

# Centrality in Social Networks: Theoretical and Simulation Approaches

Dr. Anthony H. Dekker

Defence Science and Technology Organisation (DSTO)

Canberra, Australia

dekker@ACM.org

**Abstract.** Centrality is an important concept in the study of social networks, which in turn are important in studying organisational and team behaviour. For example, “central” individuals influence information flow and decision-making within a group. However, the relationship between mathematical measures of centrality on the one hand, and the real-world phenomenon of centrality on the other, is somewhat unclear. In this paper, we provide two additional perspectives: an analysis of real-world social-network data, and a study of networks produced by a simulation process. Comparing the two perspectives leads to recommendations on when to use different centrality measures.

## 1. INTRODUCTION

Social Network Analysis [1] is an important tool for studying organisational structures. When simulating organisations, Social Network Analysis concepts are of great value in interpreting the results [2]. Within Social Network Analysis, *centrality* is an important concept [1]. High centrality scores identify actors with the greatest structural importance in networks, and these actors would be expected to have a key role in simulated and real-world behaviour. This applies to networks of many different kinds [1].

Several methods for measuring centrality exist, and this raises the question: which of these methods should be used?

The difficulty in answering this question is that centrality is a sociological concept, but there is no well-defined *a priori* way of measuring it. In this, it resembles the concept of intelligence in individual psychology. Some of the current centrality measures have a venerable tradition, while others are newer. Some have their roots in mathematical concepts from graph theory [3],[4], while others are more ad hoc.

Because there is no well-defined sociological measure of centrality independent of the networks themselves, centrality scores for real-world networks cannot be compared against “true” centrality levels. Instead, the choice of centrality measures must come from comparing different network-based measures against each other, or from simulation experiments where centrality levels are known. In this paper, we use both approaches to compare five measures of centrality (all scaled to the range 0...1):

- Closeness centrality  $C_C$  is one of the most widely used centrality measures [1]. The closeness centrality of actor  $x$  is defined as the reciprocal of the average of  $d(x, y)$ :

$$C_C(x) = \frac{n-1}{\sum_{y \neq x} d(x, y)} = \frac{1}{AVG_{y \neq x} d(x, y)} \quad (1)$$

where  $n$  is the number of actors in the network, and  $d(x, y)$  is the shortest-path distance between actors  $x$  and  $y$ .

- Valued centrality  $C_V$  was introduced as an alternative to closeness centrality [5]. Although originally intended for valued networks, with ties of varying strength, it is equally applicable to ordinary networks. It is defined similarly to closeness centrality, but is the average of the reciprocal of  $d(x, y)$ , rather than the other way around:

$$C_V(x) = \frac{1}{n-1} \left( \sum_{y \neq x} \frac{1}{d(x, y)} \right) = AVG_{y \neq x} \left( \frac{1}{d(x, y)} \right) \quad (2)$$

- Jordan centrality  $C_J$  was introduced implicitly by Hage and Harary [4], and is derived from the “Jordan centre” of a network. It uses only the largest of the distances  $d(x, y)$ :

$$C_J(x) = \frac{1}{MAX_{y \neq x} d(x, y)} \quad (3)$$

Hage and Harary suggest that identifying the actors with the highest  $C_J$  can offer useful insights into a network.

- Betweenness centrality  $C_B$  is based on counting the number of geodesics (shortest paths)  $g_{xy}$  between actors  $x$  and  $y$ , and looking at the number  $g_{xy}(z)$  which travel via actor  $z$ :

$$C_B(z) = \frac{2}{(n-1)(n-2)} \sum_{x \neq z} \sum_{x < y \neq z} \left( \frac{g_{xy}(z)}{g_{xy}} \right) \quad (4)$$

