

Technical report

Network degree distributions

Garry Robins

Philippa Pattison

Johan Koskinen

Social Networks laboratory

School of Behavioural Science

University of Melbourne

22 April 2008¹

This paper defines and summarises certain issues pertaining to degree distributions that affect the modeling of social networks. Some of these issues remain essentially unresolved in the existing research literature.

Definitions

(a) For nondirected graphs:

- *Degree* of node i : $k(i)$, the number of edges incident with i . $k(i) = \sum_j x_{ij}$
- *Degree sequence*: $k(1), k(2), \dots, k(n)$ is the (usually ordered) sequence of degrees of the nodes, indexed by the labels, $1, 2, \dots, n$ in the node set.
- *Degree distribution*: $(d_0, d_1, \dots, d_{n-1})$ where d_k is the number of nodes with degree k .

(b) For directed graphs

- *Outdegree* of node i : x_{i+} , the number of arcs directed from node i . $x_{i+} = \sum_j x_{ij}$
- *Indegree* of node i : x_{+i} , the number of arcs directed to node i . $x_{+i} = \sum_j x_{ji}$
- *In- and outdegree distributions*: distribution of counts of nodes with indegree and outdegree k , respectively.

Figure 1 presents a nondirected network and the associated degree distribution in histogram form.

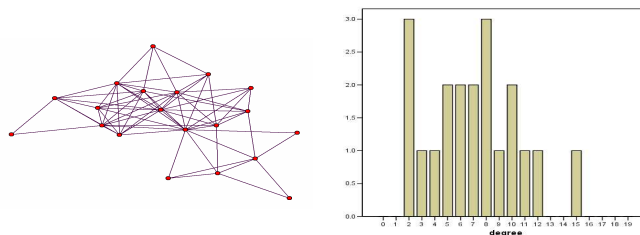


Figure 1
A nondirected network with degree distribution

¹ We would like to thank Michelle Shumate for making her data available for the purposes of this technical report.

The issues

There are three related questions that are central to our concerns:

1. How best to represent degree distribution in statistical models for social networks?
2. What is the range of convenient and plausible functional forms that a degree distribution might take?
3. How can we map a given functional form of a degree distribution into an exponential random graph parameterisation?

Probability models for the degree distributions

Let K be the degree of a *randomly chosen person* in the network². Then a statistical model for the degree distribution is represented by: $P(K=k) = f(k)$ where $f(k)$ is a probability distribution.

The question is: what is an appropriate $f(k)$? The literature provides a small range of possibilities

Degree distributions for simple random graphs

It is well known that Bernoulli random graph distributions (i.e. Erdős-Renyi graphs or simple random graphs) produce graphs with approximate Poisson degree distributions (and exact Poisson in the limit, c.f. e.g. Britton, Deijfen and Martin-Löf, 2006). Because the properties of these graphs and of the degree distribution are so well understood, they are easy to work with. Unfortunately, it is also well-known that they do not fit empirical network data well. There are two common features of empirical networks that simple random graphs do not represent well: a high level of triangulation (i.e. clustering); and positive skew on the degree distribution.

Another interesting feature attributed to empirical networks has been referred to as *degree-based assortative mixing*, whereby the degrees of adjacent nodes are correlated. This can be the outcome of high triangulation (Newman & Park, 2003). A possible feature that has not been extensively examined in empirical directed networks is the extent to which in- and out-degree distributions are correlated.

Considerable attention has been given to the variance of the degree distribution as a measure of inhomogeneity, including centrality and centralization. The degree variance is completely determined by the degree sequence and its properties, both exact and in relation to various random graph distributions (Snijders, 1981a,1981b; and Hagberg, 2000, 2003a,b,c).

² In the following, no explicit distinction is made between degree data sampled independently and degree data that emanates from a completely observed network on n nodes. In case of the former - which would include for example collections of ego-networks or survey responses to "How many sexual partners have you had in the past year" - the statement "*randomly chosen person*" might actually have a plausible interpretation in terms of the sampling mechanism. In the case of completely observed networks and their derived degree distributions, the notion of "*randomly chosen person*" is clearly tenuous and as a consequence observed frequencies for degrees cannot be modelled independently (not least because of functional dependencies between degrees in the case of for example undirected networks).

