NATIONAL RESEARCH UNIVERSITY

Descriptive Network Analysis

Social Network Analysis. MAGoLEGO course. Lecture 2

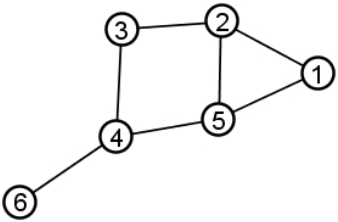
Leonid Zhukov

lzhukov@hse.ru
www.leonidzhukov.net/hse/2019/sna

National Research University Higher School of Economics
School of Data Analysis and Artificial Intelligence, Department of Computer Science

Graphs

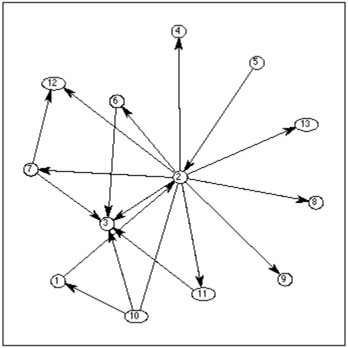
- A *graph* $G = (V, E)$ is an ordered pair of sets: a set of vertices V and a set edges E, where $n = |V|$, $m = |E|$
- An *edge* $e_{ij} = (v_i, v_j)$ is pair of vertices (ordered pair for directed graph)
- *Adjacency matrix* $A^{n \times n}$ is a matrix with nonzero element a_{ij} when there is an edge e_{ij}

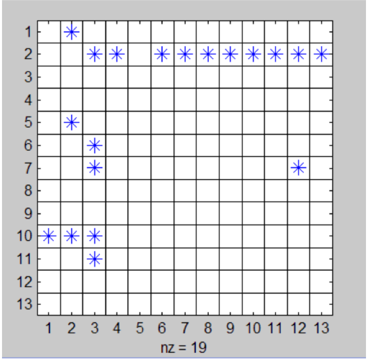


	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0	1	0	0	1	0
[2,]	1	0	1	0	1	0
[3,]	0	1	0	1	0	0
[4,]	0	0	1	0	1	1
[5,]	1	1	0	1	0	0
[6,]	0	0	0	1	0	0

Graphs and matrices

Graph $G(n, m)$, adjacency matrix $A_{ij}^{n\times n}$, edge $i \to j$, $m = nnz(A)$

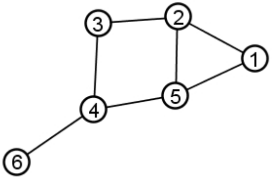




Node degree

- Two nodes/vertices are *adjacent* if they share a common edge
- An edge and a node on that edge are called *incident*.
- The *neighborhood* $\mathcal{N}(v)$ of a node v in a graph G is the set of nodes adjacent to v.
- The *degree* k_i of a nodes v_i is the total number of nodes adjacent to it, $k_i = |\mathcal{N}(v_i)|$
- Average node degree:

$$\langle k \rangle = \frac{1}{n} \sum_i k_i = \frac{2m}{n} = \frac{2|E|}{|V|}$$



Node degree

in directed networks:

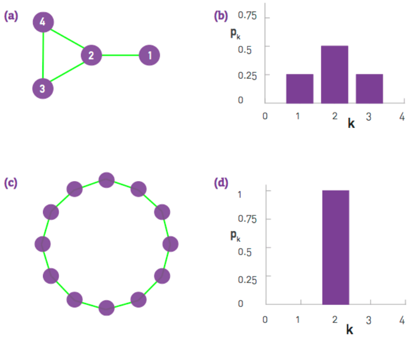
- k_i^{in} - incoming degree, number of edges/links pointing to node i
- k_i^{out} - outgoing degree, number of edges/links pointing from node i
- total node degree $k_i = k_i^{in} + k_i^{out}$
- Average in and out degrees are equal:

$$\langle k^{in} \rangle = \frac{1}{n} \sum_i k_i^{in} = \langle k^{out} \rangle = \frac{1}{n} \sum_i k_i^{out} = \frac{m}{n} = \frac{|E|}{|V|}$$

