Web-graphs. Lecture 3.

Bollobash - Riordan model

Our aim is to construct the random graph G_m^n having n vertices and mn edges.

Let $deg_G(v)$ be the node's v degree in graph G.

$\mathbf{m = 1}$

- Fix G_1^1 – one node and one loop
- G_1^{n-1} is a graph on $n-1$ vertices $\{v_1, \ldots, v_{n-1}\}$ having $n-1$ edge
- G_1^n is obtained from G_1^{n-1} by addition of one node and edge incident to it and to some node $i \in [1, \ldots, n]$ with the following probability:

$$P(i = s) = \begin{cases} \frac{deg_{G_1^{n-1}} v_s}{2n-1}, & \text{if } 1 \leq s \leq n-1 \\ \frac{1}{2n-1}, & \text{otherwise.} \end{cases}$$

- $\sum_{s=1}^{n} P(i = s) = 1$

Bollobash - Riordan model

$\mathbf{m > 1}$

- Graph G_m^n is obtained from G_1^{mn} by gluing together the following nodes:
  - $\{v_1, \ldots, v_m\} \longrightarrow v'_1$
  - $\{v_{m+1}, \ldots, v_{2m}\} \longrightarrow v'_2$
  - $\ldots$
  - $\{v_{(n-1)m+1}, \ldots, v_{nm}\} \longrightarrow v'_n$
- Let $v'(v_i)$ is the name of vertex in G_m^n corresponding to the group in which v_i falls.
- Edge $(v_i, v_j) \longrightarrow (v'(v_i), v'(v_j))$

Bollobash - Riordan model

Degree distribution

Theorem (Bollobas, Riordan, Spencer, Tusnady)
Let $m \geq 1$, G_m^n is the random graph in the Bollobash-Riordan model. Then for each $d \leq n^{1/15}$ with probability tending to 1 and $n\ to\infty$

$$\frac{|\{v \in G_m^n : degv = d\}|}{n} \sim \frac{2m(m+1)}{d(d+1)(d+2)}$$

Restriction on d ($d \leq n^{1/15}$) can be omitted (the result of Grechnikov).
Therefore, in the Bollobash-Riordan model we have power law with $\gamma = 3$.

In real web-graphs $\gamma \in (2, 3) \longrightarrow$ some modifications should be introduced.

Backley-Ostgus model

Our aim is to construct the random graph $H^n_{a,m}$ having n vertices and mn edges. a - parameter, $a > 0$

Parameter a is *initial attractiveness* of the node.

Let $deg_G(v)$ be the node's v degree in graph G.

$\mathbf{m = 1}$

- Fix $H^1_{a,1}$ – one node and one loop
- $H^{n-1}_{a,1}$ is a graph on $n-1$ vertices $\{v_1, \ldots, v_{n-1}\}$ having $n-1$ edge
- $H^n_{a,1}$ is obtained from $H^{n-1}_{a,1}$ by addition of one node and edge incident to it and to some node $i \in [1, \ldots, n]$ with the following probability:

$$P(i = s) = \begin{cases} \frac{deg_{H^{n-1}_{a,1}} v_s + a - 1}{(a+1)n-1}, & \text{if } 1 \le s \le n-1 \\ \frac{a}{(a+1)n-1}, & \text{otherwise.} \end{cases}$$

- $\sum_{s=1}^{n} P(i = s) = 1$

For $m > 1$ we act as in case of the Bollobas-Riordan model.

Backley-Ostgus model

The statements about

- diameter,
- stability and vulnerability of the giant component

hold as well as in the Bollobas-Riordan model.

