Developments in Data Analysis
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 12, Ljubljana: FDV, 1996

Centrality in Social Networks

Vladimir Batagelj*

Abstract

In the paper an introduction to main social networks centrality measures is given. A new view on these measures is proposed, based on relational algebra. All described measures are implemented in computer programs `CENTRAL` and `FLOWIND`.

Keywords: social networks, centrality measures, closeness, betweenness, flow indices.

AMS Subj.Class.(1991): 92 H 30, 90 B 12, 05 C 12, 05 C 20, 68 R 10.

1 Mathematical prelude

Let $\mathcal{E}$ be a finite *set of units*, $n = \text{card}(\mathcal{E})$. A *binary relation* on $\mathcal{E}$ is any subset R of the Cartesian product $\mathcal{E} \times \mathcal{E}$. Instead of $(x, y) \in R$ we usually write xRy. By $R(x) = \{y \in \mathcal{E} : xRy\}$ we denote the *set of immediate successors* of unit x.

For example:

$$R = \{(a, d), (a, e), (d, c), (e, c), (e, e)\}$$

is a relation on $\mathcal{E} = \{a, b, c, d, e\}$.

We can represent a given relation by graph, by a matrix, or by the list of sets of immediate successors.

The representation by a graph, see Figure 1, is used to visualize the relation (provided that there are not too many units) and to support our intuition about several notions about relations. The representation by matrix is appropriate for computations – it enables us to use matrix algebra.

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

*University of Ljubljana, Department of Mathematics, Jadranska 19, 61 111 Ljubljana, Slovenia, e-mail: `vladimir.batagelj@uni-lj.si`

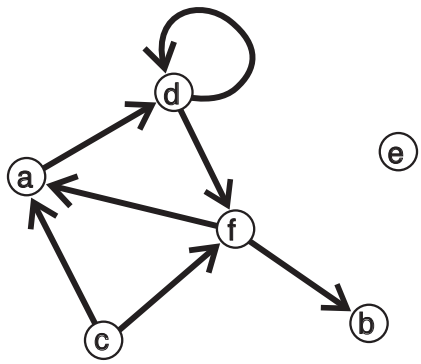

Figure 1: Graph of relation R

The representation by list of sets of immediate successors

$$\begin{array}{rclrcl} R(a) & = & \{d\} & R(d) & = & \{d, f\} \\ R(b) & = & \emptyset & R(e) & = & \emptyset \\ R(c) & = & \{a, f\} & R(f) & = & \{a, b\} \end{array}$$

is a basis for the usual input form for computer programs. Suppose that we numbered each element from the set of units $\mathcal{E}$ by its sequence number.

```
6  -1  4  -3  1  6  -4  4  6  -6  1  2  0
```

The first number equals the cardinality of set $\mathcal{E}$. A negative number $-i$ starts the list of immediate successors of node i. The description of a relation/graph is terminated by number 0.

We denote by $I = \{(x, y) : x \in \mathcal{E}\}$ the *identity* (diagonal) relation. Relation R is *reflexive* iff $I \subseteq R$; and it is *irreflexive* iff $I \cap R = \emptyset$.

Since relations are sets, all operations on sets $\cap, \cup, \setminus, \oplus$ can be used on them. Beside these, we can introduce also some special operations.

Inverse relation

$$R^{-1} = \{(x, y) : (y, x) \in R\}$$

Relation R is *symmetric* iff $R = R^{-1}$. We can represent pairs of opposite arcs (x, y) and (y, x) by an unoriented edge $(x\colon y)$.

Product of relations $R, S \subseteq \mathcal{E} \times \mathcal{E}$ is called the relation

$$R * S = \{(x, y) : \exists z \in \mathcal{E} : (xRz \wedge zSy)\}$$

The product operation $*$ is associative. Using it we can introduce the n-th *power* of relation R as $R^n = R * R * \ldots * R$. The relation R is *transitive* iff $R^2 \subseteq R$.

xR^ny iff there exists a walk of length n from x to y in graph $(\mathcal{E}, R)$.

Let Q be a property defined on relations. A Q-*closure* of relation R is the smallest relation R^Q, if it exists, which contains R and has the property Q.

The *symmetric closure* of R is denoted by $\hat{R}$. It holds

$$\hat{R} = R \cup R^{-1}$$

The relation R is symmetric iff $\hat{R} = R = R^{-1}$. In the graph of symmetric relation all arcs appear in opposite pairs – it can be viewed also as an undirected graph.