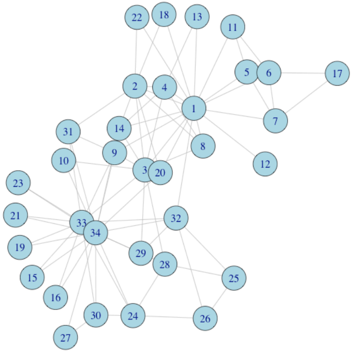
Degree distribution

- k_i - node degree, $k_i = 1, 2, \ldots k_{\text{max}}$
- n_k - number of nodes with degree k, total nodes $n = \sum_k n_k$
- Degree distribution is a fraction of the nodes with degree k

$$P(k_i = k) = P(k) = P_k = \frac{n_k}{\sum_k n_k} = \frac{n_k}{n}$$

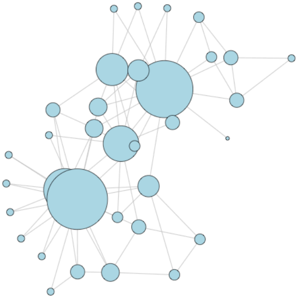


Node degrees



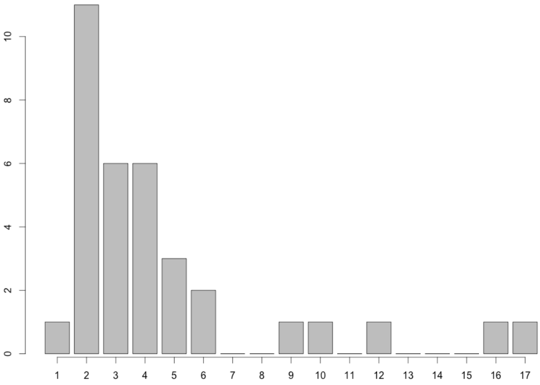
igraphdata: data(karate), igraph:plot()

Node degrees



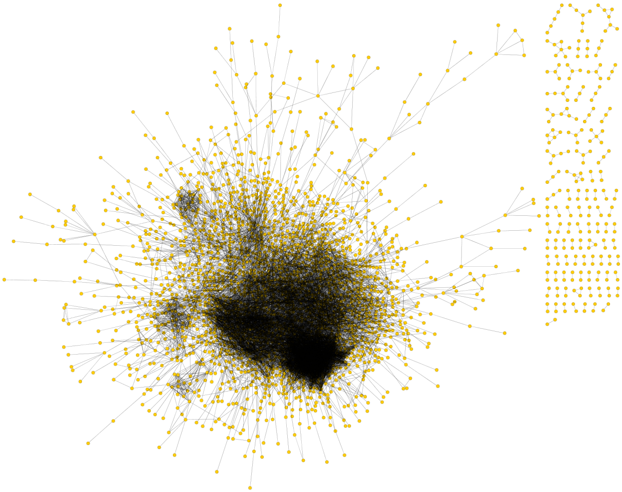
igraphdata: data(karate), igprah:plot()

Node degree histogram



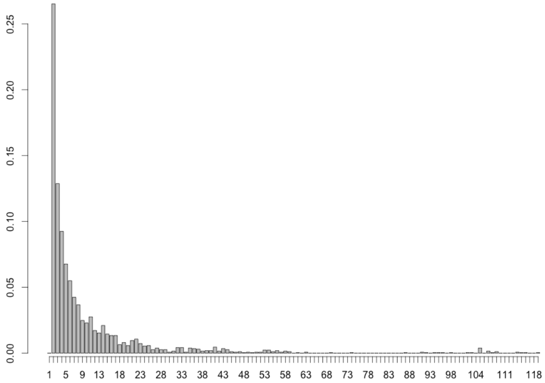
igraph: degree.distribution()

