5 Advanced Centrality Concepts

Dirk Koschützki, Katharina Anna Lehmann,* Dagmar Tenfelde-Podehl, and Oliver Zlotowski*

The sheer number of different centrality indices introduced in the literature, or even only the ones in Chapter 3, is daunting. Frequently, a new definition is motivated by the previous ones failing to capture the notion of centrality of a vertex in a new application. In this chapter we will discuss the connections, similarities and differences of centralities. The goal of this chapter is to present an overview of such connections, thus providing some kind of map of the existing centrality indices. For that we focus on formal descriptions that hold for all networks. However, this approach has its limits.

Usually such approaches do not consider the special structure of the network that might be known for a concrete application, and it might not be able to convey the intuitive appeal of certain definitions in a concrete application. Nevertheless we consider such an approach appropriate to investigate the abstract definitions of different centrality indices. This is in a certain contrast to some of the literature, that only intuitively justifies a new definition of a centrality index on small example graphs.

Such connection between different definitions have been studied before, though usually not in a mathematical setting. One typical example is the work by Holme [304]. He considers a connection of betweenness centrality and congestion of a simulated particle hopping network. The particles are routed along shortest paths, but two particles are not allowed to occupy the same vertex. He investigates two policies of dealing with this requirement, namely that a particle waits if the next scheduled vertex is occupied, thus creating the possibility of deadlocks. Alternatively the particles can be allowed to continue their journey on a detour. He finds that such a prediction is only possible if the total number of particles in the network is small. Thus shortest-path betweenness for the application of the particle hopping model is the wrong choice, as it fails to predict congestion. In retrospect this is not really surprising because the definition of betweenness does not account for one path being blocked by another path, thus assuming that the particles do not interfere with each other. In particular the possibility of spill-backs as a result of overcrowded vertices is well known for car traffic flow on road networks, as for example addressed by the traffic-simulation presented by Gawron in [242]. Nagel [437] gives a more general overview of traffic considerations.

* Lead authors

U. Brandes and T. Erlebach (Eds.): Network Analysis, LNCS 3418, pp. 83–111, 2005.

Unfortunately, the only general lesson to be learned from this is that it does matter which precise definition of centrality one uses in a concrete application. This sheds another light on our attempts to classify centrality indices, namely to help identify the 'right' centrality index for a particular application. This is perhaps not possible in general, just because we have no idea what kind of applications might be of interest, and how the network is constructed. However, for a concrete application the considerations here might give valuable ideas on how to model the situation precisely or as a reasonable approximation.

In Section 5.1 we start with some general approaches to normalize centrality indices. Many of these techniques are so general that they can be applied to all indices presented in Chapter 3. We will differentiate between approaches that facilitate the comparison of centrality values within the same graph and between different graphs.

We then consider the possibility to modify a centrality index by letting it focus on a certain subset of vertices. This set can, e.g., be a subset of Web pages that a Web surfer is most interested in. With such a subset a ranking can be personalized to the interests of an user. This idea of personalization is explained in more detail in Section 5.2. As in the case of normalization some of the techniques are virtually applicable to all centrality indices presented in Chapter 3, whereas others are designed especially for only one centrality index.

An informal approach to structure the wide field of centrality indices presented in this book is given in Section 5.3. For that we dissect these indices into different components, namely a basic term, a term operator, personalization, and normalization and thereby we define four categories of centrality indices. This approach finally leads to a flow chart that may be used to 'design' a new centrality index.

Section 5.4 elaborates on fundamental properties that any general or application specific centrality index should respect. Several such properties are proposed and discussed, resulting in different sets of axioms for centrality indices.

Finally, in Section 5.5 we discuss how centrality indices react on changes on the structure of the network. Typical examples are experimentally attained networks, where a new experiments or a new threshold changes the valuation or even existence of elements, or the Web graph, where the addition of pages and links happens at all times. For this kind of modifications the question of stability of ranking results is of interest and we will provide several examples of centrality indices and their reactions on such modifications.