Theorem (Backley, Ostgus)

Let $m \geq 1$, $a \geq 1$ $H^n_{a,m}$ is the random graph in the Backley-Ostgus model. Then for each $d \leq n^{1/(100(a+1))}$ with probability tending to 1 and n $to\infty$

$$\frac{|\{v \in H^n_{a,m} : degv = d\}|}{n} \sim \frac{\mathrm{const(a,m)}}{d^{a+2}}$$

Backley-Ostgus model

Theorem (Backley, Ostgus)

Let $m \geq 1$, $a \geq 1$ $H^n_{a,m}$ is the random graph in the Backley-Ostgus model. Then for each $d \leq n^{1/(100(a+1))}$ with probability tending to 1 and n $to\infty$

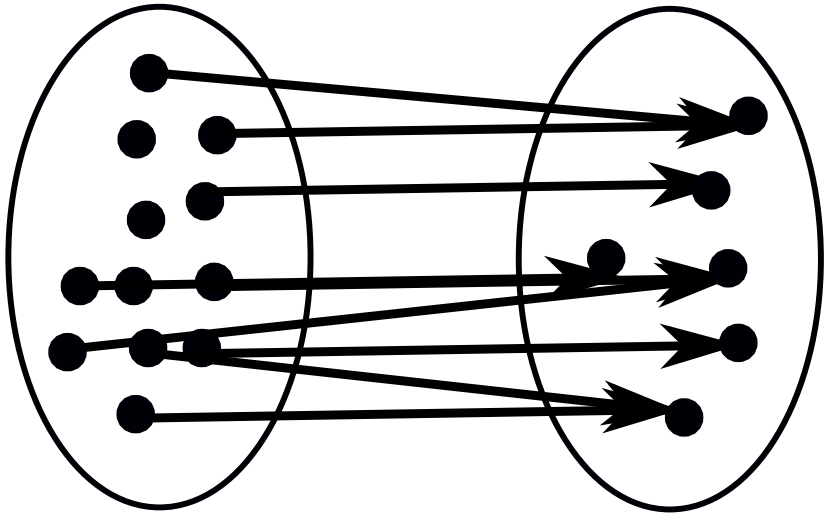
$$\frac{|\{v \in H^n_{a,m} : degv = d\}|}{n} \sim \frac{\mathrm{const(a,m)}}{d^{a+2}}$$

- We can vary the exponent of the power law by varying parameter a value.
- Restrictions on a and d can be omitted. $a > 0$

It can be shown that $a = 0.27$ is the optimal parameter value to describe the real web-graph

Link ring

Link ring is the structure in the Internet which is artificially created to grow up itself (of some other pages) rank in the search system. The mechanism is creation of tight links network.



nodes which make rating cheat

nodes which need to increase their rating

Why do we need to detect such structures?

Link ring

Why do we need to detect such structures? $\longleftarrow$ To make searching system better.

BUT! For (web) social networks appearance of such structures is a normal phenomenon.

How we can detect such objects?

1. Propose some model which properly describes probability of appearance of some properties in the real network.
2. Calculate the probability of appearance of the property in the (fragment) of real graph and in the model.
3. if the real one deviates significantly $\longrightarrow$ something is wrong.

To detect link rings the needed property is the probability to have a link between nodes of degrees k and l. Or, more accurately, the expected number of edges between the nodes with given degrees.

Link ring

Let

- $G_n = (V_n, E_n)$ is a graph,
- $d_1,\ d_2 \in \mathbb{N}$

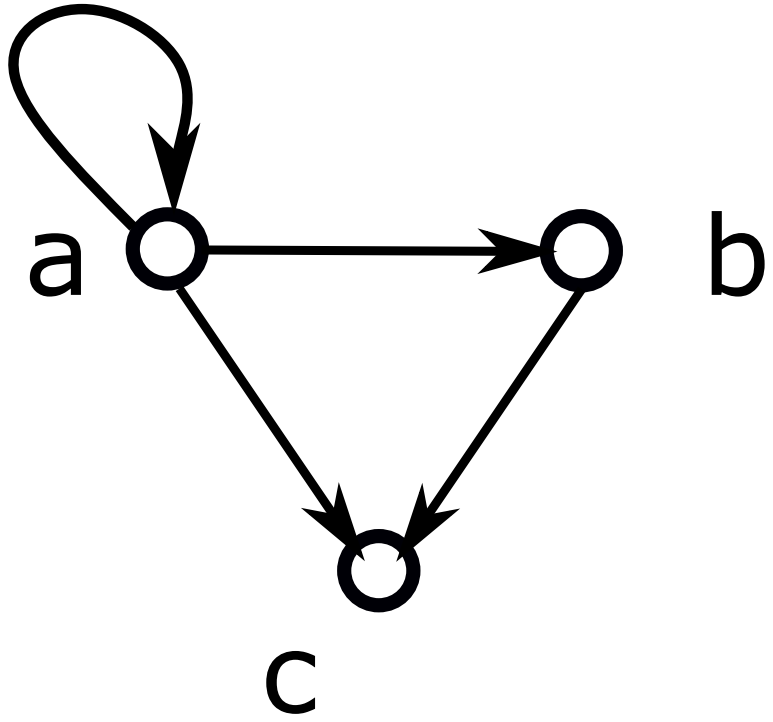
$X_n(d_1, d_2)$ is the total number of edges between all nodes having total degrees d_1 and d_2 in G_n.
It is important to note, that:

- if $d_1 = d_2$, then we should account for each edge **twice**
- each loop it considered also **two times**

$$|\{v \in V_n | \mathrm{degv} = \mathrm{d}\}| = \frac{1}{\mathrm{d}} \sum_{\mathrm{d_1}} \mathrm{X_n}(\mathrm{d_1}, \mathrm{d})$$

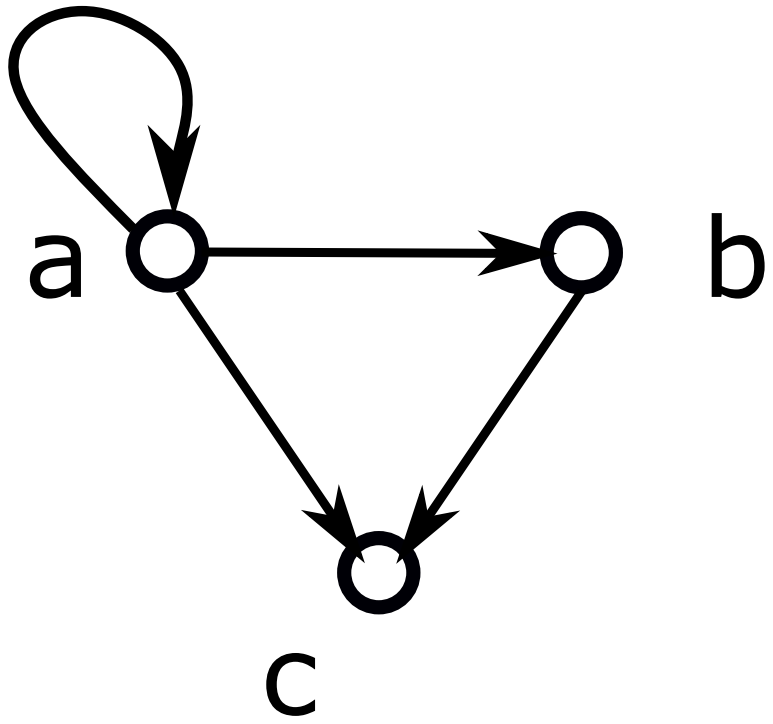
Example 1

Calculate numbers of edges between nodes with every possible combination of total degrees.



Example 1

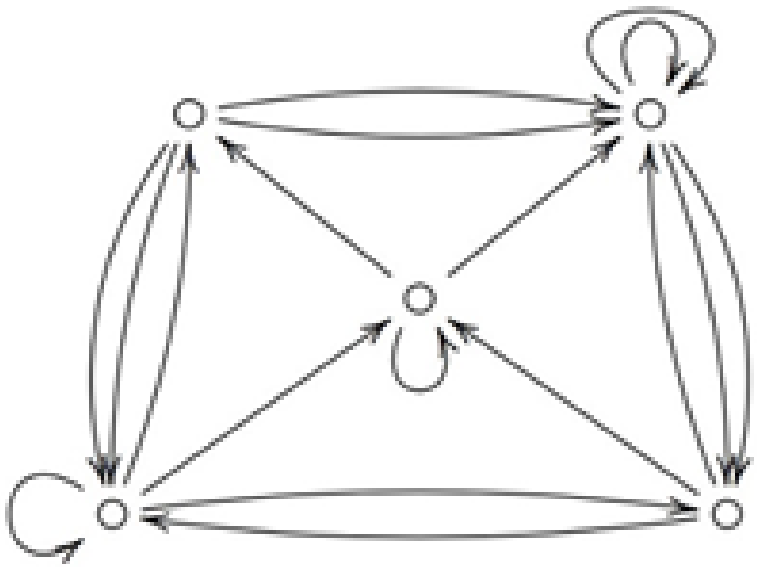
Calculate numbers of edges between nodes with every possible combination of total degrees.