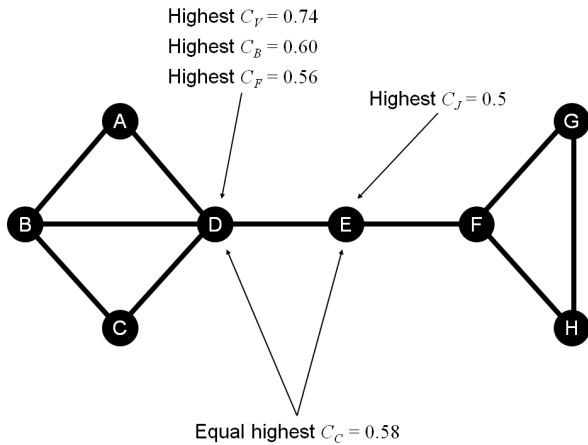
- Flow centrality  $C_F$  [6] provides an alternative to betweenness centrality which is more suitable for valued networks, based on the maximum flow [3]  $m_{xy}$  between actors  $x$  and  $y$ , and looking at the amount of flow  $m_{xy}(z)$  which travels via actor  $z$ :

$$C_F(z) = \frac{\sum_{x \neq z} \sum_{x < y \neq z} m_{xy}(z)}{\sum_{x \neq z} \sum_{x < y \neq z} m_{xy}} \quad (5)$$

The first three of these measures can be computed fairly easily, and there is an efficient algorithm for calculating betweenness centrality [7], but flow centrality is more time-consuming to compute for large networks.

There are also differences between the five centrality measures if we permit disconnected actors (or subsets of actors). If there is no path between actors  $x$  and  $y$ , we define the distance between them to be infinite, i.e.  $d(x, y) = \infty$  (taking  $1/\infty = 0$ ). Consequently, we have  $C_C(x) = C_J(x) = 0$  for every actor  $x$  in a network containing disconnected components. In addition, the betweenness centrality  $C_B(z)$  of actor  $z$  is technically undefined, because  $g_{xy} = 0$ . However, if we define  $0/0 = 0$  for the purpose of calculating betweenness centrality, meaningful betweenness scores can be obtained. For networks with disconnected actors, the measures  $C_V$  and  $C_F$  can be computed without problems.

Figure 1 shows the maximal values of the five centrality measures when applied to a simple network.



**Figure 1:** Five centrality measures applied to a simple network

Other definitions of centrality exist [8], which we have not included in our study. Degree centrality compares the degrees of nodes against the maximum possible degree for a network of the given size. However, degree centrality reflects only a local view of relationships between nodes, and does not provide information about overall network structure.

Bonacich's eigenvector centrality [8] is another interesting centrality measure which we did not include in this study, due to its complexity relative to the five measures listed above.

## 2. FACTOR ANALYSIS

Factor Analysis or Principal Components Analysis [9],[10] is a standard statistical technique used to identify relationships between different measures such as our five centrality measures. It is therefore the obvious method for comparing the centrality measures against each other. We applied Factor Analysis to the centrality scores of all the actors in a mixed set of ten connected social networks:

- the two island voyaging networks of [4] – with 12 and 10 actors respectively;
- the Florentine families network in [1] – with 15 actors, after deleting an isolate;
- two work communication networks – with 33 and 47 actors respectively;
- two Internet social networks – one from a newsgroup with 40 actors, and one from a blogging network with 25 actors;
- two random networks – with 12 actors each; and
- one binary tree network – with 15 actors.

Factor Analysis requires measures that are approximately normally distributed. The first three centrality scores do have approximately normal distributions, but betweenness centrality ( $C_B$ ) and flow centrality ( $C_F$ ) scores were strongly skewed towards zero. Such skewness requires a mathematical transformation to be applied prior to factor analysis. In the case of these two measures, the cube roots were approximately normally distributed (with absolute values of skew and kurtosis less than 0.8), and so a cube root transformation was applied prior to Factor Analysis.

Table 1 shows the correlations between the five measures (all correlations are statistically significant at the  $10^{-3}$  level or better):

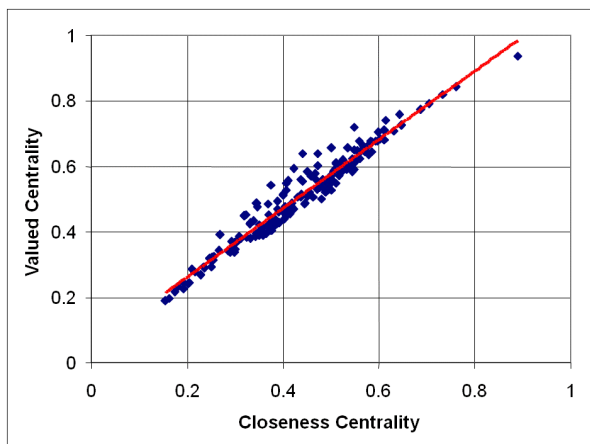
**Table 1:** Correlations between centrality measures for ten networks

