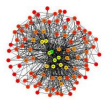
Data and Networks
Part II

Dr. Franck Kalala Mutombo

African Institute for Mathematical Sciences (AIMS) South Africa

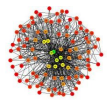
franckm@aims.ac.za

May 19, 2017



Scale free network

Scale free network



Power-Law Degree Distribution

Definition

(Non-stochastic) A finite sequence $k = (k_1, \cdots, k_n)$ of real numbers, such that $k_1 \leq k_2 \cdots \leq k_n$, is said to follow a *power-law* or *scaling relationship* if

$$r = ck_r^{-\gamma}, \tag{1}$$

where r (by definition) is the rank of k_r, c is a fixed constant, and γ is called the scaling index. The definition is said to be non-stochastic in the sense that there is no underlying probability model for the given sequence. The relationship for the rank r versus k appears as a line of slope $-\gamma$ when plotted on a log-log scale. Indeed we have

$$\log(r) = \log(c) - \gamma \log(k_r). \tag{2}$$

The relationship (1) is referred to as the size-rank (or cumulative) form of scaling.

Definition

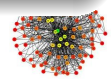
(Stochastic)

- let us assume an underlying probability model P for a non-negative random variable X,
- let $F(x) = P[X \leq x]$ for $x \geq 0$ denote the (cumulative) distribution function (CDF) of X,
- and let $F(x) = 1 - F(x)$ denote the complementary CDF (CCDF) or the *tail* function [Papaulis, 1984, Grimmett, Stirzaker].

In this context, a random variable X or its corresponding distribution function F is said to follow a *power-law* or is *scaling* with index $\lambda > 0$ if as $x \to \infty$,

$$P[X > x] = 1 - F(x) \sim cx^{-\lambda}, \tag{3}$$

for some $0 < c < \infty$ and a *tail index* $\lambda > 0$. $f(x) \sim g(x)$ as $x \to \infty$ if $f(x)/g(x) \to 1$ as $x \to \infty$.



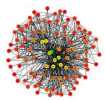
Requiring the existence of the cumulative distribution function $F(x)$, the probability density function (pdf) of the random variable X is defined to be [Papaulis, 1984, Grimmett, Stirzaker]:

$$f(x) = dF(x)/dx,$$

so for the random variable X, its stochastic cumulative form of scaling or size-rank relationship (4) has an equivalent non-cumulative or size-frequency counterpart given by

$$f(x) \sim cx^{-(1+\gamma)}$$

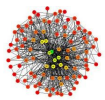
which appears similarly as a line of slope $-(1+\gamma)$ on a log-log scale.



If X is a continuous random variable, its first moment, i.e. its mean, is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = c\lambda \int_{x_{min}}^{\infty} \frac{1}{x^{\lambda}} = \frac{c\lambda}{1-\lambda} \frac{1}{x^{\lambda-1}} \Big|_{x_{min}}^{\infty} dx.$$

- if $1 < \lambda < 2$, the first moment is finite but the second moment/variance is infinite,
- if $0 < \lambda \leq 1$, both the second moment/variance and the first moment/mean are infinite. For this reason power-law distributions are sometimes called heavy tail distributions. m
- In general, all moments of F of order $\beta \geq \gamma$ are infinite.



The relation

$$P[X > x] = 1 - F(x) \sim cx^{-\lambda}, \tag{4}$$

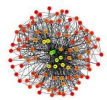
implies that

$$\log(P[X > x]) \approx log(c) - \lambda \log(x), \tag{5}$$

A doubly logarithmic plots of x versus $1 - F(x)$ yield straight lines of slope $-\lambda$, at least for large x.
If $x > u$, then the conditional distribution of X given that $X > u$ is given by

$$P\{[X > x | X > u]\} = \frac{P\{[X > x] \cap [X > u]\}}{P[X > u]} = \frac{P[X > x]}{P[X > u]} \sim c_1 x^{-\gamma},$$

where the constant c_1 is given by $1/u^{-\gamma}$ and does not depend on x.
Hence, when x is large, the conditional probability $P[X > x | X > u]$ is identical to the (unconditional) distribution $P[X > x]$, except for a change in scale. Owing to this fact, power-law distributions are often called *scaling distributions* or *scale-free distributions*.