Positively skewed degree distributions

It is common for empirical social networks to have positively skewed degree distributions. This result reflects social processes whereby there is greater variation of activity of actors in the network than would be expected by chance. For instance, in directed networks, it is not unusual to see some actors who are highly *popular* (indegree) and/or highly *active* (sometimes referred to as *expansive* – outdegree). Sometimes, but certainly not universally, empirical networks may exhibit outliers in the degree distribution with particularly high degree. These are often referred to as *hubs*.

Figure 2 presents collaboration network data (directed) for 126 non-government and inter-governmental organizations, collected by Shumate (2008). Figure 3 presents the corresponding degree distributions, which exhibit strong positive skew for both in- and out-degree distributions.

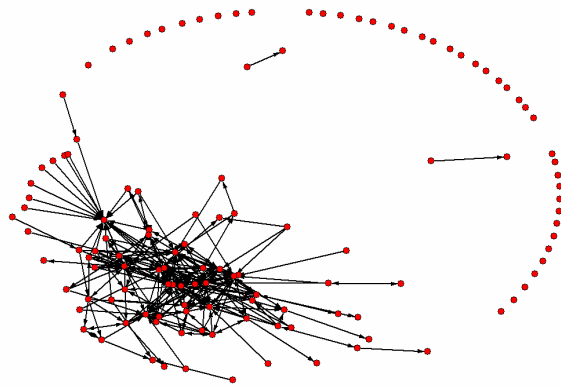


Figure 2
The NGO/IGO network (Shumate, 2008)

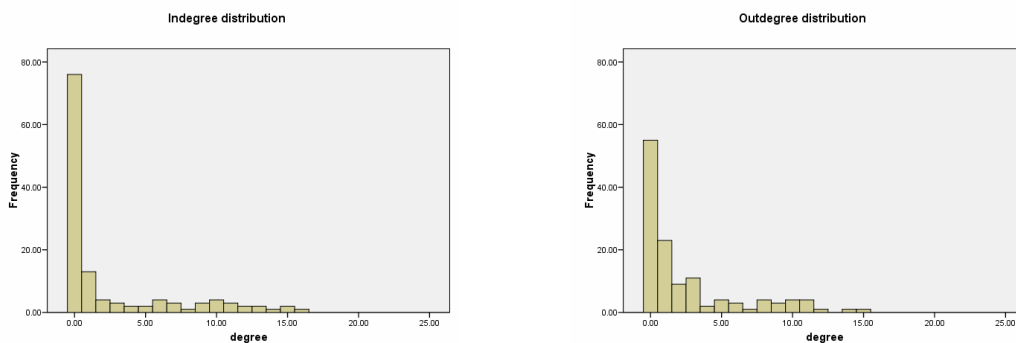


Figure 3
Degree distributions for the NGO network

Inverse power law degree distributions (“scale free”):

Barabási & Albert (1999) proposed an inverse power law degree distribution for networks with highly skewed degree distributions.

Let K be the degree of a randomly chosen node in the network. Then an inverse power law degree distribution is represented by:

$$P(K=k) \text{ is proportional to } k^{-\rho} \text{ (at least for large } k)$$

where ρ is a scaling parameter (greater than 1). It is common (e.g. Liljeros et al., 2003) to find that authors define the power law in terms of its complementary (for want of a better term; it has been referred to also as the “survival function”) *cumulative distribution function* rather than its *probability distribution function*, something which not only causes confusion but may also adversely affect resulting conclusions (Jones and Handcock, 2006).

This model for the degree distribution is influential in the physics network literature, where such networks are referred to as *scale free networks* (Barabási, 2002). A popular mechanism for generating such networks is the *preferential attachment model* (Albert & Barabási, 2002), whereby new nodes are added to the network with a new connection from the new node created to an existing node. The probability of the new connection relates to the degree of the existing nodes. This is essentially a model whereby popular nodes become more popular, thereby creating hubs. This model was first proposed in relational contexts by Simon (1955). Statistically, the preferential attachment model is represented by the Yule distribution (Yule, 1924).

Note that $\log P(K=k)$ is proportional to $-\rho \log k$. In the scale free network literature, it is common to plot an inverse power law degree distribution on a log-log scale to investigate the linearity of the relationship and to estimate ρ using ordinary least square regression. This procedure has been strongly criticized on statistical grounds by Handcock and Jones (2003). The linear fitting procedure cannot be motivated by any standard statistical principles and hence there is no clear relation between the model and fitting procedure. Methodologically sound alternatives, such as likelihood-based approaches (Handcock & Jones, 2004), are available for estimating the model parameters under the assumption that data has been generated by a power law distribution. Statistical fitting of power law distributions to actual data is however hampered by the sensitivity of the functional form of the probability distribution function to degrees close to the origin. Non-statistical approaches have avoided this issue by only stipulating power law behaviour for the tail of the distribution. The choice of the lower limit k_{\min} above which the power law is said to apply appears to be wholly arbitrary and its implication for distributional assumptions is unclear. Some rigour is introduced in the treatment of k_{\min} in Handcock and Jones (2003 and 2004) and Jones and Handcock (2006).