The *transitive closure* of R is denoted by $\overline{R}$

$$\overline{R} = \bigcup_{k \in \mathbb{N}^+} R^k$$

The relation R is transitive iff $\overline{R} = R$. $x\overline{R}y$ iff y is *reachable* from x in the graph of relation R. Therefore $\mathrm{card}(\overline{R}(x))$ equals the number of nodes reachable from the node x.

2 Node centrality measures

It seems that the most important distinction between centrality measures is based on the view/decision whether the relation is considered directed or undirected. This gives us two main types of centrality measures:

- directed case: *measures of importance*; with two subgroups: *measures of influence*, based on relation R; and *measures of support*, based on relation R^{-1};
- undirected case: *measures of centrality*, based on relation $\hat{R}$.

Since for symmetric relations $R = R^{-1} = \hat{R}$, in this case all three types of measures coincide.

Another important division of centrality measures is:

- *local* measures, which consider only immediate neighbors of a given node;
- *global* measures, which consider all nodes connected by paths with a given node.

Other general discussions of centrality measures can be found in [14, 10, 13, 15, 7, 6].

2.1 Degree

We consider some ways of constructing node measures. Let $S \subseteq \mathcal{E} \times \mathcal{E}$ be an irreflexive relation. We can base a measure on the cardinality of the set $S(x)$. To ensure comparability of measures for relations on different sets we normalize it by dividing it with the maximal possible value d_{max}. This gives us a measure

$$c(x; S) = \frac{\mathrm{card}(S(x))}{d_{max}}$$

It holds $0 \leq c(x; S) \leq 1$.

As special examples of this measure we get:

measure	definition	program `CENTRAL`
local influence	$c(x;R)$	`c_o`
local support	$c(x;R^{-1})$	`c_i`
local centrality	$c(x;\hat{R})$	`c`
global influence	$c(x;\overline{R})$	`ct_o`
global support	$c(x;\overline{R^{-1}})$	`ct_i`
global centrality	$c(x;\overline{\hat{R}})$	`ct`

We can define

$$d_{max} = \max\{\mathrm{card}(S(x)) : S \subseteq \mathcal{E} \times \mathcal{E} \wedge \mathrm{card}(\mathcal{E}) = n\}$$

In this case $d_{max} = n$; and for irreflexive relations $d_{max} = n - 1$. The relation $S \setminus I$ is always irreflexive. Note that for an irreflexive relation R its transitive closure $\overline{(R)}$ need not be irreflexive. In program `CENTRAL` we restrict to irreflexive relations by selecting the option `exclude diagonal`.

If the graph of relation R is not connected, other normalizations can be considered – for example, based on the maximal cardinality of *activity regions* $a(x;R) = \mathrm{card}(\overline{R}^{-1}(x) \cup \overline{R}(x))$ (implemented in program `CENTRAL`); or on the cardinality of the connected component to which a given node belongs.

2.2 Closeness

Let S be irreflexive. Sabidussi [17] introduced the following measure of *closeness*

$$cl(x;S) = \begin{cases} \dfrac{\mathrm{card}(\overline{S}(x))}{\sum_{y \in \overline{S}(x)} d(x,y)} \cdot \dfrac{\mathrm{card}(\overline{S}(x)) + 1}{n} & \overline{S}(x) \neq \emptyset \\ 0 & \overline{S}(x) = \emptyset \end{cases}$$

where $d(x,y)$ is the length of the shortest path from x to y. It holds $0 \leq cl(x;S) \leq 1$.

We distinguish

measure	definition	program `CENTRAL`
out-closeness	$cl(x;R)$	`cl_o`
in-closeness	$cl(x;R^{-1})$	`cl_i`
(total) closeness	$cl(x;\hat{R})$	`cl`

If graph is strongly connected, then $\mathrm{card}(\overline{S}(x)) = n - 1$. In this case

$$cl(x;S) = \frac{n-1}{\sum_{y \in \overline{S}(x)} d(x,y)}$$

2.3 Betweenness

Freeman [9, 10, 11] defined *betweenness* of a node x by

$$b(x;S) = \frac{1}{(n-1)(n-2)} \sum_{\substack{y,z \in \mathcal{E}: n(y,z) \neq 0 \\ y \neq z, x \neq y, x \neq z}} \frac{n(y,z;x)}{n(y,z)}$$

where $n(y,z)$ is the number of shortest paths from y to z and $n(y,z;x)$ is the number of shortest paths from y to z passing through x. It holds $0 \leq b(x;S) \leq 1$ and $b(x;S) = b(x;S^{-1})$.