5.1 Normalization

In Chapter 3 we saw different centrality concepts for vertices and edges in a graph. Many of them were restricted to the nonnegative reals, and some to the interval $[0, 1]$, such as the Hub- & Authority-scores which are obtained using normalization with respect to the Euclidean norm.

The question that arises is what it means to have a centrality of, say, 0.8 for an edge or vertex? Among other things, this strongly depends on the maximum

centrality that occurs in the graph, on the topology of the graph, and on the number of vertices in the graph. In this section we discuss whether there are general concepts of normalization that allow a comparison of centrality scores between the elements of a graph, or between the elements of different graphs. Most of the material presented here stems from Ruhnau [499] and Möller [430].

In the following, we restrict our investigations to the centrality concepts of vertices, but the ideas can be carried over to those for edges.

5.1.1 Comparing Elements of a Graph

We start by investigating the question how centrality scores, possibly produced by different centrality concepts, may be compared in a given graph $G = (V, E)$ with n vertices. To simplify the notation of the normalization approaches we will use here a centrality vector instead of a function. For any centrality c_X, where X is a wildcard for the different acronyms, we will define the vector $\boldsymbol{c_X}$ where $\boldsymbol{c_X}_i = c_X(i)$ for all vertices $i \in V$. Each centrality vector $\boldsymbol{c_X}$ may then be normalized by dividing the centrality by the p-norm of the centrality vector

$$\|\boldsymbol{c_X}\|_p = \begin{cases} \left(\sum_{i=1}^{n} |\boldsymbol{c_X}_i|^p\right)^{1/p}, & 1 \le p < \infty \\ \max_{i=1,\ldots,n}\{|\boldsymbol{c_X}_i|\}, & p = \infty \end{cases}$$

to produce centrality scores $\boldsymbol{c_X}_i \le 1$.

The main difference between the p-norm for $p < \infty$ and $p = \infty$ (the maximum norm) is that, when normalizing using $p = \infty$, the maximum centrality score in the graph is 1, and this value is attained for at least one vertex. Therefore, the normalization using the maximum norm yields a 'relative' centrality for each vertex in a graph. Note that this normalization is not appropriate for comparing vertices in different graphs, since the value of 1 (or -1, if negative values are allowed) is attained in each graph, independent of its topology.

For $p < \infty$, the centrality concepts that may produce negative centrality scores (e.g. Bonacich's bargaining centrality, see Section 3.9.2) have to be treated in a special way. Möller [430] proposes to separate the negative and positive components:

$$\boldsymbol{c'_X}_i = \begin{cases} \boldsymbol{c_X}_i / \left(\sum_{j:\boldsymbol{c_X}_j > 0} |\boldsymbol{c_X}_j|^p\right)^{1/p}, & \boldsymbol{c_X}_i > 0, \\ 0, & \boldsymbol{c_X}_i = 0, \\ \boldsymbol{c_X}_i / \left(\sum_{j:\boldsymbol{c_X}_j < 0} |\boldsymbol{c_X}_j|^p\right)^{1/p}, & \boldsymbol{c_X}_i < 0. \end{cases}$$

Taking $p = 1$, this means (for non-negative centralities) that each of the vertices is assigned their associated percentage of centrality within a graph. It might be worth discussing whether a similar approach is reasonable when using the maximum norm – or whether one should normalize using the maximum value instead of the maximum absolute value. The latter would have the advantage that in each graph we would obtain a 1 as the maximal normalized centrality value.

A normalization with the p-norm is in general not appropriate for comparing vertices of different graphs. We will see that the Euclidean norm ($p = 2$) forms an exception for eigenvector centralities in that the maximal value that can be attained is independent of the number of vertices, see the end of Section 5.4.2.