- $X(4,4) = 2$
- $X(2,2) = 2$
- $X(4,2) = 2$
- $X(2,4) = 2$

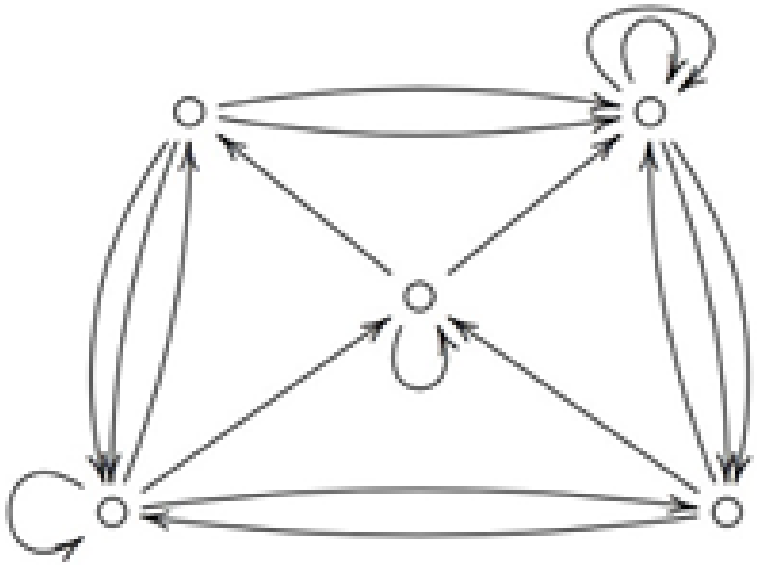
Example 2

Calculate numbers of edges between nodes with every possible combination of total degrees.



Example 2

Calculate numbers of edges between nodes with every possible combination of total degrees.



- $X(6,6) = 6$
- $X(8,6) = X(6,8) = 6$
- $X(10,6) = X(6,10) = 6$
- $X(8,8) = 2$
- $X(10,10) = 4$

Edge numbers: results

Theorem (Grechnikov)

In the Buckley-Ostgus model with probability tending to 1 while $n \to \infty$

$$\frac{X_n(d_1, d_2)}{n} \sim c(a, m)\left(\frac{(d_1 + d_2)^{1-a}}{d_1^2 d_2^2}\right)$$

Corollary

In the Bollobas-Riordan model

$$\mathbb{E}X(d_1, d_2) = \Theta\left(\frac{1}{d_1^2 d_2^2}\right)$$

In the Buckley-Ostgus model

$$\mathbb{E}X(d_1, d_2) = \Theta\left(\frac{(d_1 + d_2)^{1-a}}{d_1^2 d_2^2}\right)$$

Link spam classification

Let H is the part of the real host-graph. How can we automatically evaluate existence of link spam in it?

1. Calculate total degrees of each node in H
2. For each pair of degrees calculate expected number of edges in Buckley-Ostgus graph using the results of Grechnikov's theorem.
3. Sum up the numbers of edges obtained in the previous step.
4. Compare this number with the number of edges in H. If the last one is significantly higher $\longrightarrow$ we have spam or (web) social network community.

Degree-degree correlation

Let $G = (V, E)$ is the graph without loops, multiple edges and direction.

