Measuring the shape of degree distributions

Dr Jennifer Badham
Visiting Fellow
SEIT, UNSW Canberra
research@criticalconnections.com.au

UNSW THE UNIVERSITY OF NEW SOUTH WALES SYDNEY · CANBERRA · AUSTRALIA | Canberra

Overview

Context

- What does 'shape' mean for degree distribution
- Why measure it?
  - Compare
  - Characterise
  - Relationship with other properties
- No agreed measure

Candidates to compare

- Variance / standard deviation / coefficient of variation
- Power law exponent
- Centralisation
- Gini coefficient

Variance (and its variants)

- Standard measure in statistics
- Width of 'peak'
  - Distance from mean
- Coefficient of variation (V_k) is scale invariant
- Snijders (1981) applied to degree

$$\sigma_k^2 = \frac{1}{N}\sum_{i=1}^{N}(k_i - \mu_k)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} k_i^2 - \mu_k^2$$

$$V_k = \frac{\sigma_k}{\mu_k}$$

Power law exponent

- Parameter of fitted distribution
  - Fitted to tail only
- How quickly degree probability declines
  - Long tail with small α
- Only for very large, skewed (eg WWW)
- Poor fitting (Clauset et al 2009)

$$p_k = Ck^{-\alpha} \quad k > 0$$

Centralisation

- Network specific measure
- Extent to which most central node is more central than others
  - Centrality = degree
- Only $k_{\max}$ and μ_k considered
- Freeman (1978); Butts (2006)

$$C = \frac{k_{\max} - \mu_k}{\min\left[N - 1 - \mu_k, \frac{\mu_k (N-2)}{2} \right]}$$

Gini coefficient

- Standard in equality measure for income
- Interpretations:
  - Expected difference in degree for random pair of nodes
  - Total distance from equality (Lorenz)
- Limited attention from SNA (except Hu & Wang 2008)

$$G = \frac{1}{\mu_k} \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| k_i - k_j \right|$$

$$G = 1 - 2 \int L$$

$$L : x_A = \sum_{k=0}^{A} p_k \quad y_A = \sum_{k=0}^{A} k p_k$$

Example networks (diverse)

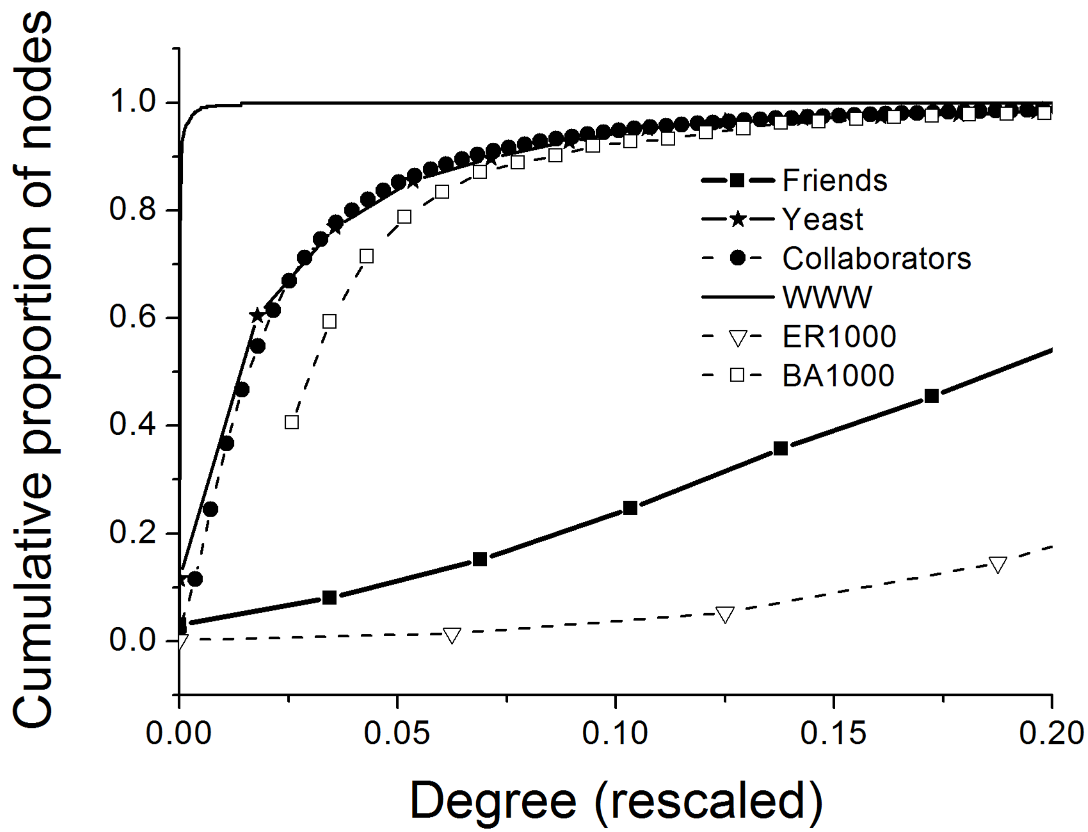
Empirical

- Friends: school
- Yeast: protein interactions
- Collaborators: condensed matter archive
- WWW: hyperlinks

