Social Networks

Practical session: plotting heterogeneous distributions, fitting power laws, and calculating network robustness.

Instructors: Markus Strohmaier, Johannes Wachs

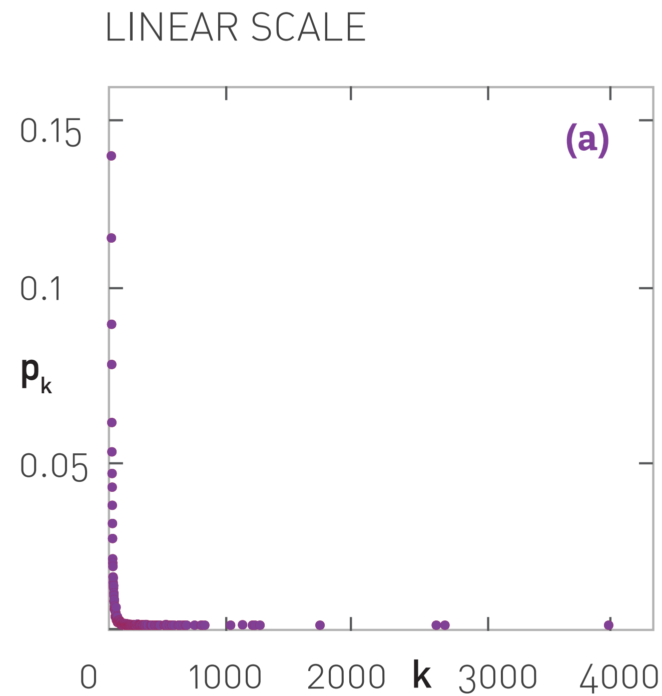
Plotting power laws

Plotting the degree distribution is an integral part of analyzing the properties of a network. The process starts with obtaining N_k, the number of nodes with degree k. This can be provided by direct measurement or by a model. From N_k we calculate $p_k = N_k/N$. The question is, how to plot p_k to best extract its properties.

Linear scale

Using a linear *k*-axis compresses the numerous small degree nodes in the small-*k* region, rendering them invisible. Similarly, as there can be orders of magnitude differences in p_k for $k = 1$ and for large *k,* if we plot p_k on a linear vertical axis, its value for large *k* will appear to be zero.

The use of a log-log plot avoids these problems.

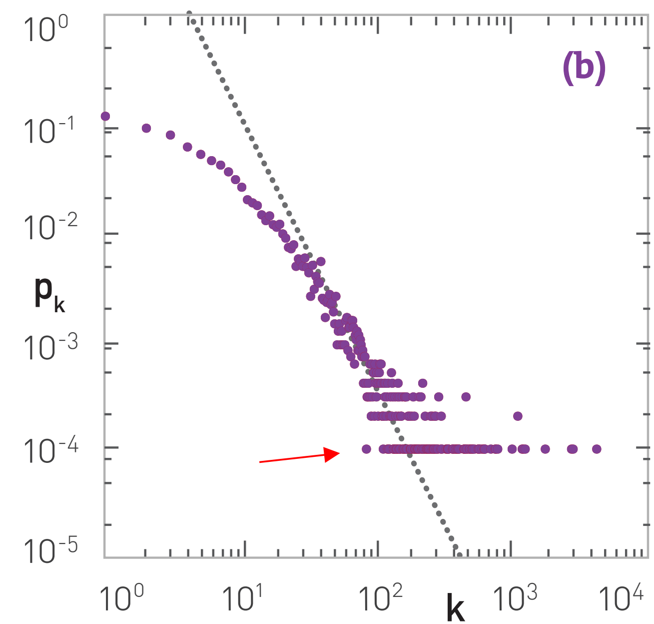


Avoid Linear Binning

The most flawed method (yet frequently seen in the literature) is to simply plot $p_k = N_k/N$ on a log-log plot. This is called ***linear binning*, as each bin has the same size Δ*k* = 1**. For a scale-free network linear binning results in an instantly recognizable plateau at large *k*, consisting of numerous data points that form a horizontal line.

This plateau has a simple explanation: Typically we have only one copy of each high degree node, hence in the high-*k* region we either have N_k=0 (no node with degree *k*) or N_k=1 (a single node with degree *k*). Consequently linear binning will either provide p_k=0, not shown on a log-log plot, or $p_k = 1/N$, which applies to all hubs, generating a plateau at $p_k = 1/N$. “Finite-size effects”.