Let us define the mean degree of the node with degree d neighbors $d_{nn}(d)$ (nn is "nearest neighbor")

$$d_{nn}(d) = \frac{1}{d|\{i : \mathrm{degi} = \mathrm{d}\}|} \sum_{i:\mathrm{degi}=\mathrm{d}} \sum_{j:(i,j)\in E} \mathrm{degj}$$

If there are loops, multiple edges and direction, the inner summation will change. We should account for loops twice, each multiple edge is considered according its multiplicity.

$$d_{nn}(d) = \frac{\sum_{d_1} d_1 X_n(d_1, d)}{\sum_{d_1} X_n(d_1, d)}$$

Assortativity

$d_{nn}(d)$ is called *assortativity*

In real networks it is often seen that

$$d_{nn}(d) \sim d^{\delta}$$

- $\delta < 0 \longrightarrow$ the network is disassortative
- $\delta > 0 \longrightarrow$ the network is assortative

Host-graph is disassortative usually, while (web) social networks are assortative ones.

Examples

Calculate $d_{nn}(4)$ for the first example and $d_{nn}(6)$ for the second one.

Assortativity

Another way to define the degree-degree correlation (assortativity) of a graph is to calculate the Pearson correlation coefficient of the degrees at either ends of a link. Table from Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. Physics reports, 424(4-5), 175-308.

Table 2.1
Basic characteristics of a number of information/communication, biological and social networks from the real-world

Network	N	$\langle k \rangle$	L	C	γ	ν	Ref.
AS2001	11,174	4.19	3.62	0.24	2.38	<0	[8,25,80]
Routers	228, 263	2.80	9.5	0.03	2.18	>0	[8,25,80]
Gnutella	709	3.6	4.3	0.014	2.19	<0	[100]
WWW	$\sim 2 \times 10^8$	7.5	16	0.11	2.1/2.7	Unknown	[101]
Protein	2,115	6.80	2.12	0.07	2.4	<0	[86]
Metabolic	778	3.2	7.40	0.7	2.2/2.1	<0	[85]
Math1999	57, 516	5.00	8.46	0.15	2.47	>0	[81,82]
Actors	225,226	61	3.65	0.79	2.3	>0	[28,83]
e-mail	59,812	2.88	4.95	0.03	1.5/2.0	Unknown	[84]

The quantities measured are: number of vertices N, characteristic path length L, clustering coefficient C, average degree $\langle k \rangle$, exponent of the degree distribution γ, and type of correlations. All networks, except the WWW, metabolic (Escherichia coli) and e-mail networks, are undirected. The two values of γ represent, respectively, the in/out-degree exponents when the network is directed.