A rather trenchant critique of this literature has been presented by Li et al (2005), who argue that scale free networks do not even apply to the supposed paradigm case of the internet. Li et al attempt to establish some firm definitions and systematic approach to this body of research. What is quite clear, however, is that the supposed universality of scale free networks, a claim that is sometimes made, is not supported. Some networks

may be scale free but in each case this is an empirical question that needs to be investigated rather than presumed. Hancock and Jones (2003) have shown that careful fitting of the degree distribution against a variety of possible statistical models indicates that scale free networks are not always the best fitting model (see also Clauset, Shalizi & Newman, 2007). It is also sometimes claimed that degree distributions determine all other features of the network, so that only knowledge of the degree distribution is necessary to understand the properties of the network. This claim does not stand in the face of plenty of counter-examples (Li et al, 2005; Robins, Pattison and Woolcock, 2005; Snijders and van Duijn, 2002; Goodreau, 2007).

Figure 4 presents log-log plots of the degree distributions for the NGO network (with zero degree nodes excluded). It can be seen that despite the strong skew of the two distributions (Figure 3), and even using the less principled OLS log-log fitting procedure, an inverse power law distribution is not particular convincing for these data.

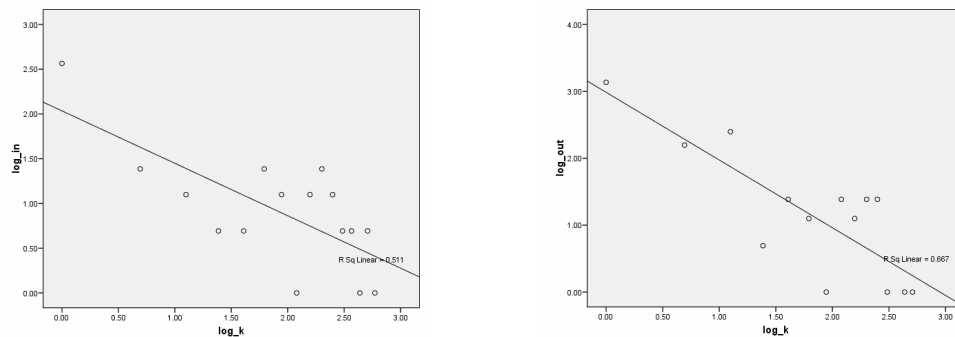


Figure 4
Log-log plots of the degree distributions for the NGO network

Controlling for degree distributions

Newman, Strogatz and Watts (2001) developed an approach to calculate a variety of graph properties for graph distributions that are random except for an arbitrary degree distribution. This is an approach that takes the degree distribution into account, rather than model it and follows a long tradition within social network analysis of using uniform graph distributions controlling for certain graph features as null models: in this case, the uniform graph distribution with fixed degree distributions ($U \mid \{x_{i+}\}, \{x_{+i}\}$ – see for instance, Snijders 1991). A generalized approach to the use of uniform graph distributions is provided by Pattison, Wasserman, Robins and Kanfer (2000).

Exponential random graph models

There is an advantage in including degree distribution parameters in a statistical network model alongside other effects. The importance of degree-based effects can then be examined alongside other possible processes such as homophily or triangulation.

Exponential random graph models are an appealing way to proceed but the degree parameterization possibilities to date are somewhat limited.

In Markov random graph models (Frank & Strauss, 1986), degree distribution information is essentially parameterised by the star parameters. The density parameter is equivalent to a parameter for mean of the degree distribution, while the 2-star parameter is equivalent to a parameter for the second moment of the distribution, the 3-star parameter equivalent to third moment, and so on. Hence, a model with full set of star parameters is equivalent to a model where degree distribution is fully parameterized (at least up to the largest degree, or the equivalent star parameter). However, this is expensive in terms of parameters and there are technical problems when the frequency of a given degree is zero. Of course, in dyadic independent versions of the model for directed networks, a full parameterization can be achieved through the p_1 model (Holland & Leinhardt, 1981).