5.1.2 Comparing Elements of Different Graphs

When vertices in different graphs have to be compared, the varying size of the graphs can be problematic. Let $\mathcal{G}_n$ be the set of connected graphs $G = (V, E)$ with n vertices. Freeman [227] proposed to define the *point-centrality*

$$\boldsymbol{c''_X}_i = \frac{\boldsymbol{c_X}_i}{\boldsymbol{c^*_X}}, \tag{5.1}$$

where $\boldsymbol{c^*_X} = \max_{G \in \mathcal{G}_n} \max_{i \in V(G)} \boldsymbol{c_X}_i$ is the maximum centrality value that a vertex can obtain taken over all graphs with n vertices.

Using the point-centrality $\boldsymbol{c''_X}_i$, the maximum value 1 is always attained by at least one vertex in at least one graph of size n. Thus, a comparison of centrality values in different graphs is possible. Unfortunately, this is often only possible in theory, since the determination of $\boldsymbol{c^*_X}$ is not trivial in general, and even impossible for some centrality concepts. Consider, for example, the status-index of Katz (see Section 3.9.1), where the centrality scores are related to the chosen damping factor. Theorem 3.9.1 states that the damping factor α is itself strongly related to the maximum eigenvalue λ_1 of the adjacency matrix. Hence, it is not clear that a feasible damping factor for the graph under investigation is also feasible for all other graphs of the same size.

Möller provides a nice example with the following two adjacency matrices:

$$A_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Since A_1^k is the zero matrix for $k \geq 2$, convergence is guaranteed for any $\alpha \in]0, 1]$. If we choose the maximum possible value $\alpha = 1$, then the infinite sum $\sum_{k=1}^{\infty} \alpha^k A_2^k$ does not converge, since it is equal to $\lim_{K \to \infty} \sum_{k=1}^{K} \mathbf{1}_2 \mathbf{1}_2^T$. This example shows that it is not clear which damping factor to choose in order to determine the value $\boldsymbol{c^*_K}$ (especially if we have to do that for different n).

Nevertheless, there are centrality concepts that allow the computation of $\boldsymbol{c^*_X}$. A very simple example is the degree centrality. It is obvious that in a simple, undirected graph with n vertices the maximum centrality value a vertex can obtain (with respect to degree centrality) is $n-1$. Another example is the shortest paths betweenness centrality (s. Section 3.4.2): The maximum value any vertex can obtain is given in a star with a value of $\frac{n^2-3n+2}{2}$ [227].

Further, the minimum total distance from a vertex i to all other vertices is attained when i is incident to all other vertices, that is, when i has maximum degree. So, it is clear that for the closeness centrality (see Section 3.2) we have $\boldsymbol{c^*_C} = (n-1)^{-1}$.

Möller shows that, in addition, the eccentricity centrality as well as the Hubs & Authorities centrality allow the calculation of the value $\boldsymbol{c}_{\boldsymbol{X}}^{*}$. For the eccentricity centrality we just note that a vertex with maximum degree has an eccentricity value of 1 and all other vertices have smaller eccentricity values, hence $\boldsymbol{c}_{\boldsymbol{E}}^{*} = 1$. Similarly, the maximum values for hub- and authority centrality values (centrality vectors are assumed to be normalized using the Euclidean norm) are 1 and they are attained by the center of a star (either all edges directed to the center of the star or all edges directed away from the center).

Shortest-path betweenness centrality and the Euclidean normalized eigenvector centrality provide other, more sophisticated, examples, see, e.g., Ruhnau [499]: These two centralities have the additional property that the maximum centrality score of 1 is attained exactly for the central vertex of a star. This property is useful when comparing vertices of different graphs, and is explained in more detail in the Section 5.4.2.

Finally we note that Everett, Sinclair and Dankelmann found an expression for the maximum betweenness in bipartite graphs, see [195].

5.2 Personalization

The motivation for a personalization of centrality analysis of networks is easily given: Imagine that you could configure your favorite Web search engine to order the WWW according to your interests and liking. In this way every user would always get the most relevant pages for every search, in an individualized way.