	Valued centrality $C_V$	Jordan centrality $C_J$	Betweenness centrality $\sqrt[3]{C_B}$	Flow centrality $\sqrt[3]{C_F}$
<b>Closeness centrality</b> $C_C$	<b>0.97</b>	<b>0.91</b>	0.42	0.56
<b>Valued centrality</b> $C_V$	–	<b>0.83</b>	0.53	0.67
<b>Jordan centrality</b> $C_J$	–	–	0.25	0.38
<b>Betweenness centrality</b> $\sqrt[3]{C_B}$	–	–	–	<b>0.91</b>

The higher correlations shown bolded in Table 1 suggest that there are two slightly different things being measured by the five centrality measures. One corresponds to the first three measures  $C_C$ ,  $C_V$ , and  $C_J$ , and the other to the last two measures  $C_B$  and  $C_F$ . This distinction is indeed made in practice within Social Network Analysis, where the terms “closeness” (generally measured by  $C_C$ ) and “betweenness” (generally measured by  $C_B$ ) are generally distinguished. Our Factor Analysis provides additional justification for this distinction. Betweenness can be identified with a possible “bridging” or “brokering” role of actors [11].

The correlation between closeness centrality ( $C_C$ ) and valued centrality ( $C_V$ ) is a high 0.97, indicating that the two forms of centrality are measuring essentially the same thing. Figure 2 shows a scatter-plot for this relationship, which closely fits the regression equation:

$$C_V \approx 1.05 C_C + 0.05 \quad (6)$$



**Figure 2:** Valued centrality and closeness centrality have a correlation of 0.97 for ordinary networks

Factor Analysis confirms the division of the five centrality measures into two blocks. Table 2 shows the two main factors (principal components) identified by Factor Analysis, and the percentage of variance explained by these factors. The factor loadings indicate the contribution to each factor by the five centrality measures:

**Table 2:** Two main factors for centrality measures on ten networks

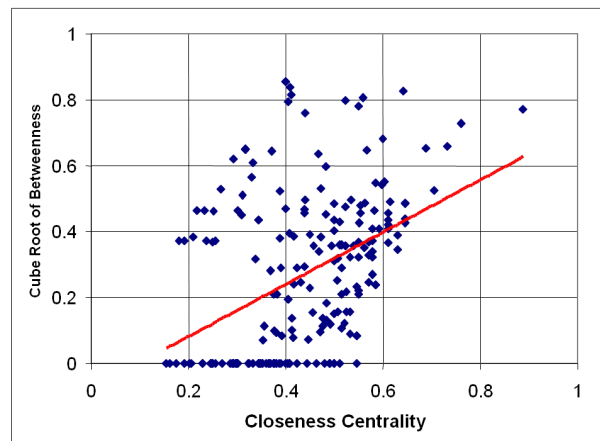
Variance	Factor Loadings				
	Closeness centrality $C_C$	Valued centrality $C_V$	Jordan centrality $C_J$	Betweenness centrality $\sqrt[3]{C_B}$	Flow centrality $\sqrt[3]{C_F}$
72%	0.49	0.51	0.43	0.37	0.43
24%	-0.31	-0.18	-0.47	0.62	0.51

Table 2 tells us that 72% of the variance in the data corresponds to a single concept, which we might call simply “centrality.” The factor loadings show that valued centrality ( $C_V$ ) is marginally the best single

measure for this concept (with a factor loading of 0.51), with closeness centrality ( $C_C$ ) coming a very close second (with a factor loading of 0.49).

A further 24% of the variance corresponds to the difference between the measures  $C_C$ ,  $C_V$ , and  $C_J$  on the one hand, and the measures  $C_B$  and  $C_F$  on the other hand – exactly the distinction between the “closeness” and “betweenness” concepts. As would be expected, betweenness centrality ( $C_B$ ) is the best single measure of the “betweenness” concept (with a factor loading of 0.62). The scatter-plot in Figure 3 conforms the distinction between the “closeness” and “betweenness” concepts.

