Network Measures

Ulrik Brandes Thomas Erlebach (Eds.), *Network Analysis - Methodological Foundations*, Springer-Verlag Berlin Heidelberg, 2005

Centrality indices are to quantify an intuitive feeling that in most networks some vertices or edges are more central than others.

It should be noted that the term ‘centrality’ is by no means clearly defined. What is it that makes a vertex central and another vertex peripheral? For example, a vertex can be regarded as central if it is heavily required for the transport of information within the network or if it is connected to other important vertices. These few examples from a set of dozens other possibilities show that the interpretation of ‘centrality’ is heavily dependent on the context.

1.Degree

Undirected network:

$$c_{iD}(v) = d(v)$$

Directed network:

$$c_{iD}(v) = d^{-}(v)$$ in-degree centrality

$$c_{oD}(v) = d^{+}(v)$$ out-degree centrality

2. Facility Location Problems

The first family consists of those problems that use a minimax criterion. As an example, consider the problem of determining the location for an emergency facility such as a hospital. The main objective of such an emergency facility location problem is to find a site that minimizes the maximum response time between the facility and the site of a possible emergency.

2. Facility Location Problems

The second family of location problems optimizes a minisum criterion which is used in determining the location for a service facility like a shopping mall. The aim here is to minimize the total travel time.

A third family of location problems deals with the location of commercial facilities which operate in a competitive environment. The goal of a competitive location problem is to estimate the market share captured by each competing facility in order to optimize its location.

2.1 Eccentricity

The aim of the first problem family is to determine a location that minimizes the maximum distance to any other location in the network. Suppose that a hospital is located at a vertex $u \in V$. We denote the maximum distance from u to a random vertex v in the network, representing a possible incident, as the eccentricity $e(u)$ of u, where $e(u) = \max\{d(u, v): v \in V\}$. The problem of finding an optimal location can be solved by determining the minimum over all $e(u)$ with $u \in V$.

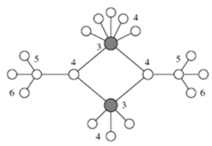

Fig. 3.1. Eccentricity values of a graph. Vertices in the center are colored in grey

2.1 Eccentricity – Equivalent Definition

$$c_E(u) = \frac{1}{e(u)} = \frac{1}{\max\{d(u,v) : v \in V\}}$$

This measure is consistent with our general notion of vertex centrality, since $e(u)^{-1}$ grows if the maximal distance of u decreases. Thus, for all vertices $u \in V$ of the center of G: $c_E(u) \geq c_E(v)$ for all $v \in V$.

2.2 Closeness

We consider the second type of location problems – the minisum location problem, often also called the **median problem** or **service facility location problem**. Suppose we want to place a service facility, e.g., a shopping mall, such that the total distance to all customers in the region is minimal. This would make traveling to the mall as convenient as possible for most customers. We denote the sum of the distances from a vertex $u \in V$ to any other vertex in a graph $G = (V,E)$ as the total distance

$$\sum_{v \in V} d(u,v)$$

The problem of finding an appropriate location can be solved by computing the set of vertices with minimum total distance.

2.2 Closeness

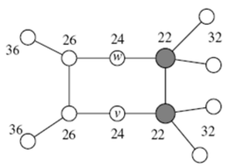

Fig. 3.2. Total distances of a graph. Lowest valued vertices are colored in grey. Note, the vertices v and w are more important with respect to the eccentricity

In social network analysis a centrality index based on this concept is called *closeness*. The focus lies here, for example, on measuring the closeness of a person to all other people in the network. People with a small total distance are considered as more important as those with a high total distance.

2.2 Closeness

$$c_C(u) = \sum_{v \in V} \frac{1}{d(u,v)}$$

$$c_R(u) = \frac{\sum_{v \in V} (\Delta_G + 1 - d(u,v))}{n-1}$$

where Δ_G and n denote the diameter of the graph and the number of vertices, respectively. The index measures how well a vertex is integrated in a network. The better a vertex is integrated the closer the vertex must be to other vertices. The primary difference between c_C and c_R is that c_R reverses the distances to get a closeness-based measure and then averages these values for each vertex.

2.3 Centroid Values

The last centrality index presented here is used in competitive settings: Suppose each vertex represents a customer in a graph. The service location problem considered above assumes a single store in a region. In reality, however, this is usually not the case. There is often at least one competitor offering the same products or services. Competitive location problems deal with the planning of commercial facilities which operate in such a competitive environment.

