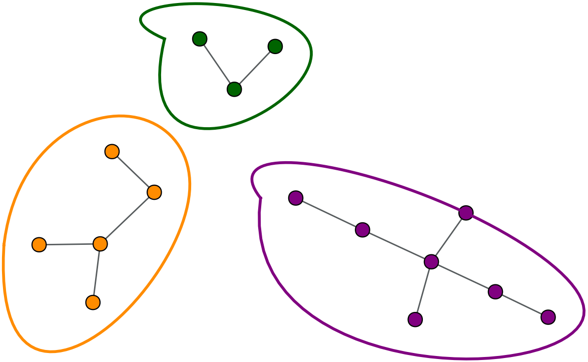
Small components are trees

$$\binom{s}{2} - (s-1) = \frac{1}{2}(s-1)(s-2)$$

The total number of extra edges:

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





For degree 3 vertex,

$$P(s|k=3) = \pi_{s_1}\pi_{s_2}\pi_{s_3}$$

$$s_1 + s_2 + s_3 = s - 1$$

For degree 3 vertex,

$$P(s|k=3) = \pi_{s_1}\pi_{s_2}\pi_{s_3}$$

$$s_1 + s_2 + s_3 = s - 1$$

$$P(s|k) = \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\cdots\sum_{s_k=0}^{\infty}\left[\prod_{j=1}^{k}\pi_{s_j}\right]\delta(s-1,\sum_j s_j)$$

For degree 3 vertex,

$$P(s|k=3) = \pi_{s_1}\pi_{s_2}\pi_{s_3}$$

$$s_1 + s_2 + s_3 = s - 1$$

$$P(s|k) = \sum_{s_1=0}^{\infty}\sum_{s_2=0}^{\infty}\cdots\sum_{s_k=0}^{\infty}\left[\prod_{j=1}^{k}\pi_{s_j}\right]\delta(s-1, \sum_j s_j)$$

$$\pi_s = \sum_{k=0}^{\infty} p_k P(s|k)$$

$$< s >= \sum_{s=0}^{\infty} s\pi_s$$

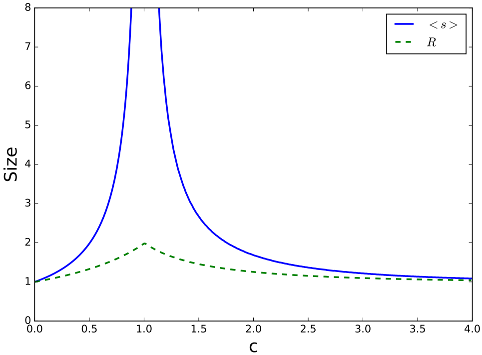
$$< s >= \frac{1}{1 - c + cS}$$

$$< s >= \frac{1}{1 - c + cS}$$

$$R = \frac{2}{2 - c + cS}$$

$$< s >= \frac{1}{1 - c + cS}$$

$$R = \frac{2}{2 - c + cS}$$

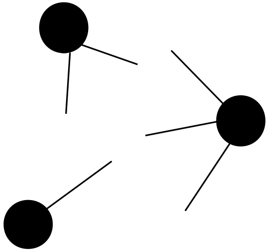


Configuration model

- Model with a given degree sequence
- More realistic and more flexible
- Can be solved exactly in the limit

Configuration model

- Specify k_i for $i = 1, 2, \cdots, n$
- Every vertex i has k_i half-edges/stubs
- $\sum_i k_i = 2m$
- Randomly connect stubs to each other
- Every matching appears with equal probability



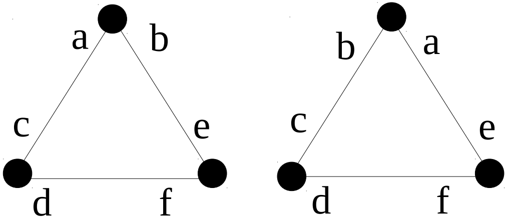
Configuration model

Do all graphs appear with equal probability?

Configuration model

Do all graphs appear with equal probability?

More than one matchings can correspond to the same graph

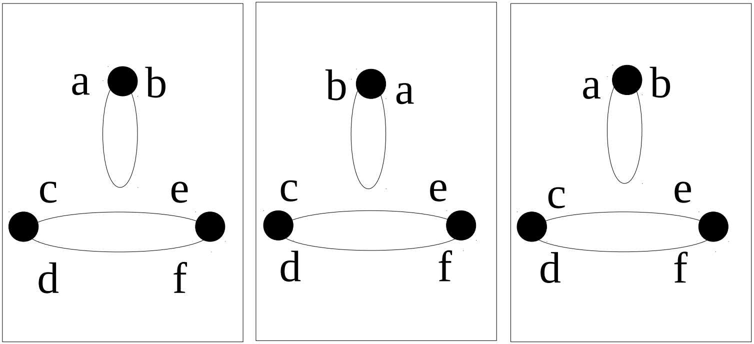


Configuration model

Number of matchings for the network N: $N(\{k_i\}) = \prod_i k_i!$

Thus, networks do occur with equal probability!

Configuration model



Configuration model

$$N = \frac{\prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!!}$$

$$n!! = n(n-2)(n-4)\cdots 2$$

Graphs do not appear with equal probability!

Summary

- Many real networks have fat-tailed degree distributions
- Naturally occuring networks tend be dissortative by degree
- Social networks tend to be assortative by degree
- Random graphs provide a powerful way to describe the large-scale structure of complex networks

`www.snehalshekatkar.com/serc2018`