Factor Analysis does not support Jordan centrality ( $C_J$ ) as a necessary addition to the set of centrality measures, although Hage and Harary [4] do provide an example where this measure does offer an additional insight. Flow centrality ( $C_F$ ) was of course developed by Freeman et al. [6] as an alternative to betweenness centrality ( $C_B$ ) for valued networks, so it is hardly surprising that it does not measure anything fundamentally different from “betweenness.”



**Figure 3:** Betweenness centrality and closeness centrality measure different things

### 3. VALUED NETWORKS

We next turned to valued networks (i.e. networks with ties of varying strength), to see if these conclusions would still hold. We used a mixed set of ten valued (and connected) networks:

- five work communication networks – with 20, 30, 47, 63, and 93 actors respectively;
- a subset of the Internet newsgroup network of [5] – with 182 actors;
- four random networks – two with 20 actors, and two with 30 actors.

We used the same five measures as before, but as indicated by Freeman et al. [6], betweenness centrality ( $C_B$ ) is not actually suitable for valued networks. One reason for this is that valued networks generally (at least 80% of the time) have only a single geodesic between a

pair of actors [5]. These geodesics are very sensitive to the strength of ties: a change in the value of a single tie changes which paths are geodesics, and thus alters the betweenness centrality ( $C_B$ ) for several actors. The process of collecting Social Network data always produces some slight uncertainties in tie strengths, due to variability in subjects filling out survey forms, or to observers failing to record interactions between actors 100% accurately. The combination of inevitable uncertainties and sensitivity to tie changes means that betweenness centrality scores for valued networks will contain a substantial random component. For this reason, flow centrality ( $C_F$ ) is a better choice for valued networks.

Table 3 shows the correlations between the five measures. The general pattern is the same as in Table 1. The correlation between closeness centrality ( $C_C$ ) and valued centrality ( $C_V$ ) is even higher than before (0.99 vs 0.97), with a slightly different regression equation:

$$C_V \approx 1.17 C_C \quad (7)$$

For valued networks, valued centrality ( $C_V$ ) is an extremely good alternative to closeness centrality ( $C_C$ ), as it was designed to be, while having the advantage of permitting disconnected actors. The correlation between the cube roots of betweenness centrality ( $C_B$ ) and flow centrality ( $C_F$ ) is lower than before (0.83 vs 0.91), indicating that the differences between them become more significant for valued networks.

**Table 3:** Correlations between centrality measures for ten valued networks

	Valued centrality $C_V$	Jordan centrality $C_J$	Betweenness centrality $\sqrt[3]{C_B}$	Flow centrality $\sqrt[3]{C_F}$
Closeness centrality $C_C$	<b>0.99</b>	<b>0.96</b>	0.35	0.50
Valued centrality $C_V$	–	<b>0.93</b>	0.40	0.55
Jordan centrality $C_J$	–	–	0.26	0.43
Betweenness centrality $\sqrt[3]{C_B}$	–	–	–	<b>0.83</b>

Table 4 shows the two main factors from Factor Analysis. The results are similar to Table 3: 70% of the variance corresponds to a single concept, “centrality.” A further 24% of the variance corresponds to the difference between “closeness” and “betweenness” concepts. In spite of the factor loadings, flow centrality ( $C_F$ ) should be used as the appropriate measure of the “betweenness” concept for valued networks, because of the problems with betweenness centrality we have noted.

**Table 4:** Two main factors for centrality measures on ten valued networks

Variance	Factor Loadings				
	Closeness centrality $C_C$	Valued centrality $C_V$	Jordan centrality $C_J$	Betweenness centrality $\sqrt[3]{C_B}$	Flow centrality $\sqrt[3]{C_F}$
70%	0.50	0.51	0.48	0.33	0.40
24%	–0.29	–0.23	–0.36	0.67	0.54

On both ordinary and valued networks, our Factor Analysis has thus supported the generally accepted view that the concept of “centrality” in Social Networks can be divided into two fundamentally different concepts, namely “closeness” and “betweenness.” The latter concept is best measured by betweenness centrality ( $C_B$ ) for ordinary (non-valued) networks, and flow centrality ( $C_F$ ) for valued networks, as recommended by Freeman et al. [6].