The usual way of referring to power-law networks is scale-free networks meaning that there exists a power-law relationship between the probability density function (or probability mass function for discrete random variables) and the node degree often translates as

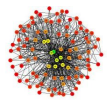
$$p(k) = Bk^{-\gamma}. \tag{6}$$

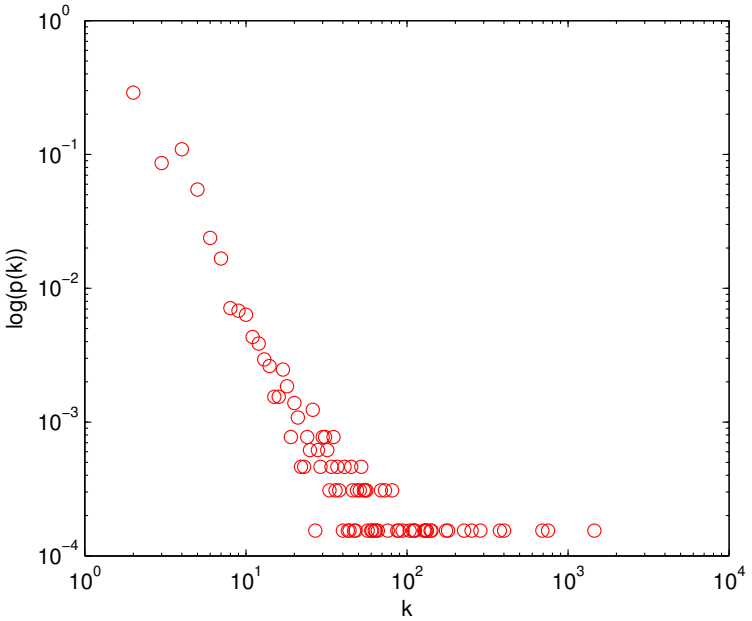
If we scale the degree by a constant factor, say c, then this produces only a proportionate scaling of the function, i.e.

$$p(k, c) = B(ck)^{-\gamma} = Bc^{-\gamma}.p(k)$$

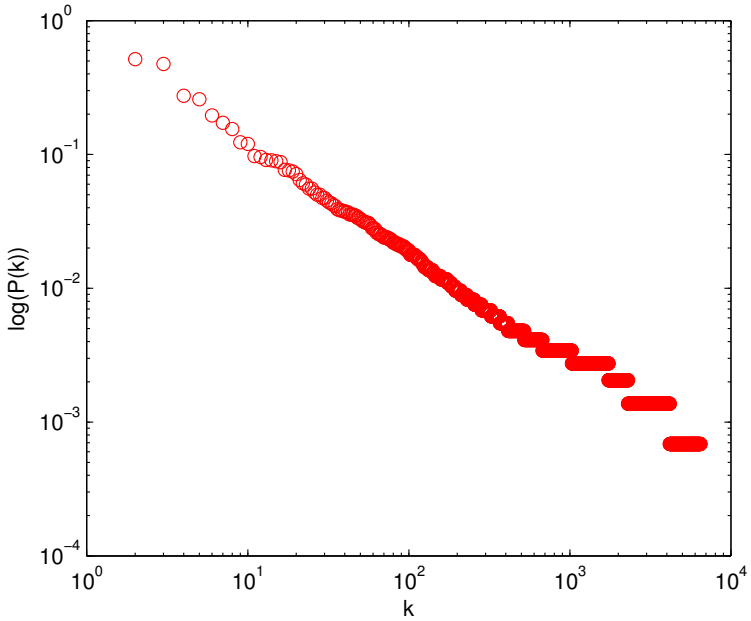
which is identical to $p(k)$ except for a change of scale. In a $\log - \log$ scale equation (6) results in a straight line of slope $-\gamma$.
In many scenarios the power-law relationship (6) is satisfied only in the *tail* of the distribution, when the value of k tends to infinity but not for small values of k. Therefore, we usually write

$$p(k) \sim k^{-\gamma}.$$



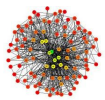


(a)



(b)

Figure : Probability (a) and cumulative distribution functions (b) for the version of the internet at autonomous system (AS) level displaying a power-law degree distribution.