2.3 Centroid Values

For reasons of simplicity, we assume that the competing facilities are equally attractive and that customers prefer the facility closest to them. Consider now the following situation: A salesman selects a location for his store knowing that a competitor can observe the selection process and decide afterwards which location to select for her shop. Which vertex should the salesman choose?

2.3 Centroid Values

Given a connected undirected graph G of n vertices. For a pair of vertices u and v, $\gamma_u(v)$ denotes the number of vertices which are closer to u than to v, that is $\gamma_u(v) = |\{w \in V : d(u,w) < d(v,w)\}|$. If the salesman selects a vertex u and his competitor selects a vertex v, then he will have the following number of customers:

$$\gamma_u(v)+\frac{1}{2}\left(n-\gamma_u(v)-\gamma_v(u)\right)=\frac{1}{2}n+\frac{1}{2}\left(\gamma_u(v)-\gamma_v(u)\right)$$

2.3 Centroid Values

Thus, letting $f(u, v) = \gamma_u(v) - \gamma_v(u)$, the competitor will choose a vertex v which minimizes $f(u, v)$. The salesman knows this strategy and calculates for each vertex u the worst case, that is

$$c_F(u) = \min\{f(u, v): v \in V - u\}$$

$c_F(u)$ is called the *centroid* value and measures the advantage of the location u compared to other locations, that is the minimal difference of the number of customers which the salesman gains or loses if he selects u and a competitor chooses an appropriate vertex v different from u.

2.3 Centroid Values

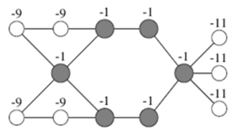

Fig. 3.4. All centroid values are negative. There is no profitable location for the salesman

Notice that for each vertex $u \in V$ in graph shown here $c_F(u) \leq -1$. Here, the salesman loses his advantage to choose as first. The strategy "choose after the leader has chosen" would be optimal.

3. Structural Properties

Center of a Graph. The eccentricity of a vertex $u \in G$ was defined as $e(u) = \max\{d(u, v): v \in V\}$. Recall, that by taking the minimum over all $e(u)$ we solve the emergency location problem. In graph theory, this minimum is called the ***radius*** $r(G) = \min\{e(u): u \in V\}$. Using the radius of G the ***center C(G) of a graph G*** is $C(G) = \{u \in V : r(G) = e(u)\}$.

3. Structural Properties

Median of a Graph. The service facility problem was solved by determining the set of vertices with minimum total distance. If the minimum total distance of G is denoted by $s(G) = \min\{s(u): u \in V\}$, the median $M(G)$ of G is given by

$$M(G) = \{u \in V : s(G) = s(u)\} .$$

3. Structural Properties

Centroid of a Graph. The computation of the centroid of a graph is a maximin optimization problem. We defined the centroid value for a given vertex u by $c_F(u) = \min\{f(u, v): v \in V - u\}$. In addition we call the objective function value $f(G) = \max\{c_F(u): u \in V\}$ the centroid value of G and denote by $Z(G) = \{u \in V : f(G) = c_F(u)\}$, the set of vertices representing the centroid of G.

4. Shortest Paths

4.1 Stress Centrality

$$c_S(v) = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \sigma_{st}(v)$$

where $\sigma_{st}(v)$ denotes the number of shortest paths containing v.

The number of shortest path that contain an element x gives an approximation of the amount of 'work' or 'stress' the element has to sustain in the network.

4. Shortest Paths

4.2 Shortest-Path Betweenness Centrality

$$c_B(v) = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$$

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$$

$\delta_{st}(v)$ denote the fraction of shortest paths between s and t that contain vertex v and σ_{st} denotes the number of all shortest-path between s and t.

4. Shortest Paths

4.2 Shortest-Path Betweenness Centrality

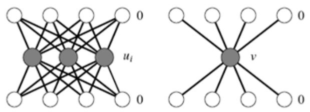

Fig. 3.6. $c_S(u_i) = 16$ and $c_B(u_i) = \frac{1}{3}$, $i = 1, 2, 3$ and $c_S(v) = 16$ but $c_B(v) = 1$. The graph shows on an example that stress centrality is not designed to evaluate how much communication control a vertex has

5. Vitality

Definition (Vitality Index). Let G be the set of all simple, undirected and unweighted graphs $G = (V, E)$ and $f : G \to R$ be any real-valued function on $G \in$ G. A vitality index $V(G, x)$ is then defined as the difference of the values of f on G and on G without element x: $V(G, x) = f(G) - f(G\backslash\{x\})$.