LINEAR BINNING

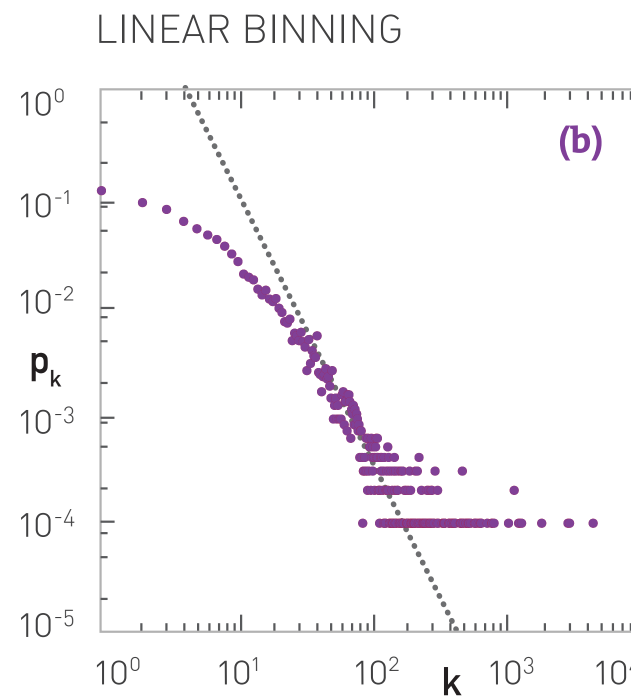


Avoid Linear Binning

This plateau affects our ability to estimate the degree exponent γ using linear binning, the obtained γ is quite different from the real value.

The reason is that under linear binning we have a large number of nodes in small k bins, allowing us to confidently fit p_k in this regime.

In the large-k bins we have too few nodes for a proper statistical estimate of p_k. Instead the emerging plateau biases our fit. Yet, it is precisely this high-k regime that plays a key role in determining γ. Increasing the bin size will not solve this problem. It is therefore recommended to avoid linear binning for fat tailed distributions.



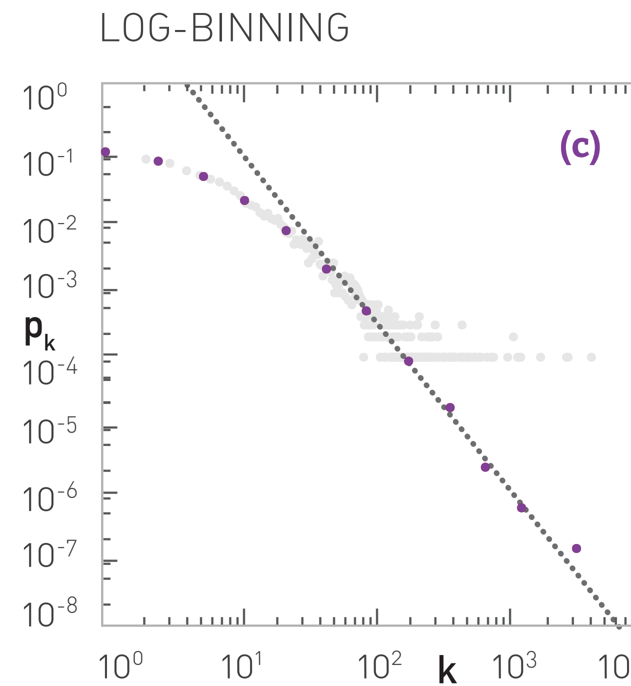
Logarithmic Binning

Logarithmic binning **correcs the non-uniform sampling of linear binning.**

For log-binning we let the bin sizes increase with the degree, making sure that each bin has a comparable number of nodes.

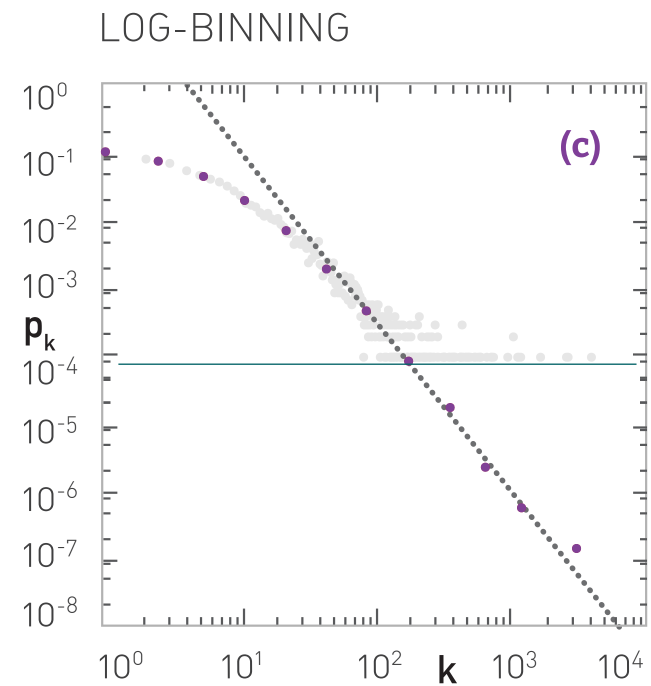
For example, we can choose the bin sizes to be multiples of 2, so that the first bin has size b_0=1, containing all nodes with k=1; the second has size b_1=2, containing nodes with degrees k=2, 3; the third bin has size b_2=4 containing nodes with degrees k=4, 5, 6, 7. By induction the n_{th} bin has size 2^{n-1} and contains all nodes with degrees $k=2^{n-1}, 2^{n-1+1}, ..., 2^{n-1-1}$.