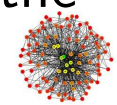
We can see that the tail of the distribution is very noisy and one way to solve this problem is to consider the cumulative distribution function (CDF) given by

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

which represents the fraction of nodes having degree k or greater or a probability of choosing a node with degree greater than or equal to k. For this case we can show that $P(k)$ also shows a power-law decay with the degree. In fact

$$P(k) = \sum_{k'=k}^{\infty} p(k') = C \sum_{k'=k}^{\infty} k'^{-\gamma} \simeq \int_{k}^{\infty} k'^{-\gamma} dk' = \frac{C}{\gamma - 1} k^{\gamma - 1},$$

where $p(k'^{-\gamma}) = Ck'^{-\gamma}$ and $k \geq k_{kmin}$. In the Figure 1 (b) we illustrate a plot for the case of the (AS) version of the internet. As we can see the CDF plot significantly reduced the noise compared to the plot in (a) of the same figure.



Another approach to reduce the noise in the tail of the distribution is to use the logarithmic binning of the form $a^{n-1} \leq k < a^n$ where $a = 2$ and n runs in the range of nodes. For example the first bin for $n = 1$ is $1 \leq k < 2$ and all nodes of degree 1 fall in that bin.

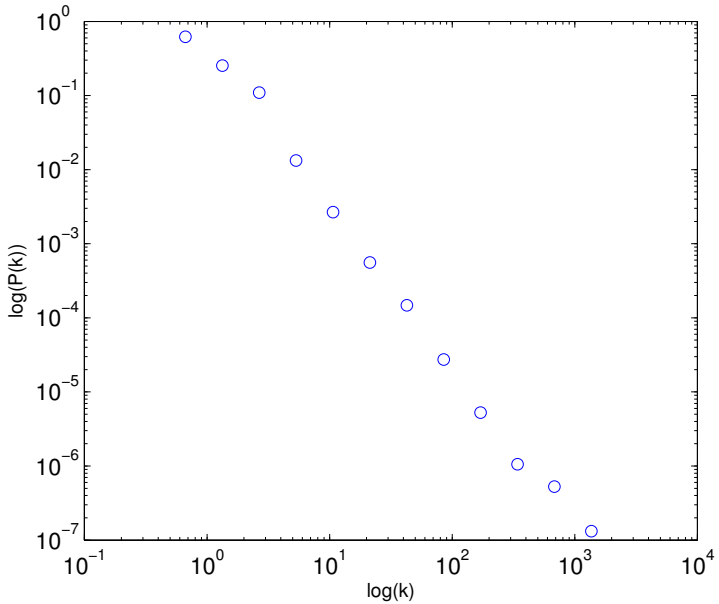
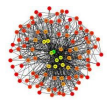


Figure : Cumulative distribution functions for the version of the internet using logarithmic bins displaying a power-law degree distribution.