The challenge remains to develop a parameterisation that expresses a given functional form to the degree distribution. In part, the combination of star parameters into the alternating k -star parameter, proposed by Snijders, Pattison, Robins and Handcock (2006) attempts to do this. As they show, this is equivalent to a parameterization for an exponentially weighted degree distribution (see also Hunter, 2007). These models can capture degree distributions well in the right circumstances (e.g. Goodreau, 2007; Robins, Pattison, & Wang, 2008).

The alternating k -star parameter can also produce skewed distributions. Figure 5 shows simulated degree distributions for models for 20 node graphs with fixed density (0.20) and an alternating k -star parameter with values -3 , 0 and $+3$. The boxplots present the frequency of nodes with a given degree across the simulated graphs. As can be seen, the positive parameter results in a small number of nodes with high degree and a large number of nodes with zero degree.

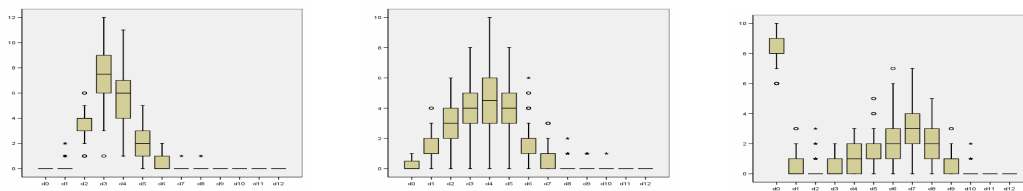


Figure 5
Simulations of degree distributions

Experience has also shown that a negative alternating k -star parameter together with a positive Markov 2-star parameter can result in positively skewed degree distributions without the appearance of many isolates, as in Figure 5 (Zhao, 2007). In addition, for directed networks, the inclusion of parameters specifically for isolated nodes, *sources* (nodes with zero indegree) and *sinks* (nodes with zero outdegree) can improve fit of the

degree distribution (Robins, Pattison & Wang, 2008). For directed networks, inclusion of a Markov 2-path (or mixed 2-star) parameter into a model may also improve fit of the correlation between in- and out-degree distributions. In principle it is possible that homogeneous Markov graph distributions may consistently (in the sense of expectancies for Markov graph models that are not *model* degenerate) produce graphs with skewed degree-distributions, for example with suitably chosen star-parameters up to the order $n-1$, but these may not always be estimable with anything less than parameter restrictions in a vein similar to the alternating k -star formulation.

Yet it is not universally the case that fit of degree distributions is good, especially for directed networks. For example, Table 1 provides parameter estimates and standard errors for a model for the NGO network. The model conditions on the density, as models that included a density parameter did not converge. This model fits many features of the network well, including the standard deviation and skew of the outdegree distribution, but not so for the indegree distribution. The observed standard deviation and the skew of the indegree distribution were both more than four standard deviations above the mean from a simulation of graphs from the model (for this goodness of fit approach, see Hunter, Goodreau & Handcock (2008); also Robins, Pattison and Wang, 2008). Nevertheless, models with additional parameters either do not converge or do not improve fit on the indegree distribution.

Table 1
Parameter estimates for the NGO network

<u>Parameter</u>	<u>Estimate</u>	<u>S.E.</u>
Reciprocity	2.65	0.24
Source	-0.87	0.42
Indegree (Alt. k -instar)	1.59	0.16
Outdegree (Alt. k -outstar)	0.38	0.20
Triangulation (AKT-TDU)	0.51	0.08

Conclusions and research issues

Despite attention paid to degree distributions in recent years, it is our conclusion that possible forms of the degree distribution of empirical networks have not been completely examined, and the resulting implications for networks not fully understood. Inclusion of degree-based parameters in statistical models is desirable but further work is necessary for better or more extended options for parameterization. For instance, Snijders et al (2006) proposed a possible alternative form of the sum of ascending factorials of degrees, in line with the Yule distribution, a suggestion that has not to date been taken up. In particular, there is not a clear comparison of the forms of the degree distribution and counterpart parameterization that might produce those forms, except in a small number of simpler cases. More work is needed to extend parameterization to enable better fit of degree distributions, especially for directed networks. The goal would be to achieve this fit without adding a huge number of star parameters that are difficult to interpret and may in practice prove difficult to converge.