There are two major approaches to this task: The first is to change weights on the vertices (pages) or edges (links) of the Web graph with a personalization vector $\boldsymbol{v}$. The weights on vertices can describe something like the time spent each day on the relevant page and a weight on the edge could describe the probability that the represented link will be used. With this, variants of Web-centrality algorithms can be run that take these personal settings into account. The other approach is to choose a 'rootset' $R \subseteq V$ of vertices and to measure the importance of other vertices and edges relative to this rootset.

We will see in Section 5.3 that these two approaches can be used as two operators. The first approach changes the description of the graph itself and the corresponding operator is denoted by $P_{\boldsymbol{v}}$. Then the corresponding term for each vertex (or edge) is evaluated on the resulting graph. The second personalization approach chooses a subset of all terms that is given by the rootset R. This operator is denoted by P_R.

We will first discuss personalization approaches for distance and shortest paths based centralities and then discuss approaches for Web centralities.

5.2.1 Personalization for Distance and Shortest Paths Based Centralities

All centralities that were presented in Chapter 3 rank every vertex relative to all other vertices in the graph. In this subsection we will be concerned with

variants of these centralities that determine the relative importance of vertices with respect to a set R of root vertices. R is chosen such that the vertices in R are assumed to be important and the question is how all other vertices should be ranked in importance with respect to R. The approach presented by White and Smith in [580] is very general and deserves some attention.

Let $c(v)$ be some centrality index on vertices. Then, $c(v|R)$ denotes the relative importance of vertex v with respect to the given rootset R. Let $P(s,t)$ denote any well defined set of paths between vertex s and t. The authors suggest different kinds of path sets:

- a set of shortest paths
- a set of k-shortest paths, defined as the set of all paths with length smaller than a given k
- a set of k-shortest vertex-disjoint paths[1]

The set of shortest paths is used e.g. in the shortest-path betweenness centrality (see Section 3.4.2). The *relative betweenness centrality* $c_{RBC}(v)$ can be defined in three ways. In the first variant we define a vertex v as important if the fraction of shortest paths leaving a vertex r from R contains v. We will denote this *source relative betweenness centrality* by

$$c_{sRBC}(v) = \sum_{r \in R} \sum_{t \in V} \delta_{rt}(v) \ . \tag{5.2}$$

If an element v is important if it is contained in a large fraction of shortest paths ending in a vertex r of R we denote the *target relative betweenness centrality* as

$$c_{tRBC}(v) = \sum_{s \in V} \sum_{r \in R} \delta_{sr}(v) \ . \tag{5.3}$$

In the last case, an element is supposed to be important if it is contained in a large fraction of shortest paths leading from R to R, denoted by

$$c_{RBC}(v) = \sum_{r_s \in R} \sum_{r_t \in R} \delta_{r_s r_t}(v) \ . \tag{5.4}$$

If any other set of paths $P(s,t)$, e.g. the set of k-shortest paths, is chosen, then the definition of $\delta_{st}(v)$ has to be changed, denoted by

$$\delta_{st|P}(v) = \frac{\sigma_{st}(v)}{|P(s,t)|} \tag{5.5}$$

where $\sigma_{st}(v)$ denotes the number of paths $p \in P(s,t)$ that contain vertex v.

This example demonstrates the general idea behind this kind of personalization. It can be easily expanded to all centralities that are based on distance.

[1] We just want to note that this set of paths is not unique in most graphs. For a deterministic centrality it is of course important to determine a unique path set, so this last path set should only be used on graphs where there is only one set for each vertex pair.

5.2.2 Personalization for Web Centralities

Consider again the random surfer model (see Section 3.9.3) for Web centralities and assume the random surfer arrived at a page where there is no outlink or where the existing out links are not relevant. The original assumption in this case is a jump to a random page where each page has equal probability. It is obvious that the assumption of equal probability is not very realistic: some surfers prefer Web pages about sports if they get stuck in a sink, others continue with a news-page etc. The question at hand is hence how to model the many different types of Web users.