We distinguish

measure	definition	program `CENTRAL`
betweenness	$b(x;R)$	`b`
(weak) betweenness	$b(x;\hat{R})$	`bw`

2.4 Flow measures

Degree, closeness and betweenness measures are based on the *economy assumption* – communications use the geodesics. Bonacich [3, 4] and Stephenson and Zelen [18] proposed two indices which consider all possible communication paths.

Let $a(x,y;m)$ denote the number of all walks from node x to node y of length not exceeding m. $A(m) = [a(x,y;m)]$. Therefore $A(1) = R$. Then the number of all walks from node x of length not exceeding m equals

$$v(x;m) = \sum_{y \in \mathcal{E}} a(x,y;m)$$

Let $v^*(m) = \max\{v(x;m) : x \in \mathcal{E}\}$ then the *Bonacich index* for node x is defined by

$$B(x) = \lim_{m \to \infty} \frac{v(x;m)}{v^*(m)}$$

It can be shown that $B(x)$ equals to the x-component of normalized ($\max_x B(x) = 1$) eigenvector of relational matrix R corresponding to the largest eigenvalue λ_{max}.

Stephenson and Zelen index is also based on the total flow in a symmetric and connected network, weighting each walk by the reciprocial of its length. It is difficult to give a short explanation of this index; see [18] for details.

Here is its definition in the general case. Let $A = [a(x,y)]$ be a symmetric network matrix. We compute from it a new matrix $B = [b(x,y)]$ determined by

$$b(x,y) = \begin{cases} 1 + \sum_{z \in \mathcal{E}} a(x,z) & x = y \\ 1 - a(x,y) & x \neq y \end{cases}$$

Let $C = [c(x,y)] = B^{-1}$ and define

$$k = \frac{1}{n} \sum_{y \in \mathcal{E}} (c(y,y) - 2c(x,y))$$

The value of k is identical for all $x \in \mathcal{E}$. Then the *Stephenson and Zelen* index for node x is determined by

$$I(x) = \frac{1}{k + c(x,x)}$$

This index is defined only for symmetric and connected networks.

Both indices can be computed by program `FLOWIND`.

3 Network centralization measures

Network centralization measures measure the extent to which the network supports/is dominated by a single node. They are usually constructed combining the corresponding node values. We have to consider two problems:

- is a node measure defined also for nonconnected networks; isolated nodes?
- is a measure comparable over different networks?

There are two general approaches.

3.1 Extremal approach

Let $p(x;S)$ be a node measure. We introduce the quantities

$$\begin{aligned} p^\star(S) &= \max_{x\in\mathcal{E}} p(x;S) \\ D(S) &= \sum_{x\in\mathcal{E}} (p^\star(S) - p(x;S)) \\ D^\star &= \max\{D(S) : S \subseteq \mathcal{E}\times\mathcal{E} \wedge \mathrm{card}(\mathcal{E}) = n\} \end{aligned}$$

Then we can define

$$p_{max}(S) = \frac{D(S)}{D^\star}$$

It can be shown [10, 3]

		$D^\star$	
measure	$p(x;S)$	general $S = R$	symmetric $S = \hat{R}$
degree/all	$c(x;S)$	$n-1$	$n-1$
degree/irreflexive	$c(x;S\setminus I)$	$(n-1)^2/n$	$n-2$
closeness	$cl(x;S\setminus I)$	$n-1$	$(n-1)(n-2)/(2n-3)$
betweenness	$b(x;S)$	$n-1$	
Bonacich	$B(x;S)$	$n-1-\sqrt{n-1}$	

and that among connected graphs in all these cases the (directed) star (with a loop in the root) is a maximal graph ($p_{max} = 1$) and the complete graph is a minimal graph ($p_{max} = 0$). Since activity $a(x;S)$ is defined only for general relations we have $D^* = \frac{(n-1)^2}{n}$ for all relations; and $D^* = \frac{(n-1)(n-2)}{n}$ for irreflexive relations.