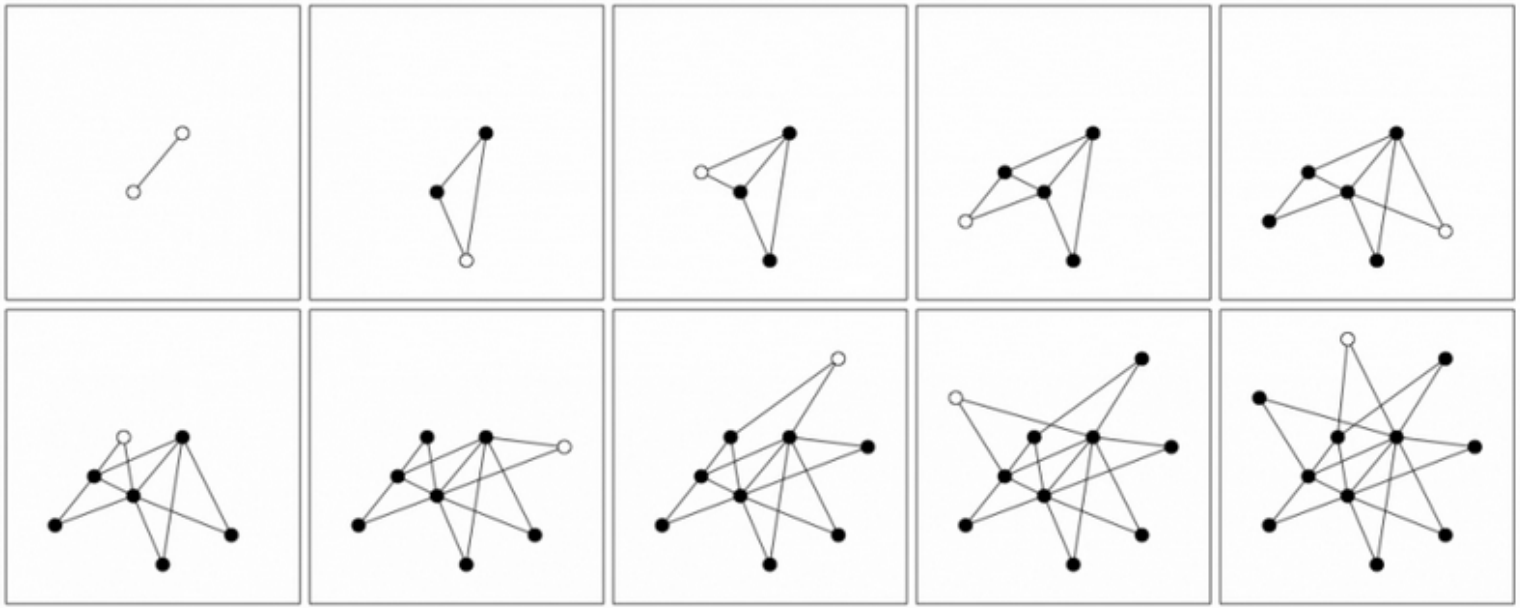
The second bin is $2 \leq k < 4$ and contains nodes of degree 2 and 3 and so on.

Scale Free Networks: [Barabási and Albert 1999]

- There are some real-world phenomena that small-world phenomena cannot capture, the most relevant one being evolution. (Small world models will be studied later on).
- Several empirical results demonstrate that many large networks are scale free, that is, their degree distribution follows a power law for large k.
- The important question is then: what is the mechanism responsible for the emergence of scale free networks? Answering this question requires a shift from modelling network topology to modelling the network assembly and evolution.
- While the goal of the former models is to construct a graph with correct topological features, the modelling of scale-free networks will put the emphasis on capturing the network dynamics.



BARABASI ALBERT MODEL



Evolution of the Barabsi-Albert Model

The sequence of images shows nine subsequent steps of the Barabsi-Albert model. Empty circles mark the newly added node to the network, which decides where to connect its two links (m=2) using preferential attachment.