Artificial

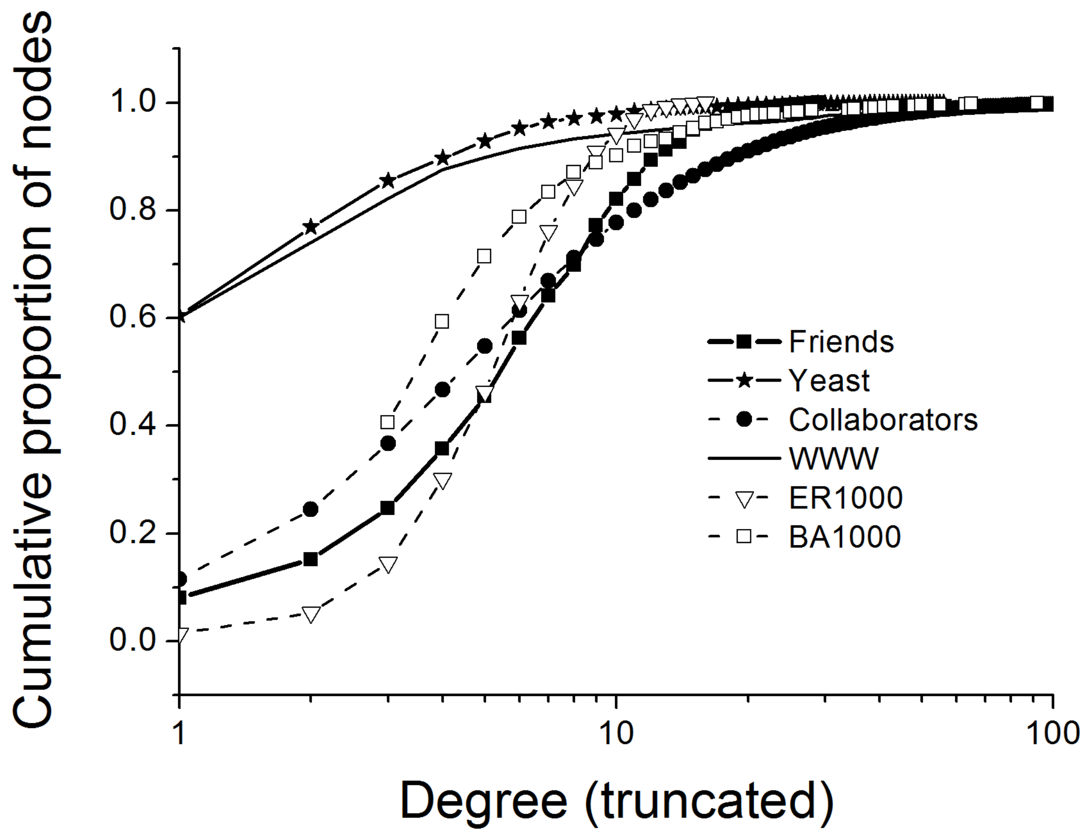
- BA1000: preferential attachment
- ER1000: fixed probability of edge
- Star1000: star with 1000 nodes