A very intuitive approach is to replace $\mathbf{1}_n$ (cf. Equation 3.44) by a *personalization vector* $\boldsymbol{v}$ satisfying $v_i > 0 \; \forall \; i$ and $\sum_i v_i = 1$. White and Smyth [580], for example, proposed to score the vertices relative to a kernel set R using

$$v_i = \begin{cases} \frac{1-\varepsilon}{|R|}, & i \in R \\ \frac{\varepsilon}{n-|R|}, & i \notin R \end{cases},$$

where $0 < \varepsilon << 1$.

They also proposed a very similar approach for the Hubs & Authorities algorithm. Instead of applying the iterative procedure given in Algorithm 2 on page 55 they added in each step a portion of the personalization vector and obtained the following modified equations:

$$\begin{aligned} \boldsymbol{c}^k_{\text{HA-H}} &= dA_\sigma \boldsymbol{c}^{k-1}_{\text{HA-A}} + (1-d)\boldsymbol{v} \\ \boldsymbol{c}^k_{\text{HA-A}} &= dA^T_\sigma \boldsymbol{c}^k_{\text{HA-H}} + (1-d)\boldsymbol{v} \\ \boldsymbol{c}^k_{\text{HA-H}} &= \frac{\boldsymbol{c}^k_{\text{HA-H}}}{\|\boldsymbol{c}^k_{\text{HA-H}}\|} \\ \boldsymbol{c}^k_{\text{HA-A}} &= \frac{\boldsymbol{c}^k_{\text{HA-A}}}{\|\boldsymbol{c}^k_{\text{HA-A}}\|} \end{aligned},$$

where $d \in [0,1]$ is chosen to control the influence of $\boldsymbol{v}$.

Going back to the PageRank algorithm it is clear that as long as all elements of $\boldsymbol{v}$ are positive and $\boldsymbol{v}$ is a stochastic vector, the associated Markov chain is still irreducible hence the convergence of the PageRank algorithm is not touched. Thus, at a first glance, this approach seems to be appealing. But there is one big disadvantage: As already known the computations of PageRank vectors for the non-personalized version is very time consuming, there is, at least at the moment, no chance to compute PageRank centralities for many different types of Web users. Nevertheless there are some promising approaches to obtain personalized PageRank vectors in an adequate amount of time.

To this end we give a general approach of personalization for PageRank, taken from Haveliwala et al. [291]. As noted above the personalized PageRank vector is given as the solution of the following equation

$$\boldsymbol{c}_{\text{PR}} = dP^T \boldsymbol{c}_{\text{PR}} + (1-d)\boldsymbol{v}.$$

Since $\left(I - dP^T\right)$ is a strictly diagonally dominant matrix, it is invertible and hence

$$\boldsymbol{c}_{\mathrm{PR}}^{\boldsymbol{v}} := \boldsymbol{c}_{\mathrm{PR}} = (1-d)\left(I - dP^T\right)^{-1}\boldsymbol{v} =: Q\boldsymbol{v}. \tag{5.6}$$

(We write $\boldsymbol{c}_{\boldsymbol{PR}}^{\boldsymbol{v}}$ to emphasize the dependence of $\boldsymbol{c}_{\boldsymbol{PR}}$ on $\boldsymbol{v}$.)

If we choose $\boldsymbol{v}$ to be the ith unit vector $\boldsymbol{v} = \boldsymbol{e}^i$, then $\boldsymbol{c}_{\mathrm{PR}}^{\boldsymbol{e}^i} = Q_{\cdot j}$, hence the set of columns of Q may be seen as a basis for the personalized PageRanks.