5. Vitality

5.1 Flow Betweenness Vitality

define the max-flow betweenness vitality for a vertex $u \in V$ by

$$c_{mf}(u) = \sum_{\substack{s,t \in V \\ u \neq s, u \neq t \\ f_{st} > 0}} \frac{f_{st}(u)}{f_{st}}$$

where $f_{st}(u)$ is the amount of flow which must go through u. We determine $f_{st}(u)$ by $f_{st}(u) = f_{st} - \tilde{f}_{st}$ where $\tilde{f}_{st}$ is the maximal s-t-flow in G \ u. That is, $\tilde{f}_{st}$ is determined by removing u form G and computing the maximal s-t-flow in the resulting network G \ u.

5. Vitality

5.2 Closeness Vitality

Wiener Index $I_W(G) = \sum_{v \in V} \sum_{w \in V} d(v, w)$

$$c_{CV}(x) = I_W(G) - I_W(G \setminus \{x\})$$

Let the distance between two vertices represent the costs to send a message from s to t. Then the closeness vitality denotes how much the transport costs in an all-to-all communication will increase if the corresponding element x is removed from the graph.

6. Feedback

Now we present centralities in which a node is the more central the more central its neighbors are.

6.1 Counting All Paths – The Status Index of Katz

To determine the importance or *status* of an individual in a social network where directed edges (*i, j*) can, for example, be interpreted as "*i votes for j*", it is not enough to count direct votes. If, e.g., only two individuals *k* and *l* vote for *i* but all other persons in the network vote either for *k* or for *l*, then it may be that *i* is the most important person in the network – even if she got only two direct votes. All other individuals voted for her indirectly.

6.1 Counting All Paths – The Status Index of Katz

To take the number of intermediate individuals into account, a damping factor $\alpha > 0$ is introduced: the longer the path between two vertices *i* and *j* is, the smaller should its impact on the status of *j* be.

The associated mathematical model is hence an unweighted (i.e. all weights are 1) directed simple graph $G = (V,E)$ without loops and associated adjacency matrix *A*.

6.1 Counting All Paths – The Status Index of Katz

Using the fact that $(A^k)_{ji}$ holds the number of paths from *j* to *i* with length *k* we hence have as status of vertex *i*

$$c_K(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k \left(A^k\right)_{ji}$$

In matrix notation we have

$$c_K = \sum_{k=1}^{\infty} \alpha^k \left(A^T\right)^k \mathbf{1}_n$$

$\mathbf{1}_n$ is the *n*-dimensional vector where every entry is 1.

6.1 Counting All Paths – The Status Index of Katz

Assuming convergence we find a closed form expression for the status index of Katz:

$$c_K = \sum_{k=1}^{\infty} \alpha^k \left(A^T\right)^k \mathbf{1}_n = \left(I - \alpha A^T\right)^{-1} \mathbf{1}_n$$

or, in another form

$$\left(I - \alpha A^T\right) c_K = \mathbf{1}_n$$

an inhomogeneous system of linear equations emphasizing the feedback nature of the centrality: the value of $\mathbf{c}_K(i)$ depends on the other centrality values $\mathbf{c}_K(j)$, $j \neq i$.

6.2 General Feedback Centralities

Bonacich's Eigenvector Centrality. In 1972 Phillip Bonacich introduced a centrality measure based on the eigenvectors of adjacency matrices. He presented three different approaches for the calculation and all three of them result in the same valuation of the vertices, the vectors differ only in a constant factor.

6.2 General Feedback Centralities

In the following we assume that the graph *G* to be analyzed is undirected, connected, loop-free, simple, and unweighted. As the graph is undirected and loop-free the adjacency matrix *A*(*G*) is symmetric and all diagonal entries are 0.

The three methods of calculation are:

a. the factor analysis approach ($\mathbf{s}^a$),

b. the convergence of an infinite sequence ($\mathbf{s}^b$), and

c. the solving of a linear equation system ($\mathbf{s}^c$).