Degree distribution



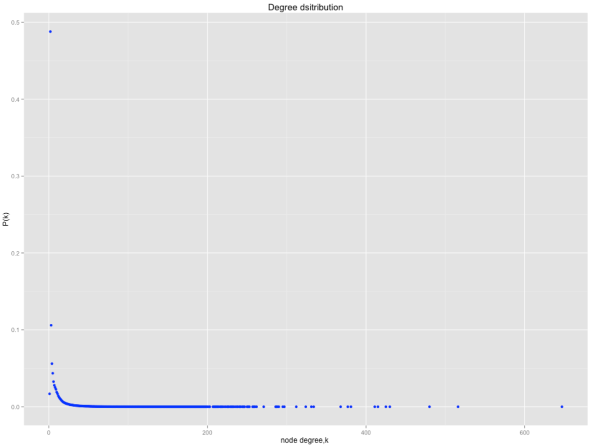
igraphdata: data(yeast)

Degree distribution



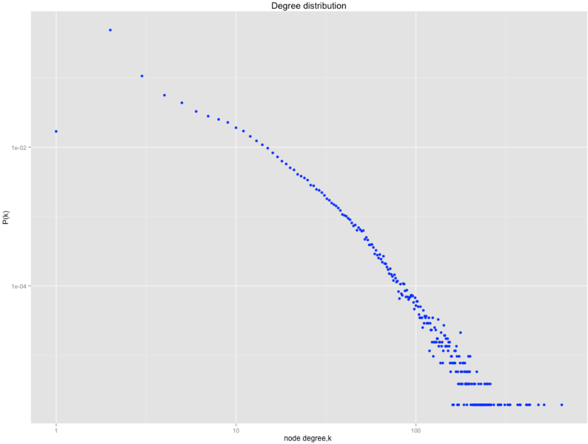
igraph: degree.distribution()

Power law degree distribution



Power law degree distribution

log-log scale



Discrete power law distribution

- Power law distribution

$$P(k) = Ck^{-\gamma} = \frac{1}{k^{\gamma}}C$$

- Log-log coordinates

$$\log P(k) = -\gamma \log k + \log C$$

$$y = -\gamma x + b$$

Distribution parameter estimation

- Maximum likelihood estimation of parameter γ:

$$\gamma = 1 + n \left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{\min}} \right]^{-1}$$

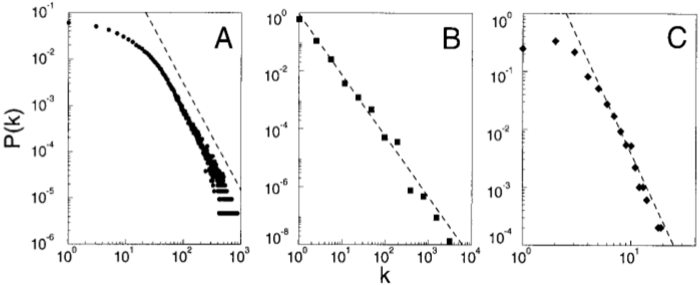
- error estimate

$$\sigma = \sqrt{n} \left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{\min}} \right]^{-1} = \frac{\gamma - 1}{\sqrt{n}}$$

- Optimal value of k_{min} can be found using Kolmogorov-Smirnov test for optimal distribution fitting

igraph:power.law.fit()

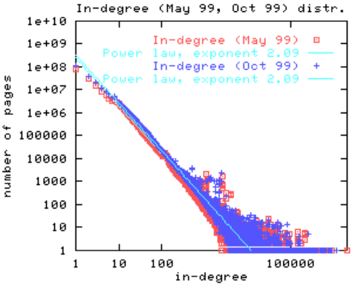
Power law networks



Actor collaboration graph, N=212,250 nodes, $\langle k \rangle = 28.8, \gamma = 2.3$
WWW, N = 325,729 nodes, $\langle k \rangle = 5.6, \gamma = 2.1$
Power grid data, N = 4941 nodes, $\langle k \rangle = 5.5, \gamma = 4$