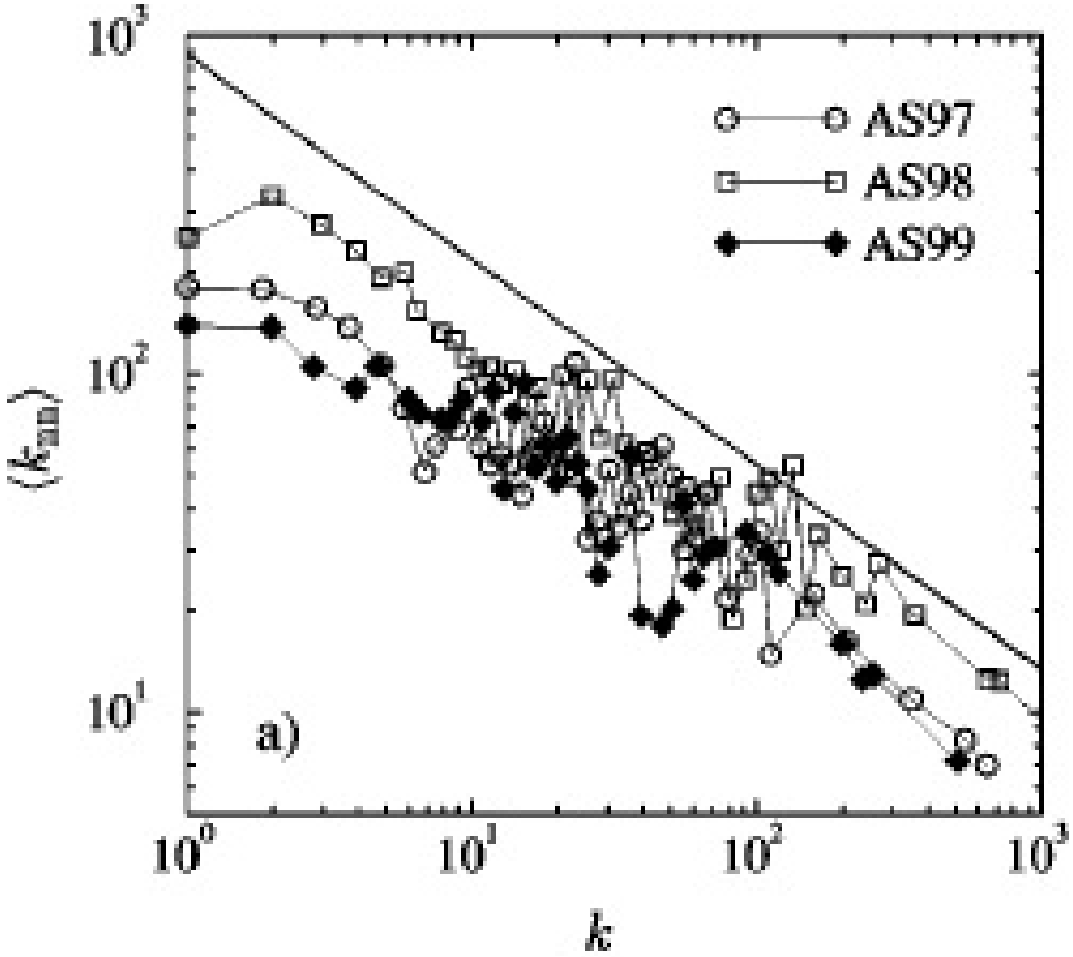
Assortativity

Table from Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. Physics reports, 424(4-5), 175-308.

The first two *information/communication networks* considered are two examples of the Internet [8,25,80]. In both networks, the nodes are the hosts and the links represent the physical connections among them. AS2001 stands for the Internet at the autonomous system (AS) level as on April 16th, 2001 [97], while Routers indicate the router level graph representation of the Internet [98]. Gnutella is a peer-to-peer [99] network provided by Clip2 Distributed Search Solutions [100], and is an example of networks that are becoming popular as a way to connect thousands of computers that allow sharing of files (for instance, music or video) between users over local and wide-area networks. Finally, the World Wide Web (WWW) is a network formed by the hyperlinks between different Web pages, and, with more than 10^8 nodes, is the largest network ever studied. Differently from the previous ones, this network is directed: each Web page

Assortativity

Vázquez, A., Pastor-Satorras, R., & Vespignani, A. (2002). Large-scale topological and dynamical properties of the Internet. Physical Review E, 65(6), 066130.