6.2 General Feedback Centralities

We declare that $\mathbf{s}^a_i\, \mathbf{s}^a_j$ should be close to a_{ij} and interprete the problem as the minimization of the least squared difference. We are therefore interested in the vector $\mathbf{s}^a$ that minimizes the following expression:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(s_i^a s_j^a - a_{ij}\right)^2$$

6.2 General Feedback Centralities

A second approach presented by Bonacich is based on an infinite sequence. For a given λ_1 we define

$$\mathbf{s}^{b_0} = \mathbf{1}_n \qquad \mathbf{s}^{b_k} = A\frac{\mathbf{s}^{b_{k-1}}}{\lambda_1} = A^k\frac{\mathbf{s}^{b_0}}{\lambda_1^k}$$

According to the Theorem, the sequence

$$\mathbf{s}^b = \lim_{k\to\infty}\mathbf{s}^{b_k} = \lim_{k\to\infty} A^k\frac{\mathbf{s}^{b_0}}{\lambda_1^k}$$

converges towards an eigenvector $\mathbf{s}_b$ of the adjacency matrix A if λ_1 equals the largest eigenvalue.

6.2 General Feedback Centralities

The third approach follows the idea of calculating an eigenvector of a linear equation system. If we define the centrality of a vertex to be equal to the sum of the centralities of its adjacent vertices, we get the following equation system:

$$s_i^c = \sum_{j=1}^{n} a_{ij} s_j^c \qquad \mathbf{s}^c = A\mathbf{s}^c$$

This equation system has a solution only if $\det(A-I) = 0$. We solve $\lambda\mathbf{s} = A\mathbf{s}$, the eigenvalue problem for A, instead.

6.2 General Feedback Centralities

We have seen three methods for the calculation of the solution vectors $\mathbf{s}^a$, $\mathbf{s}^b$, $\mathbf{s}^c$. These vectors differ only by a constant factor. The eigenvector centrality is therefore (independently from the solution method) defined by:

$$c_{EV} = \frac{\left|\mathbf{s}^d\right|}{\left\|\mathbf{s}^d\right\|}$$

6.2 General Feedback Centralities

Everett and Borgatti explain this behavior via their *core-periphery model*, where in the idealized case the core corresponds to a complete subgraph and the nodes in the periphery do not interact with each other. To measure how close a graph is to the ideal core-periphery structure (or, in other words, how *concentrated* the graph is) they define the ρ-measure

$$\rho = \sum_{i,j} a_{ij}\delta_{ij}$$

with $\delta_{ij} = c_i c_j$, where a_{ij} are the components of the adjacency matrix and c_i measures the *coreness* of a node, $c_i \in [0, 1]$.

6.2 General Feedback Centralities

To determine the coreness of the nodes, the authors propose to minimize the sum of squared distances of a_{ij} and the product $c_i c_j$, which is nothing else than one approach to compute Bonacich's Standard Centrality. Hence nothing else then computing the principal eigenvector of the adjacency matrix. Thus, only the core-vertices get high c-values, nodes in smaller clusters not belonging to the core will get values near zero.

7. WEB Page Centrality - PageRank

PageRank PR is the prestige measure used by Google to rank Web pages. It is supposed to simulate the behavior of a user browsing the Web. Most of the time, the user visits pages just by surfing, i.e., by clicking on hyperlinks of the page he is on; otherwise, the user will jump to another page by typing its URL on the browser, or going to a bookmark, etc.

7. WEB Page Centrality - PageRank

On a graph, this process can be modeled by a simple combination of a random walk with occasional jumps toward randomly selected nodes. This can be described by the simple set of implicit relations:

$$p(i) = \frac{q}{n} + (1-q)\sum_{j:j\to i}\frac{p(j)}{k_{out}(j)} \qquad i = 1,2,\dots n$$

Here, n is the number of nodes of the graph, p(i) is the PR value of node i, k_{out}(j) the outdegree of node j, and the sum runs over the nodes pointing toward i. The *damping factor q* is a probability that weighs the mixture between random walk and random jump.

7. WEB Page Centrality - Eigenvector centrality

The EV is also based on the principle that the importance of a node depends on the importance of its neighbors. In this case the relationship is more straightforward than for PR: the prestige x_i of node i is just proportional to the sum of the prestiges of the neighboring nodes pointing to it,

$$\lambda x_i = \sum_{j:j\to i} x_j = \sum_j A_{ji} x_j = \left(\mathbf{A}^T\mathbf{x}\right)_i$$

we see that x_i is just the i component of the eigenvector of the transpose of the adjacency matrix with eigenvalue λ.