Barabasi et.al, 1999

Power law networks

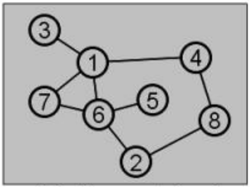


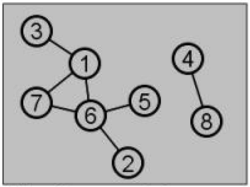
In- and out- degrees of WWW crawl 1999

Broder et.al, 1999

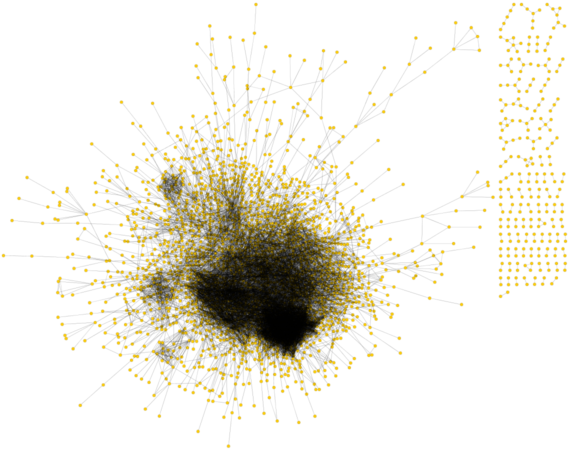
Graph connectivity

- A *path* from v_i to v_j is a sequence of edges that joins two vertices. (It also ordered list of vertices such that that there is an edge to the next vertex on the list)
- A graph is *connected* if there a paths between any two vertices.
- *Connected component* is a maximal connected subgraph of G





Graph connectivity



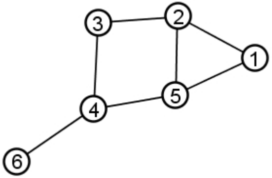
Connected components: 92

Component sizes: $2375\ 7\ 7\ 7\ 6\ 5\ 5\ 5\ 5\ 5\ 5\ 4\ 4\ 4\ 4$

Graph connectivity

- The *distance* $d_G(v_i, v_j)$ between two vertices is the number of edges in the shortest path from v_i to v_j
- Graph *diameter* is the largest shortest path: $D = \max_{i,j} d_G(v_i, v_j)$
- Average path length:

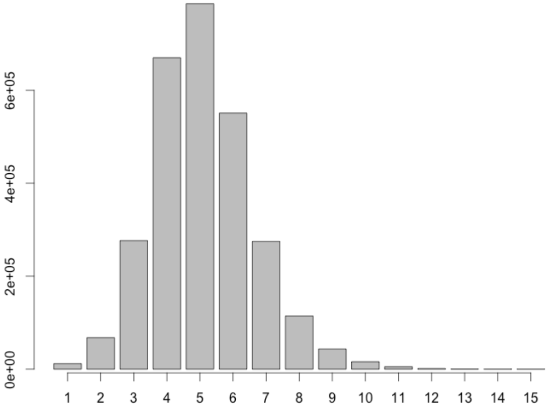
$$\langle L \rangle = \frac{1}{n(n-1)} \sum_{i \neq j} d_G(v_i, v_j)$$



igraph: shortest.paths(), diameter(), average.path.length(), path.length.hist()

Graph average path length

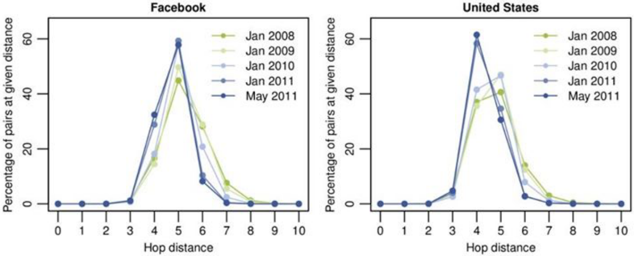
"Yeast" graph, $n = 2617$, $m = 11855$



Diameter $D = 15$, average path length $\langle L \rangle = 5.1$

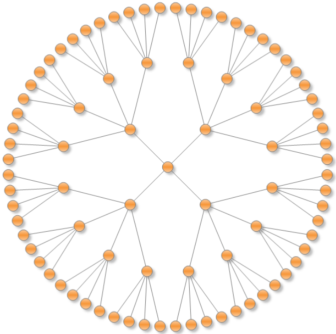
Small world

- Email graph:
  D. Watts (2001), 48,000 senders, $\langle L \rangle \approx 6$
- MSN Messenger graph:
  J. Lescovec et al (2007), 240mln users, $\langle L \rangle \approx 6.6$
- Facebook graph:
  L. Backstrom et al (2012), 721 mln users, $\langle L \rangle \approx 4.74$



figures from L.Backstrom, 2012

Simple model



An estimate: $z^d = N$, $d = \log N / \log z$
$N \approx 6.7$ bln, $z = 50$ friends, $d \approx 5.8$.