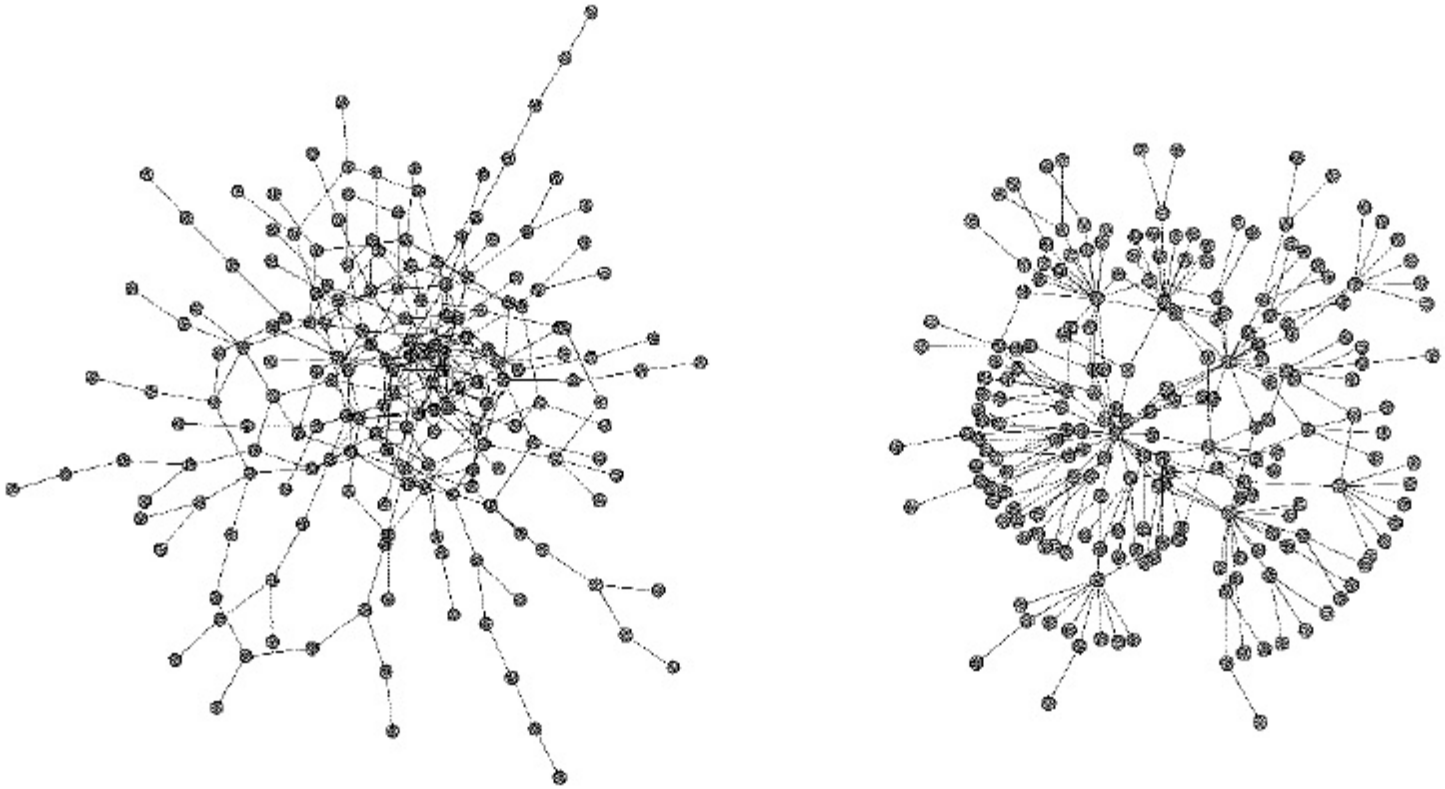
Assortativity

From M. Newman. Networks. An introduction

	Network	Type	n	m	c	S	ℓ	α	C	C_{ws}	r	Ref(s).
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3	0.20	0.78	0.208	16, 323
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–	0.59	0.88	0.276	88, 253
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–	0.15	0.34	0.120	89, 146
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–	0.45	0.56	0.363	234, 236
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–	0.088	0.60	0.127	234, 236
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1				9, 10
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0		0.16		103
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–	0.17	0.13	0.092	248
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029	34
	Sexual contacts	Undirected	2 810					3.2				197, 198
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4	0.11	0.29	−0.067	13, 28
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7				56
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–				280
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–	0.13	0.15	0.157	184
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7		0.44		97, 116
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5	0.035	0.39	−0.189	66, 111
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–	0.10	0.080	−0.003	323
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–		0.69	−0.033	294
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4	0.070	0.082	−0.016	239
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–	0.033	0.012	−0.119	315
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0	0.010	0.030	−0.154	115
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1	0.012	0.011	−0.366	6, 282
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240	166
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156	164
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263	160
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326	209
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226	323, 328

Assortativity

From M. Newman. Networks. An introduction



(a) (b)

Figure 7.12: Assortative and disassortative networks. These two small networks are not real networks—they were computer generated to display the phenomenon of assortativity by degree. (a) A network that is assortative by degree, displaying the characteristic dense core of high-degree vertices surrounded by a periphery of lower-degree ones. (b) A disassortative network, displaying the star-like structures characteristic of this case. Figure from Newman and Girvan [249]. Copyright 2003 Springer-Verlag Berlin Heidelberg. Reproduced with kind permission of Springer Science and Business Media.