#### 4. SIMULATION EXPERIMENTS

We complement this factor analysis with four simulation experiments, creating networks where a designated node will be most central. The test for the various centrality measures is how easily they can recognise this most central node. In other words, the simulation experiments have well-defined *a priori* indications of centrality to which we can compare various centrality measures.

We will use four different kinds of (non-valued) networks, each with 41 nodes and an average degree of 4. The different kind of networks are illustrated in Figure 4. Type (a) networks result from placing links by a random process, with the designated node four times as likely to be chosen, and hence (on average) having four times the degree of the other nodes.

Type (b) networks are Scale-Free networks produced by the preferential attachment process of Barabási and Albert [12]. Here nodes are added one by one, attaching themselves to existing nodes chosen with probability proportional to their degree, but with double the probability for the designated node.

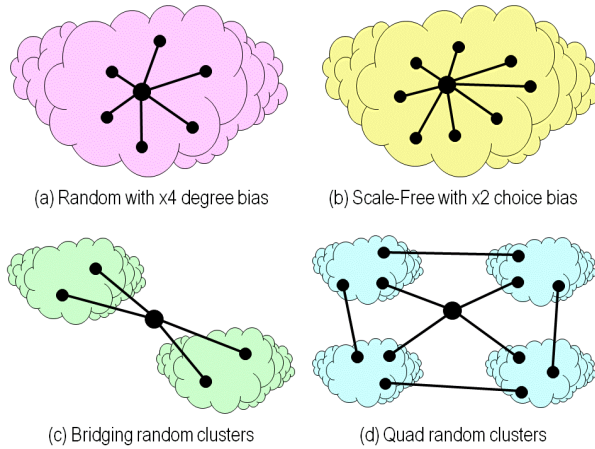
Type (c) networks are intended to have high “betweenness.” The designated node “bridges” two random clusters of 20 nodes each, and its degree is equal to the average of these clusters.

Finally, type (d) networks consist of four random clusters of 10 nodes each. The designated node bridges them, but does not provide the only connection between them.

For each of these processes, 100 different random networks were generated. If the resulting network was disconnected, the process was repeated, since otherwise analysis would simply reveal the fact that closeness centrality ( $C_C$ ) cannot be used with disconnected networks.

For each network, centrality scores were calculated for  $C_C$ ,  $C_V$ ,  $C_J$ , and  $C_B$ . Flow centrality ( $C_F$ ) was not

used, due to the computational effort required, and because it is not an appropriate measure for non-valued networks.



**Figure 4:** Four kinds of simulated network

In order to test the ability of different centrality measures to recognise the centrality of the designated node, we need to compare the calculated centrality of the designated node with the centrality scores of the population of remaining nodes. A standard way of making this comparison is to use a  $t$ -test. The  $t$ -statistic for a value  $x$  compared against a population of size  $n$ , with mean  $\mu$  and standard deviation  $\sigma$  is:

$$t = \frac{(x - \mu)\sqrt{n}}{\sigma} \quad (8)$$

For a population of size 40, values of  $t$  greater than 2.4 represent a difference statistically significant at the 1% level.

Table 5 shows the worst-case  $t$  values for the different centrality measures over the different groups of simulated networks. The two best performances in each row are highlighted.

**Table 5:** Worst-case  $t$  scores for centrality recognition over four sets of 100 simulated networks

Network	Closeness centrality $C_C$	Valued centrality $C_V$	Jordan centrality $C_J$	Betweenness centrality $C_B$
Degree bias (a)	<b>3.6</b>	<b>4.0</b>	-3.9	2.5
Scale-free (b)	5.2	<b>6.3</b>	-1.1	<b>6.9</b>
Bridge (c)	<b>14.9</b>	6.7	8.7	<b>16.1</b>
Quad (d)	<b>12.3</b>	7.7	6.7	<b>8.5</b>

Jordan centrality ( $C_J$ ) is, not surprisingly, unable to recognise centrality in type (a) networks, since centrality in those networks is purely a consequence of a higher

degree. Betweenness centrality ( $C_B$ ) is marginal for recognising centrality in these networks.

