Thresholding normally distributed data creates complex networks

George Cantwell[1], Yanchen Liu[2], Benjamin Maier[3], Carlos A. Serván[4], Alice C. Schwarze[5], Jordan Snyder[6], and Guillaume St-Onge[7]

[1]Department of Physics, University of Michigan, Ann Arbor, Michigan, USA
[2]Center for Complex Network Research, Northeastern University, Boston, Massachusetts, USA
[3]Robert Koch Institute, Nordufer 20, D-13353 Berlin, Germany
[4]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA
[5]Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom
[6]Department of Mathematics, University of California, Davis, California, USA
[7]Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec, Canada

ABSTRACT

Network data sets are often constructed by some kind of thresholding procedure. The resulting networks frequently possess properties such as heavy-tailed degree distributions, non-zero clustering, large connected components and short average shortest path lengths. These properties are considered typical of complex networks and appear in many contexts, prompting consideration of their universality. Here we introduce a very simple model for continuous valued relational data and study the effect of thresholding it. We find that some, but not all, of the properties associated with complex networks can be seen after thresholding, even when the underlying data is not "complex".

We examine the properties of networks created by thresholding relational data. To do this we introduce a basic null model of the underlying data, which is then thresholded. The model is derived from three assumptions: (1) all nodes are statistically identical; (2) any correlations are local; and (3) the underlying data is normally distributed. If there are n nodes, the underlying data will be an $n \times n$ symmetric matrix, $\boldsymbol{X}$. The listed assumptions lead to a simple one parameter normal model for the data:

$$\boldsymbol{X} \sim \mathscr{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \qquad \text{where,} \qquad \Sigma_{(i,j),(i,j)} = 1; \qquad \Sigma_{(i,j),(i,k)} = \rho; \qquad \Sigma_{(i,j),(k,l)} = 0. \tag{1}$$

Networks are created by thresholding $\boldsymbol{X}$. For each pair of nodes i, j, if $X_{ij} \geq t$ we say there is an edge. Despite the fact that the underlying data is fully connected and normally distributed (a situation not usually considered "complex") we find the thresholded networks display some of the properties typically ascribed to complex networks. As an example, below we show some degree distributions for real-world networks along with the normal model predictions.

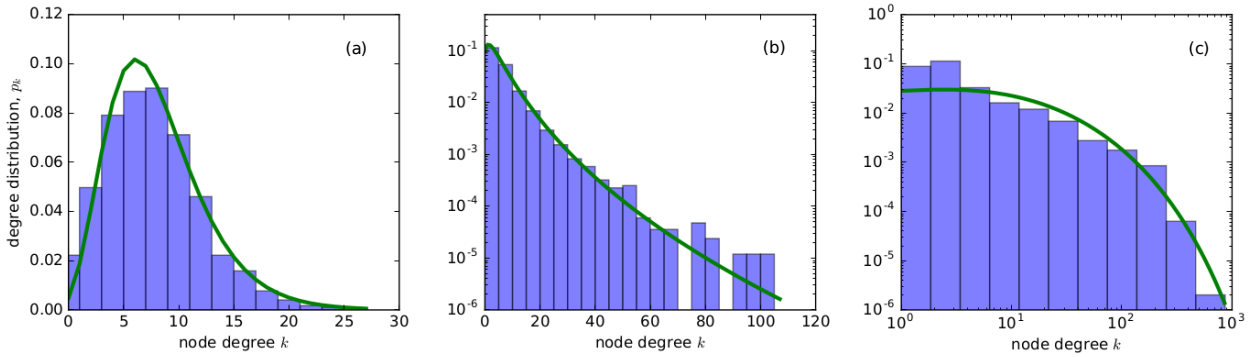

Figure 1. Degree histograms for three real-world networks along with fitted distributions from the threshold model. We show (a) A high school friendship network, (b) a co-authorship network between scientists, and (c) a protein–protein interaction network. These networks were chosen for their different degree distributions – note the different scales on the axes: linear, log-linear, and log-log.