Triads, transitivity and clustering

Facebook friendship

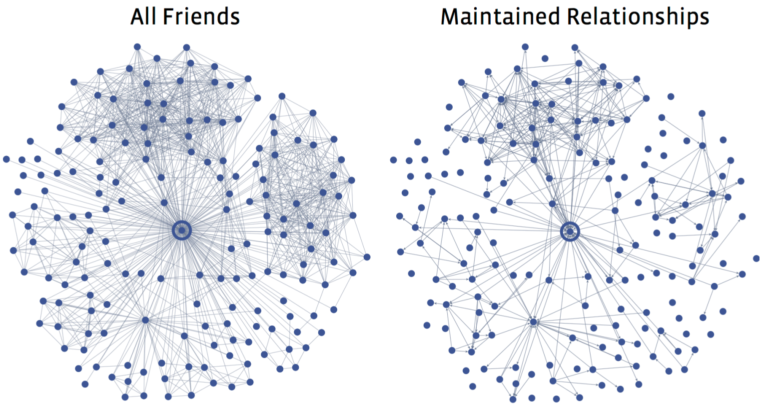


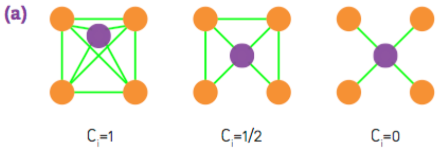
image from Cameron Marlow, Facebook

Clustering coefficient

How neighbors of a given node connected to each other

- *Local clustering coefficient* (per vertex):

$$C_i = \frac{\text{number of links in } \mathcal{N}_i}{k_i(k_i - 1)/2}$$



- Average clustering coefficient:

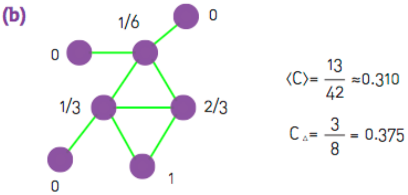
$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$$

igraph:transitivity(type="local")

Clustering coefficient

- *Global clustering coefficient*:

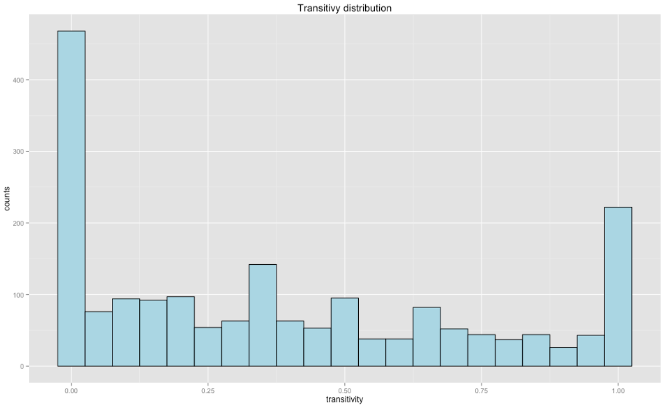
$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}}$$



igraph:transitivity(type="global")

Clustering coefficient

Yeast graph



Global clustering coefficient: $C = 0.468$

Statistical properties

- Power-law degree distribution
- Small average path length
- High clustering coefficient (transitivity)
- Gigantic connected component

References

- Statistical Analysis of Network Data with R. Eric Kolaczyk, Gabor Csardi. Springer, 2014
- Social Network Analysis: Methods and Applications. S. Wasserman, K. Faust, Cambridge University Press, 1994
- Networks: An Introduction. Mark Newman. Oxford University Press, 2010.
- Power laws, Pareto distributions and Zipf's law, M. E. J. Newman, Contemporary Physics, pages 323–351, 2005.
- Power-Law Distribution in Empirical Data, A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Review, Vol 51, No 4, pp. 661-703, 2009.