Measuring the shape of degree distributions

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

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

\[
\sigma_k^2 = \frac{1}{N} \sum_{i=1}^{N} k_i^2 - \mu_k^2
\]

\[
V_k = \frac{\sigma_k}{\mu_k}
\]
Power law exponent

- Parameter of fitted distribution
  - Fitted to tail only
- How quickly degree probability declines
  - Long tail with small $\alpha$
- 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 $k_{\text{max}}$ and $\mu_k$ considered
- Freeman (1978); Butts (2006)

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

\[
G = \frac{1}{\mu_k} \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |k_i - k_j|
\]

\[
G = 1 - 2 \int L
\]

\[
L : x_A = \sum_{k=0}^{A} p_k \quad y_A = \sum_{k=0}^{A} kp_k
\]
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
Example networks: distribution

Cumulative proportion of nodes

Degree (truncated)
Comparison: example networks

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

* A missing value for \( \alpha \) 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

<table>
<thead>
<tr>
<th>Measure</th>
<th>Transfer</th>
<th>Addition</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desirable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ($\mu_k$)</td>
<td>No change</td>
<td>Increase</td>
<td>No change</td>
</tr>
<tr>
<td>Coefficient of variation ($V_k$)</td>
<td>Decrease</td>
<td>Decrease</td>
<td>No change</td>
</tr>
<tr>
<td>Pareto proportion ($P_{0.1}$)</td>
<td>Varies</td>
<td>Decrease</td>
<td>No change</td>
</tr>
<tr>
<td>Herfindahl-Hirschman Index ($H^2$)</td>
<td>Decrease</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov ($KS'$)</td>
<td>Varies</td>
<td>No change</td>
<td>No change</td>
</tr>
<tr>
<td>Normalised Hierarchization ($h_2$)</td>
<td>Decrease</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
<tr>
<td>Normalised Centralization ($C'$)</td>
<td>No change</td>
<td>Increase</td>
<td>Decrease</td>
</tr>
<tr>
<td>Gini coefficient ($G'$)</td>
<td>Decrease</td>
<td>Decrease</td>
<td>No change</td>
</tr>
</tbody>
</table>
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 \( \alpha \) 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.
References (example networks)