The Problem that occurs is that the determination of Q needs to invert a matrix which is very time consuming if the matrices are large. To reduce the computational complexity Q is approximated by $\hat{Q} \in \mathbb{R}^{n\times K}$ and hence we consider only a subset of K basis vectors (independent columns of Q) taking a convex combination to obtain an estimate for

$$\boldsymbol{c}_{\mathrm{PR}}^{\boldsymbol{w}} = \hat{Q}\boldsymbol{w}$$

where $\boldsymbol{w} \in \mathbb{R}^K$ is a stochastic vector, $w_i > 0\ \forall\ i$.

Haveliwala et al. show that the following three personalization approaches can be subsumed under the general approach described above:

– Topic sensitive PageRank [289],
– Modular PageRank [326],
– BlockRank [339].

They only differ in how the approximation is conducted. We describe these approaches briefly in the following subsections.

Topic Sensitive PageRank. Haveliwala [289] proposes to proceed in a combined offline-online algorithm where the first phase (offline) consists of the following two steps

1. Choose the K most important topics $t_1, \ldots, t_K$ and define v_i^k to be the (normalized) degree of membership of page i to topic t_k, $i = 1, \ldots, n,\ k = 1, \ldots, K$.
2. Compute $\hat{Q}_{\cdot k} = \boldsymbol{c}_{\mathrm{PR}}^{\boldsymbol{v}^k}$, $k = 1, \ldots, K$

The second phase that is run online is as follows

1. For query σ compute (normalized) topic-weights $w_1^\sigma, \ldots, w_K^\sigma$
2. Combine the columns of $\hat{Q}$ with respect to the weights to get

$$\boldsymbol{c}_{\mathrm{PR}}^{\sigma} = \sum_{k=1}^{K} w_k^\sigma \hat{Q}_{\cdot k}.$$

Note that to apply this approach it is important that

– K is small enough (e.g. $K = 16$) and
– the range of topics is broad enough.

Modular PageRank. A second approach was proposed by Jeh and Widom [326]. Their algorithm consists of an offline and an online step. In the offline step K pages $i_1, \ldots, i_K$ with high rank are chosen. These high-ranked pages form the set of hubs.

Using personalization vectors $\boldsymbol{e}^{i_k}$, the associated PageRank vectors called *basis vectors* or *hub vectors* $\boldsymbol{c}_{\mathrm{PR}}^{\boldsymbol{e}^{i_k}}$ are computed. By linearity for each personalization vector $\boldsymbol{v}$ that is a convex combination of $\boldsymbol{e}^{i_1}, \ldots, \boldsymbol{e}^{i_K}$ the corresponding personalized PageRank vector can be computed as a convex combination of the hub vectors. But if K gets larger, it is neither possible to compute all hub vectors in advance nor to store them efficiently. To overcome this deficiency, Jeh and Widom propose a procedure using *partial vectors* and a *hubs skeleton*. They are able to show that in contrast to the hub vectors it is possible to compute and store the partial vectors efficiently. These partial vectors together with the hubs skeleton are enough to compute all hub vectors and hence (by transitivity) the final personalized PageRank. Essentially the idea is to reduce the computations to the set of hubs, which is much smaller than the Web graph (but $K \geq 10^4$ is possible). Note that the larger K may be chosen, the better the Q-matrix is represented.

The online step then consists of determining a personalization vector $\boldsymbol{v}^\sigma = \sum_{k=1}^{K} \alpha_k^\sigma \boldsymbol{e}^{i_k}$ and the corresponding PageRank vector

$$\boldsymbol{c}_{\mathrm{PR}}^{\sigma} = \sum_{k=1}^{K} \boldsymbol{c}_{\mathrm{PR}}^{\alpha_k^\sigma \boldsymbol{e}^{i_k}}$$

(again by using partial vectors and the hubs skeleton).