References

- Albert, R., & Barabási, A-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics*, 74, 47-97.
- Barabási, A-L. (2002). *Linked: the new science of networks*. Cambridge, MA: Perseus.
- Barabási, A-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.
- Britton, T., Deijfen, M. and Martin-Löf, Anders, (2006). Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124, 1377-1397.
- Clauset, A., Shalizi, C., & Newman, M. (2007). Power law distributions in empirical data. Submitted to *SIAM Review*.
- Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81, 832-842.
- Goodreau S (2007) Advanced in exponential random graph (p^*) models applied to a large social network. *Social Networks*, 29, 231-248.
- Hagberg, Jan (2000). *Centrality Testing and Distribution of the Degree Variance in Bernoulli Graphs*, Research Report 2000:8, Department of Statistics, Stockholm University (<http://gauss.stat.su.se/site/pdf/jhlicnov30.pdf>)
- Hagberg, Jan (2003a). *Extreme Values and Other Attained Values of the Degree Variance in Graphs*, Research Report 2003:9, Department of Statistics, Stockholm University (http://gauss.stat.su.se/site/pdf/RR2003_9.pdf)
- Hagberg, Jan (2003b). *Random Graph Distributions of Degree variance*, Research Report 2003:8, Department of Statistics, Stockholm University (http://gauss.stat.su.se/site/pdf/RR2003_8.pdf)
- Hagberg, Jan (2003c). *General Moments of Degrees in Random Graphs*, Research Report 2003:7, Department of Statistics, Stockholm University (http://gauss.stat.su.se/site/pdf/RR2003_7.pdf)
- Handcock, M.S., Jones, J., 2004. Likelihood-based inference for stochastic models of sexual network formation. *Theoretical Population Biology* 65, 413–422.
- Handcock, M., & Jones, J. (2003). An assessment of preferential attachment as a mechanism for human sexual network formation. *Proceedings of the Royal Society, B*, 270, 1123-1128.
- Holland, P.W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.
- Hunter D (2007) Curved exponential family models for social networks. *Social Networks*, 29, 216-230.
- Hunter, D., Goodreau, S., & Handcock, M. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103, 248-258.
- Jones, J.H., and Handcock, M (2006). Interval estimates for epidemic thresholds in two-sex network models. *Theoretical Population Biology*, 70 (2), 125-134
- Li, L., Alderson, D., Tanaka, R., Doyle, J., & Willinger, W. (2005). *Towards a theory of scale-free graphs: Definition, properties, and implications (extended version)*. California Institute of Technology, Engineering and Applied Sciences Division, Technical Report CIT-CDS-04-006.
- Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E. & Åberg, Y. (2003). Authors' reply. *Nature* 423 (6940), 606
- Newman, M., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 69, 036122.
- Newman, M.E.J., Strogatz, S.H., & Watts, D.J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64, 026118.
- Pattison, P., Wasserman, S., Robins, G.L., & Kanfer, A. (2000). Statistical evaluation of algebraic constraints for social networks. *Journal of Mathematical Psychology*, 44, 536-568.
- Robins, G., Pattison, P., & Wang, P. (2008). Closure, connectivity and degrees: New specifications for exponential random graph (p^*) models for directed social networks. *Working paper: University of Melbourne*.
- Robins, G.L., Pattison, P.E., & Woolcock, J. (2005). Social networks and small worlds. *American Journal of Sociology*, 110, 894-936.
- Shumate, M. (2008). *The NGO/IGO network*. Personal communication.
- Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, 42, 435-450.
- Snijders, T.A.B. (1981a). The degree variance: An index of graph heterogeneity, *Social Networks*, 3, 163-174.
- Snijders, T.A.B. (1981b). *Maximum value and null moments of the degree variance*. TW-report 229, Departments of Mathematics, University of Groningen.

- Snijders, T. (1991). Enumeration and simulation models for 0-1 matrices with given marginals. *Psychometrika*, 56, 397-417.
- Snijders, T.A.B., Pattison, P., Robins, G.L., & Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36, 99-153.
- Snijders, T.A.B., & van Duijn, M.A.J. (2002). Conditional maximum likelihood estimation under various specifications of exponential random graph models. In Jan Hagberg (Ed.), *Contributions to social network analysis, information theory, and other topics: A festschrift in honour of Ove Frank* (pp. 117-134). University of Stockholm: Department of Statistics.
- Yule, G. (1924). A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, FRS. *Philosophical transactions of the Royal Society of London Series B – Biological Sciences*, 213, 21.
- Zhao, Y. (2007). *Multiple networks in organizations*. PhD thesis: University of Melbourne.