The degree distribution is given by $p_{\langle kn \rangle}=N_n/b_n$, where N_n is the number of nodes found in the bin n of size b_n and $\langle k_n \rangle$ is the average degree of the nodes in bin b_n.



Logarithmic Binning

Note that now the scaling extends into the high-*k* plateau, invisible under linear binning. Therefore logarithmic binning extracts useful information from the rare high degree nodes as well.

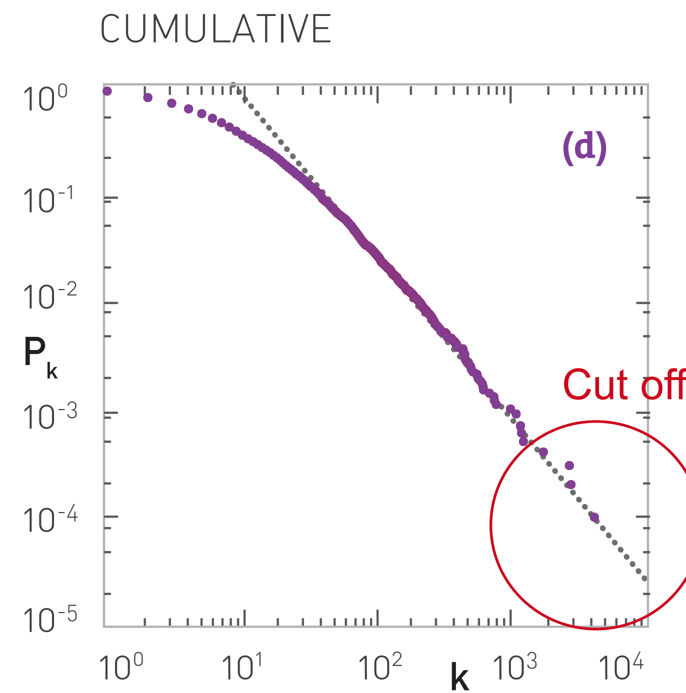


Cumulative Distribution

Another way to extract information from the tail of p_k is to plot the complementary cumulative distribution

which again enhances the statistical significance of the high-degree region.

The cumulative distribution again eliminates the plateau observed for linear binning and leads to an extended scaling region, **allowing for a more accurate estimate of the degree exponent**.



Power laws, Pareto distributions and Zipf's law, M. E. J. Newman

Linear scale

Logarithmic scale

Even with logarithmic binning there is still some noise in the tail, although it is sharply decreased