BlockRank. This approach of Kamvar et al. [339] was originally invented for accelerating the computation of PageRank, see Section 4.3.2. It consists of a 3-phase-algorithm where the main idea is to decompose the Web graph according to hosts. But, as already proposed by the authors, this approach may also be applied to find personalized PageRank scores: In the second step of the algorithm the host-weights have to be introduced, hence the algorithm is the following:

1. (offline) Choose K blocks (hosts) and let v_i^k be the local PageRank of page i in block k, $i = 1, \ldots, n,\ k = 1, \ldots, K$. Compute $\hat{Q}_{\cdot k} = \boldsymbol{c}_{\mathrm{PR}}^{\boldsymbol{v}^k}$ (the authors claim that $K \geq 10^3$ is possible if the Web structure is exploited).
2. (online) For query σ find appropriate host-weights to combine the hosts.
3. (online) Apply the (standard) PageRank algorithm to compute the associated centralities. Use as input the local PageRank scores computed in the first step, weighted by the host-weights of step 2.

Both, the concept of personalization from this section and normalization from the previous section will be rediscussed in the following two sections to introduce the four dimensions of centrality indices.

5.3 Four Dimensions of a Centrality Index

In this section we present a four dimension approach which is an attempt to structure the wide field of different centrality measures and related personalization and normalization methods presented so far. The idea to this model emerged from the observation that there is currently no consistent axiomatic schema that captures all the centrality measures considered in Chapter 3, for more details see Section 5.4. But it is important to note, that the following contribution does not constitute a formal approach or even claims completeness. Nevertheless, we believe that it may be a helpful tool in praxis.

The analysis of the centrality measures in Chapter 3 has led to the idea of dividing the centralities into four categories according to their fundamental computation model. Each computation model is represented by a so-called basic term. Given a basic term, a term operator (e.g. the sum or the maximum), and several personalization and normalization methods may be applied to it. In the following we want to discuss the idea in more detail. At the end of this section we provide a scheme based on our perception that helps to classify new centrality measures, or helps to customize existing ones.

Basic Term. The classification of the centrality measures into four categories and the representation of each category by a basic term constitutes the first dimension of our approach. Once again, we want to mention that this classification is only a proposal which emerged form the analysis of existing measures described so far.

Reachability. The first category of centrality measures is based on the notion of 'reachability'. A vertex is supposed to be central if it reaches many other vertices. Centrality measures of this category are the degree centrality (cf. Section 3.3.1), the centrality based on eccentricity and closeness (cf. Section 3.3.2), and the random walk closeness centrality (cf. Section 3.8.3). All of these centralities rely on the distance concept $d(u, v)$ of two vertices u and v. In the degree centrality, for example, we count the number of vertices that can be reached within distance 1. The closeness of a vertex u is measured by the reciprocal of the sum over the distances to all other vertices v. The same is true for the centrality based on eccentricity, where the maximum is taken instead of the total distance. In the case of the random walk closeness centrality the notion of distance is equivalently given as the mean first passage time from vertex u to all other vertices v in a random walk.

Amount of flow. The second category of centrality measures is based on the amount of flow $f_{st}(x)$ from a vertex s to a vertex t that goes through a vertex or an edge x. This can be easily seen at centrality measures based on current flow processes (cf. Section 3.7) and random walks as described in Section 3.8.1 and 3.8.2. But also measures based on the enumeration of shortest paths belong to this category. The stress centrality presented in Section 3.4.1 may also be

interpreted as measuring the amount of flow going through an element x if every vertex s sends to every other vertex t one unit flow along each shortest path connecting them. In the same context, the shortest-path betweenness centrality introduced in Section 3.4.2 measures the expected fraction of times a unit flow goes through the element if every vertex s sends one unit flow consecutively to every other vertex t, and each time choosing one of all shortest paths connecting them uniformly, independently at random. The basic term covering these measures is $f_{st}(x)$.

Vitality. A third category of centrality measures is based on the vitality as defined in Section 3.6. Here, the centrality value of an element x is defined as the difference of a real-valued function f on G with and without the element. Recall, a general vitality measure was denoted by $\mathcal{V}(G, x) = f(G) - f(G \backslash \{x\})$. The maximum flow betweenness vitality presented in Sect. 3.6.1 belongs to this category.