3.2 Variational approach

The other approach is based on variance. First we compute the average node centrality

$$\overline{p}(S) = \frac{1}{n}\sum_{x\in\mathcal{E}} p(x;S)$$

and then define

$$p_V(S) = \frac{1}{n}\sum_{x\in\mathcal{E}} (p(x;S) - \overline{p}(S))^2$$

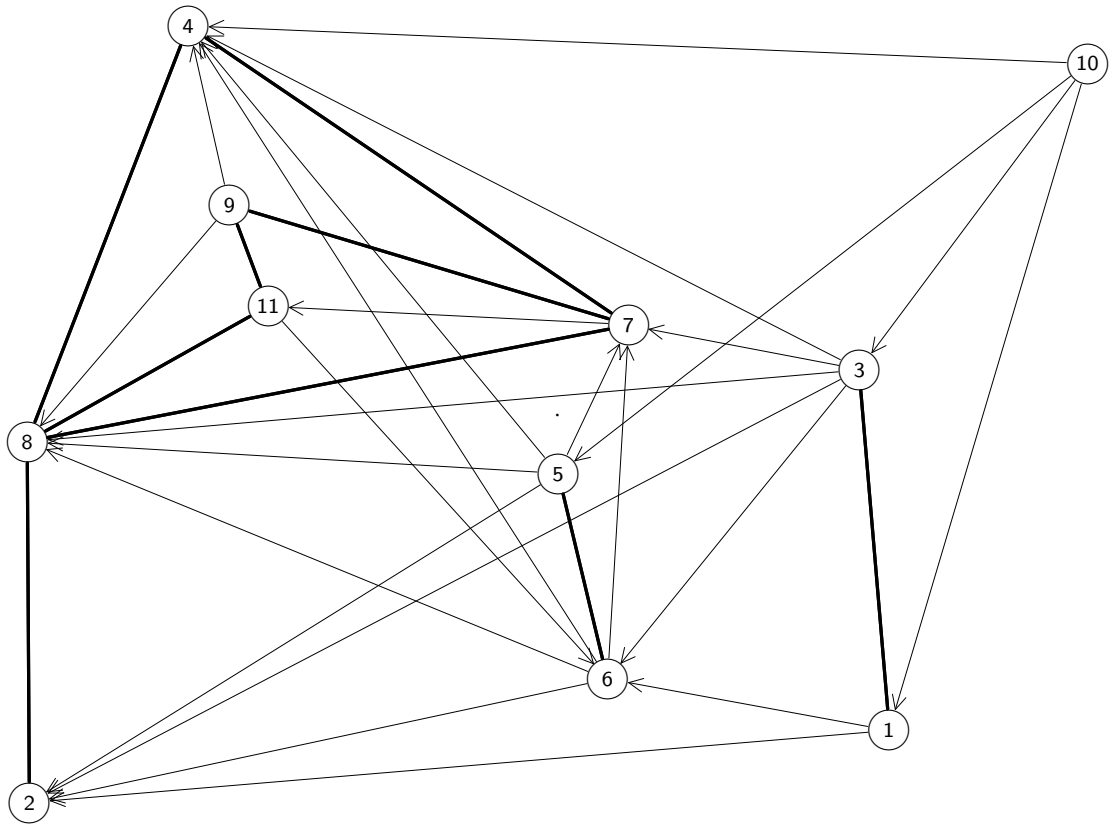

Figure 2: Network – discussion, recall

4 Algorithmic aspects

It can be shown [2] that transitive closure matrix and matrices of number and length of geodesics can be computed by Fletcher's algorithm [8] as closure matrices over corresponding semiring.

In computing of flow indices we need an algorithm for determining the largest eigenvalue and the corresponding eigenvector, and an algorithm for inverting a matrix. These algorithms can be found in any book on numerical analysis [5, 16].

5 Example

For an example the presented indices were applied to the Student Government Data, Discussion, recall [12]. The network is presented in Figure 2.

In Table 1 all degree indices are given, while the advanced indices are presented in Table 2. The relation was considered over all irreflexive relations and the indices were normalized over all units. All values of indices are multiplied by 100 – presented as percentages. For the largest eigenvalue of relation R we obtained $\lambda_{max}(R) = 3.28637$. The Stephenson-Zelen index is not computed since the network is not symmetric.