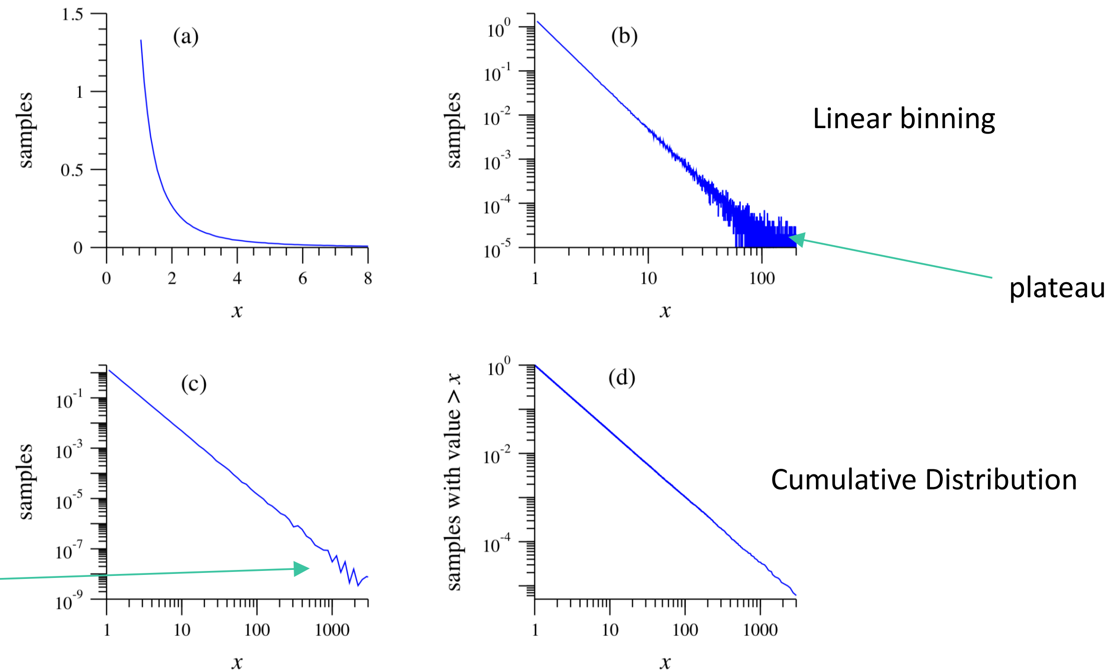
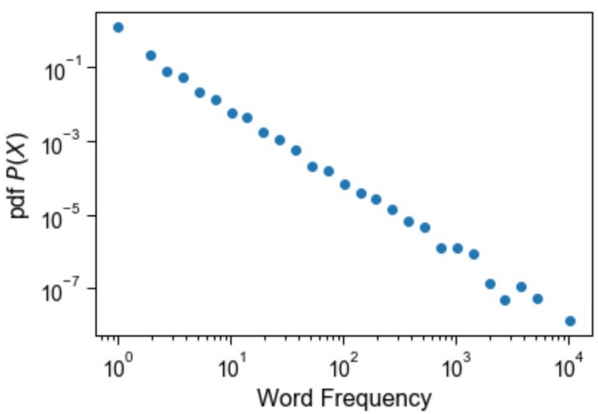


FIG. 3 (a) Histogram of the set of 1 million random numbers described in the text, which have a power-law distribution with exponent $\alpha = 2.5$. (b) The same histogram on logarithmic scales. Notice how noisy the results get in the tail towards the right-hand side of the panel. This happens because the number of samples in the bins becomes small and statistical fluctuations are therefore large as a fraction of sample number. (c) A histogram constructed using "logarithmic binning". (d) A cumulative histogram or rank/frequency plot of the same data. The cumulative distribution also follows a power law, but with an exponent of $\alpha - 1 = 1.5$.

 https://arxiv.org/abs/cond-mat/0412004

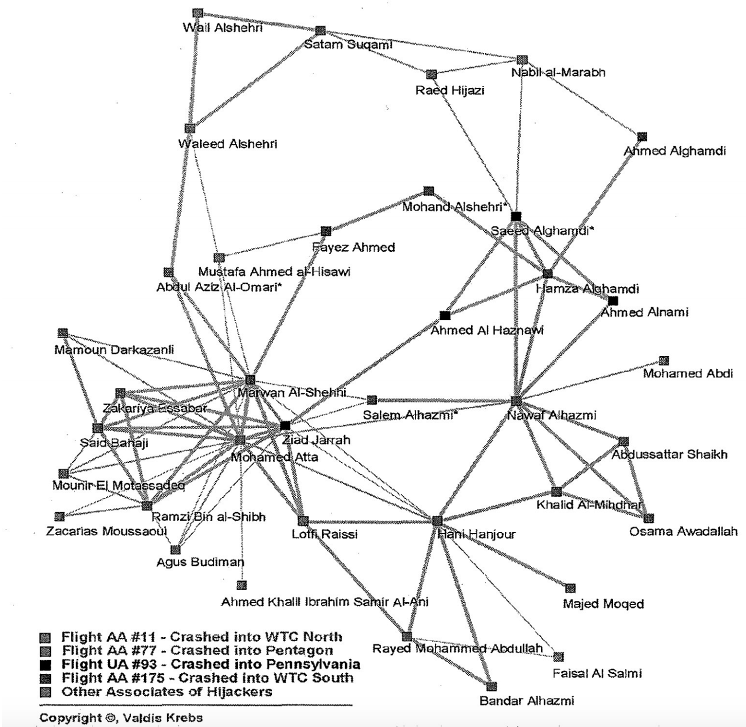
Hands-on

Today we will use a notebook generously shared by Prof. Michael Szell of IT University of Copenhagen which shows how to plot the degree distribution and also how to fit powerlaw distributions to real data.



Hands-on Part 2

In the second part of the practical we will have a look at the collaboration network of the September 11 hiiackers and test its robustness.



Krebs, Valdis E. "Mapping networks of terrorist cells." Connections 24.3 (2002): 43-52.

Social Networks

Thanks!

Instructors: Markus Strohmaier, Johannes Wachs