For type (b) networks, betweenness centrality ( $C_B$ ) and valued centrality ( $C_V$ ) perform best, with closeness centrality ( $C_C$ ) also good. Jordan centrality ( $C_J$ ) is again unable to recognise centrality in these networks.

For type (c) networks, which are designed to have high “betweenness,” betweenness centrality ( $C_B$ ) performs best, followed by closeness centrality ( $C_C$ ).

Finally, for type (d) networks, closeness centrality ( $C_C$ ) performs best, followed by betweenness centrality ( $C_B$ ).

Overall, Jordan centrality ( $C_J$ ) does not appear to be a particularly helpful centrality measure. Closeness centrality ( $C_C$ ) and valued centrality ( $C_V$ ) are the measures with broadest utility. Over all four kinds of network, valued centrality ( $C_V$ ) had the highest worst-case  $t$  score (4.0), with closeness centrality ( $C_C$ ) not far behind.

The specific concept of “betweenness,” highlighted in case (c), is best measured by betweenness centrality ( $C_B$ ).

## 5. DISCUSSION

Which centrality measure should analysts and simulators of organisational structure use? Our Factor Analysis and our simulation experiments have told complementary versions of the same story, supporting the generally accepted division of centrality into “closeness” and “betweenness” concepts. The latter concept is best measured by betweenness centrality ( $C_B$ ) for ordinary (non-valued) networks, and flow centrality ( $C_F$ ) for valued networks, as recommended by Freeman et al. [6], even though it is more time-consuming to compute.

For “closeness,” on the other hand, there are two measures that are both more or less equally good, for both valued and non-valued networks. They are the commonly used closeness centrality ( $C_C$ ) and the valued centrality ( $C_V$ ) introduced in [5]. The second of these measures has the advantage that it can deal with disconnected actors (or subsets of actors). It should therefore be used when disconnection is a potential issue. Neither Factor Analysis nor simulation support the use of Jordan centrality ( $C_J$ ), although Hage and Harary [4] do provide an example where this measure does offer a useful insight.

## REFERENCES

1. Wasserman, S. & Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

2. Dekker, A. (2007), "Studying Organisational Topology with Simple Computational Models,"
3. Gibbons, A. (1985), *Algorithmic Graph Theory*, Cambridge University Press.
4. Hage, P. & Harary, F. (1995), "Eccentricity and centrality in networks," *Social Networks*, **17**, 57–63.
5. Dekker, A.H. (2005), "Conceptual Distance in Social Network Analysis," *Journal of Social Structure*, **6** (3), [www.cmu.edu/joss/content/articles/volume6/dekker/](http://www.cmu.edu/joss/content/articles/volume6/dekker/)
6. Freeman, L.C., Borgatti, S.P. & White, D.R. (1991), "Centrality in valued graphs: A measure of betweenness based on network flow," *Social Networks*, **13**, 141–154.
7. Brandes, U. (2001), "A Faster Algorithm for Betweenness Centrality," *Journal of Mathematical Sociology*, **25** (2), 163–177: [www.inf.uni-konstanz.de/algo/publications/b-fabc-01.pdf](http://www.inf.uni-konstanz.de/algo/publications/b-fabc-01.pdf)
8. Borgatti, S. & Everett, M. (2006), "A Graph-theoretic perspective on centrality," *Social Networks*, **28** (4), 466–484.
9. Cohen, R.J., Swerdlik, M.E. & Phillips, C.M. (1988), *Psychological Testing and Assessment*, 3rd edition, Mayfield.
10. Kline, P. (1994), *An Easy Guide to Factor Analysis*, Routledge.
11. Krebs, V.E. (2002), "Uncloaking Terrorist Networks," *First Monday*, **7** (4), available online at [firstmonday.org/issues/issue7\\_4/krebs/](http://firstmonday.org/issues/issue7_4/krebs/)
12. Barabási, A.-L. & Albert, R. (1999), "Emergence of scaling in random networks," *Science*, **286**, 509–512.