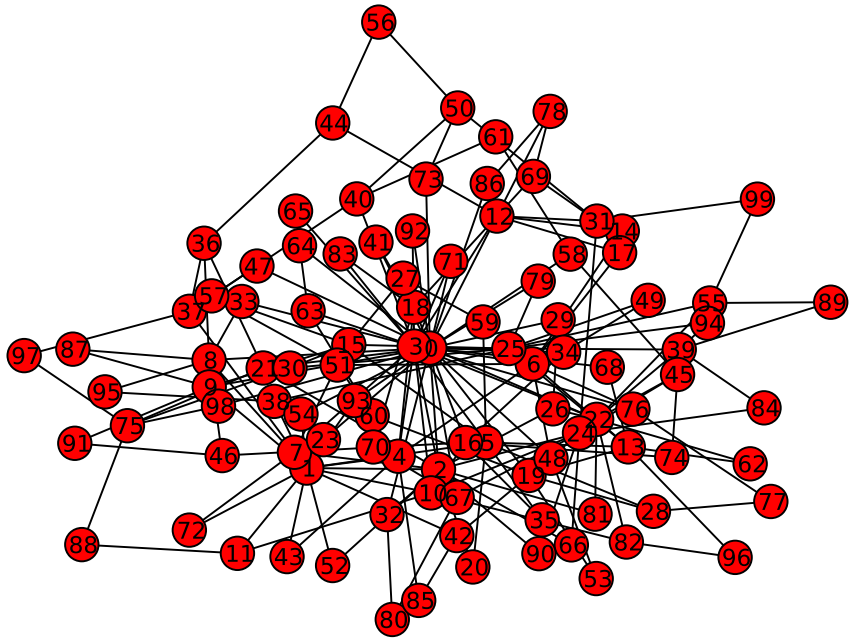
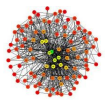


Figure : Example of a random network obtained with the preferential attachment method of Barabási and Albert with $n = 100$ and $m_0 = 2$.

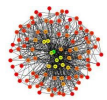


The Barabási-Albert (BA) Model

Two ingredients, growth and preferential attachment, inspired the introduction of the Barabási-Albert model (BA), which led for the first time to a network with a power-law degree distribution. The algorithm of the BA model is the following [Barabási and Albert 1999]:

1. Growth: Starting with a small number (m_0) of nodes, at every time step, we add a new node with $m(\leq m_0)$ edges that links the new node to m different nodes already present in the system.
2. Preferential attachment: When choosing the nodes to which the new node connects, we assume that the probability $\prod_i$ that a new node will be connected to node i depends on the degree k_i, such that

$$\prod_i = \frac{k_i}{\sum_j k_j}.$$



The Barabási and Albert (1999) network produces a network with the following approximate probability distribution for the degrees within the network:

$$p(k) = \frac{2m^2}{k^3}, \ k = m, \ m+1, \ \cdots, \ n.$$

with $\gamma = 3$. An exact result for the degree distribution due to Dorogovtsev et al. (2000) is that

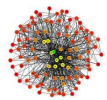
$$p(k) = \frac{2m(m+1)}{k(k+1)(k+2)}, \ k = m, \ m+1, \ \cdots, \ n.$$

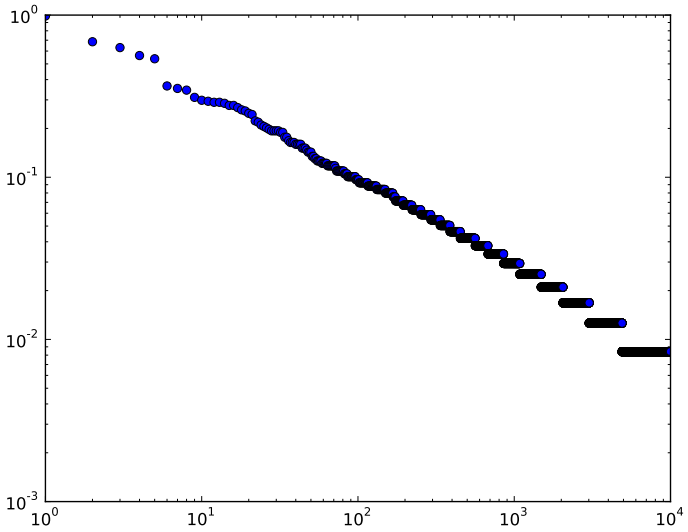
For large k, i.e. when $k \to \infty$, we have that