Example networks: distribution



UNSW | Canberra

Example networks: distribution



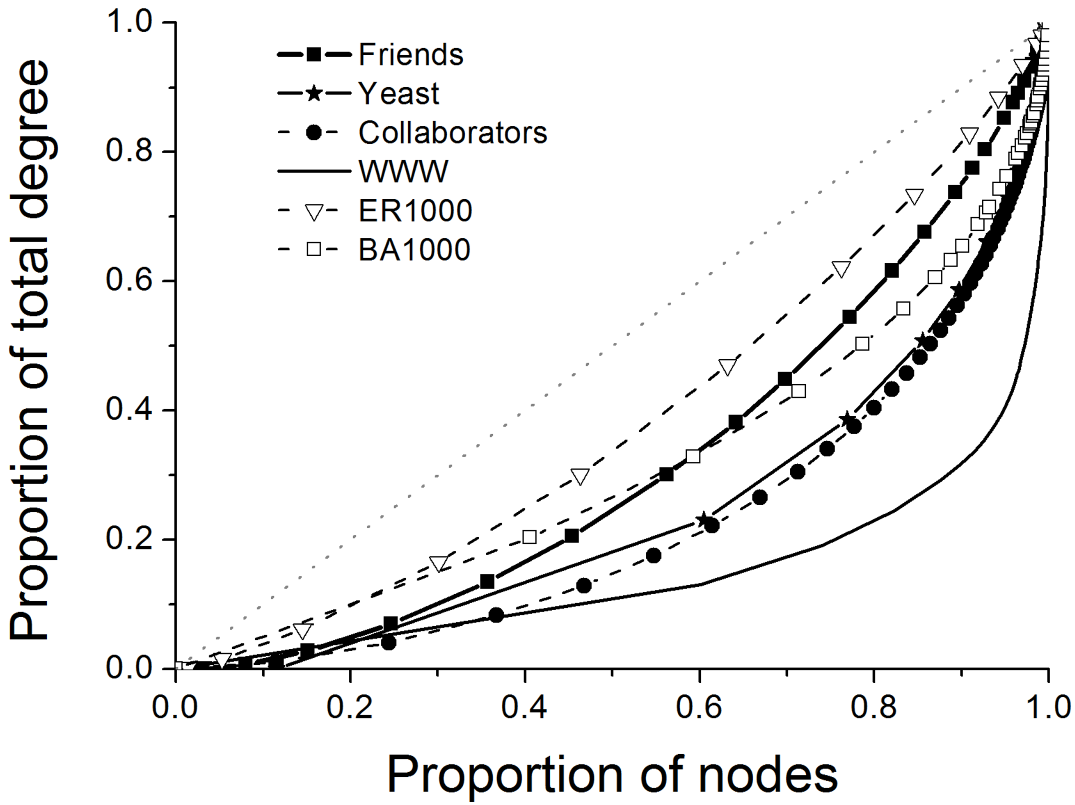
UNSW | Canberra

Comparison: example networks

Network	Size	h_2	C	V_k	J	α*	G
WWW (in)	N=325,729 μ_k=4.6	0.132	0.03	8.5	0.047	2.1	0.71
Collaborators	N=40,421 μ_k=8.7	0.055	0.01	1.5	0.031	-	0.55
Yeast (in)	N=2,114 μ_k=2.1	0.072	0.03	1.4	0.066	2.4	0.51
BA1000	N=1,000 μ_k=6.0	0.357	0.11	1.2	0.209	3.0	0.37
Friends (in)	N=859 μ_k=6.8	0.034	0.03	0.7	0.090	-	0.37
ER1000	N=1,000 μ_k=6.0	0.013	0.01	0.4	0.046	-	0.23
Star1000	N=1,000 μ_k=2.0	0.400	1.00	0.5	1.000	-	0.50

* A missing value for α indicates that it is not available in the literature, which may occur because the power law functional form is inappropriate or because it may be appropriate but was not reported. Unlike other measures in the table, a higher value indicates lower heterogeneity.

Example networks: Lorenz curves



Shape measure principles

- Objective is comparability: Must be sensible for all potential degree distribution shapes
- Relevant principles drawn from systematic evaluation for income inequality (Cowell 2000)
  - Transfer: Moving edges from high degree node to lower degree node reduces inequality (no reversal)
  - Addition: Increase all nodes by same number of edges should reduce (relative) or maintain (absolute) inequality
  - Replication: Multiple copies of all nodes has no effect

Comparison: principles

Measure	Transfer	Addition	Replication
Desirable	Decrease	Not Increase	No change
Mean (μ_k)	No change	Increase	No change
Coefficient of variation (V_k)	Decrease	Decrease	No change
Pareto proportion ($P_{0.1}$)	Varies	Decrease	No change
Herfindahl-Hirschman Index (H^2)	Decrease	Decrease	Decrease
Kolmogorov-Smirnov (KS)	Varies	No change	No change
Normalised Hierachization (h_2)	Decrease	Decrease	Decrease
Normalised Centralization (C)	No change	Increase	Decrease
Gini coefficient (G)	Decrease	Decrease	No change

Conclusion

- Only Gini (G) and Coefficient of Variation (V_k) meet principles
  - Centralisation unresponsive to transfers
  - Power law cannot always be fitted
- V_k not meaningful for skewed distributions, researchers use α for networks, G for income
- G intuitive mathematically (difference) and graphically (comparison to equality)
- Also relevant to other distributions (eg shortest path, betweenness, clustering coefficient)

References (measures)

- Butts CT. Exact bounds for degree centralization. Social Networks 2006; 28:283-96.
- Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. SIAM Review 2009; 51(4):661-703.
- Cowell FA. Measurement of inequality. In: Atkinson AB, Bourguignon F, eds. Handbook of Income Distribution: Elsevier; 2000. p. 87-166.
- Freeman LC. Centrality in social networks: conceptual clarification. Social Networks 1978; 1:215-39.
- Hu HB, Wang XF. Unified index to quantifying heterogeneity of complex networks. Physica A: Statistical Mechanics and its Applications 2008; 387(14):3769-80.
- Snijders TAB. The degree variance: An index of graph heterogeneity. Social Networks 1981; 3(3):163-74.

UNSW THE UNIVERSITY OF NEW SOUTH WALES SYDNEY · CANBERRA · AUSTRALIA | Canberra

References (example networks)

- Friends: Rapoport A, Horvath WJ. A study of a large sociogram. Behavioral Science 1961; 6(4):279-91.
- Yeast: Jeong H, Mason SP, Barabási A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001; 411(6833):41-42.
- Collaborators: Newman MEJ. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences of the United States of America 2001; 98:404-09.
- WWW: Albert R, Jeong H, Barabási A-L. Diameter of the World Wide Web. Nature 1999; 401(9 September 1999):130-31.
- BA1000: Barabási A-L, Albert R. Emergence of scaling in random networks. Science 1999; 286(5439):509-12.
- ER1000: Erdös P, Rényi A. On the evolution of random graphs. Publications of the Institute of Mathematics, Hungarian Academy of Science 1960; 5:17-60.