The interpretation of results is left to the reader.

	unit	$c(R^{-1})$	$c(R)$	$c(\hat{R})$	$c(\overline{R^{-1}})$	$c(\overline{R})$	$c(\overline{\hat{R}})$	$a(R)$
1	minister 1	20	30	40	20	90	100	100
2	minister 2	50	10	50	100	70	100	100
3	minister 3	20	60	70	20	90	100	100
4	minister 4	70	20	70	100	70	100	100
5	minister 5	20	50	60	100	70	100	100
6	minister 6	40	50	80	100	70	100	100
7	minister 7	60	40	70	100	70	100	100
8	minister 8	80	40	80	100	70	100	100
9	adviser 1	20	40	40	100	70	100	100
10	adviser 2	0	40	40	0	100	100	100
11	adviser 3	30	30	40	100	70	100	100
centralization		51.7	27.5	26.4	28.6	28.6	0.0	0.0

Table 1: Degree indices

	unit	$cl(R^{-1})$	$cl(R)$	$cl(\hat{R})$	$b(R)$	$b(\hat{R})$	$B(R)$
1	minister 1	27.27	48.13	58.82	0.74	1.50	60.66
2	minister 2	66.67	29.95	66.67	0.59	2.10	17.79
3	minister 3	27.27	68.18	76.92	4.70	7.83	100.00
4	minister 4	76.92	33.94	76.92	1.48	8.22	39.35
5	minister 5	40.00	56.57	71.43	2.22	4.49	81.54
6	minister 6	55.56	56.57	83.33	12.67	11.98	81.54
7	minister 7	71.43	46.28	76.92	12.11	5.54	70.84
8	minister 8	83.33	46.28	83.33	15.85	10.24	58.46
9	adviser 1	47.62	46.28	58.82	0.74	0.76	70.84
10	adviser 2	0.00	55.56	58.82	0.00	2.10	85.67
11	adviser 3	52.63	46.28	58.82	12.22	0.80	64.16
centralization		36.80	21.60	14.58	11.10	7.62	53.99

Table 2: Closeness, betweenness and Bonacich's index

References

[1] Batagelj V. (1991,1993): STRAN - programs for STRucture ANalysis. Manual, Ljubljana.

[2] Batagelj V. (1991): Semirings for Social Networks Analysis. Manuscript.

[3] Bolland J.M. (1988): Sorting out centrality: An Analysis of the performance of four cenrality models in real and simulated networks. Social Networks **10**, 233-253.

[4] Bonacich P. (1971): Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology **2**, 113-120.

[5] Conte S.D., de Boor C. (1981): Elementary numerical analysis, Third edition. Singapore: McGraw-Hill.

[6] Fabič-Petrač I., Jerman-Blažič B., Batagelj V. (1991): Study of Computation, Relatedness and Activity Prediction of Topological Indices. Journal of Mathematical Chemistry **8**, 121-134.

[7] Faust K., Wasserman S. (1991): Centrality and Prestige: A Review and Synthesis. Manuscript, revision May 16.

[8] Fletcher J.G. (1980): A more general algorithm for computing closed semiring costs between vertices of a directed graph. CACM, 350-351.

[9] Freeman L.C. (1977): A Set of Measures of Centrality Based on Betweenness. Sociometry, **40**/1, 35-41.

[10] Freeman L.C. (1978/79): Centrality in Social Networks, Conceptual Clarification. Social Networks, **1**, 215-239.

[11] Gould R.V. (1987): Measures of Betweenness in Non-symmetric Networks. Social Networks **9**, 277-282.

[12] Hlebec V. (1993): Recall versus recognition: Comparison of two alternative procedures for collecting social network data. Manuscript.

[13] Brigham R.C., Dutton R.D. (1985,1991): A Compilation of Relations between Graph Invariants. Networks **15**, 73-107; Supplement I. Networks **21**, 421-455.

[14] Høivik T., Gleditsch N.P. (1975): Structural parameters of graphs: A theorethical investigation. In H.M. Blalock, et al. (eds.), Quantitative Sociology, 203-223. New York: Academic Press.

[15] Hummon N.P., Doreian P., Freeman L.C. (1990): Analyzing the Structure of the Centrality-Productivity Literature Created Between 1948 and 1979. Knowledge: Creation, Diffusion, Utilization **11**/4, 459-480.

[16] Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. (1986): Numerical Recipes; The art of scientific computing. Cambridge: Cambridge University Press.

[17] Sabidussi G. (1966): The centrality index of a graph. Psychometrika, **31**, 581-603.

[18] Stephenson K., Zelen M. (1989): Rethinking Centrality: Methods and Examples. Social Networks **11**, 1-37.