$$p(k) \sim k^{-3}$$

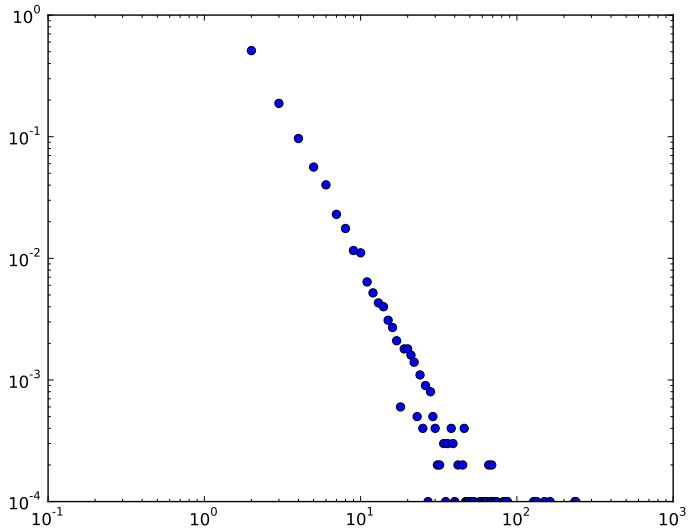
which immediately implies that the cumulative degree distribution is given by

$$P(k) \sim k^{-2}.$$



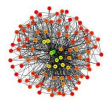


(a)



(b)

Figure : Cumulative degree distribution (a) and probability distribution (b) for a *SF* with $n = 10000$, constructed according to the BA model. For each node entering the network, 2 new edges are placed.



Since the algorithm always adds m links at each of the $n - m_0$ steps, the total number of (undirected) links in the final network is always $m_0(m_0 - 1)/2 + m(n - m_0)$, and therefore the mean degree is [GPrettejohn et al. 2011]

$$\overline{k} = 2\frac{0.5m_0(m_0 - 1) + m(n - m_0)}{n} = \frac{m_0(m_0 - 1) + 2m(n - m)}{n},$$

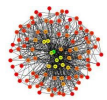
and when $n \to \infty$, we get

$$\overline{k} \simeq 2m.$$

No exact expression for the average path length is known, however, the approximate scaling $L \sim (\log(n)/\log(\log(n)))$ is derived in [Fronczak et al. (2004)] and the expression for the approximate pathlength

$$L \simeq \frac{\log(n) - log(m/2) - 1 - 0.577}{\log(\log(n)) + \log(m/2)} + 1.5$$

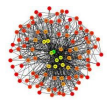
is derived in [GPrettejohn et al. 2011].



The clustering coefficient is not known exactly. It was shown in [Barabási and Albert 1999] that the clustering coefficient decreases with n, but less slowly than it decreases for an Erdös-Rényi network. Recently, the approximate expression for the clustering coefficient

$$\overline{C} \sim \frac{m \log(n)^2}{8n}$$

has been derived [**?**]. However, since the clustering coefficient decreases with n we can expect that the clustering coefficient will become close to zero as n increases.



References

Papaulis A. (1984)

Probability, random variables, and stochastic processes

McGraw-Hill Book Co., New York.

Grimmett, G. R. and Stirzaker, D. R. (1992)

Probability and random processes

The Clarendon Press Oxford University Press, New York.

Barabási, A.-L. and Albert, R. (1999)

Statistical mechanics of complex networks

Science, 286, 509–512.

GPrettejohn, B. J. and Berryman, M. J. and McDonnell, M. D. (2011)

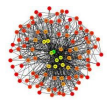
Methods for generating complex networks with selected structural properties for simulations: A review and tutorial for neuroscientists

Front. Comput. Neurosci.,5

Fronczak, A. and Fronczak, P. and Holyst, J. A. (2004)

Average path length in random networks.

Phys. Rev. E.,70, 056110.



The End