Feedback. A fourth category of centrality measures is based on a implicit definition of a centrality (cf. Section 3.9). These measures might be subsumed by the abstract formula $c(v_i) = f(c(v_1), \ldots, c(v_n))$, where the centrality value of a certain vertex v_i depends on the centrality values of all vertices $v_1, \ldots, v_n$.

Term Operator. The second dimension is represented by the term operator. Consider the first three categories: here we observed that often a set of suitable operators can be applied to a basic term to obtain meaningful centrality measures. We want to illustrate this idea on some centrality measures: If we have carefully defined the distance for a given application, we can choose whether the centrality index is given by the maximum of all distances from u to any other vertex v (as in the eccentricity), or the sum over all distances (as in the closeness centrality), or the average distance to all other vertices (as a normalized closeness centrality). In some cases even a special operator as the variance of all the distance might led a meaningful centrality index. Thus, for all centrality indices of the first three categories, it makes sense to separate the choice of a term operator from the basic term.

Personalization. The third dimension is given by the methods that help to personalize centrality measures. In Section 5.2 we differentiate two variants of personalization. The first approach, denoted by $P_{\boldsymbol{v}}$, is applicable to all centrality measure that can deal with vertex or edge weights. This personalization applies a weight vector $\boldsymbol{v}$ to V, E, or to the transition matrix of the random surfer model in the case of the Web centralities. The second personalization method, denoted by P_R, considers a subset of vertices, the so called rootset R. The centrality of a vertex is measured with respect to this rootset. This method is applicable to all distance based centrality indices. Both personalization methods and all other approaches to personalization build the third dimension.

Normalization. All of the centrality measures presented in this book can be normalized. Thus, the normalization forms a fourth dimension. Recall, a common normalization applicable to most centrality measures is to divide every value by the maximum centrality value. In Section 5.1 several normalization methods were considered.

Independence of the Dimensions. All of these four dimensions: basic term, term operator, personalization, and normalization are independent of each other and we have outlined that the centrality measures presented in this book can be meaningfully dissected into them. Of course, we cannot claim that all centrality indices ever published will fall into one of these categories or can be dissected as demonstrated. Moreover, since we lack any strict definition of centrality indices, we cannot ensure that every possible combinations will result in meaningful centrality index. Our aim is to provide a model that helps to structure the design of a suitable centrality index according to our four-dimensional approach.

Designing a Centrality Index. The diagram in Figure 5.1 shows an approach that demonstrates how an appropriate centrality can be found or adapted for a given application. The first step in choosing an appropriate centrality index is to find the question that should be answered by the centrality measure. That determines the category and the corresponding basic term. In general, however, the basic term refers only to an abstract concept. The distance between two vertices, for example, could be measured by the mean first passage time in a random walk or by the classic definition of distance on shortest paths. Thus a concrete computational model must be developed for the chosen basic term. After this step, a first personalization might be applied. This personalization leads to a personalized graph with modified or added weights on the vertices or edges, respectively. Afterwards, a second personalization might be applicable by choosing a 'rootset' if the basic term corresponds to one of the categories *reachability, amount of flow or vitality.* The centrality of a vertex is then measured with respect to this rootset. If the resulting term belongs to the first three categories, 'reachability', 'amount of flow', or 'vitality', we have to chose a term operator which will be applied to the term with respect to the personalized graph. We want to mention here as examples the maximum-operator or the summation over all terms.

If the chosen centrality index is a feedback centrality a personalization with a rootset is not always applicable. Thus, the route through the diagram follows a special path for these indices. The next step here is to determine the appropriate linear equation system and to solve it.

In all four categories the resulting centrality values might be normalized, as discussed in Section 5.1. Usually this normalization is performed by a multiplication with a scalar.

As a tool for describing, structuring, and developing centrality measures our four dimension approach provides a flexible alternative to classical approaches even though more formalization and refinement is needed. In the next section

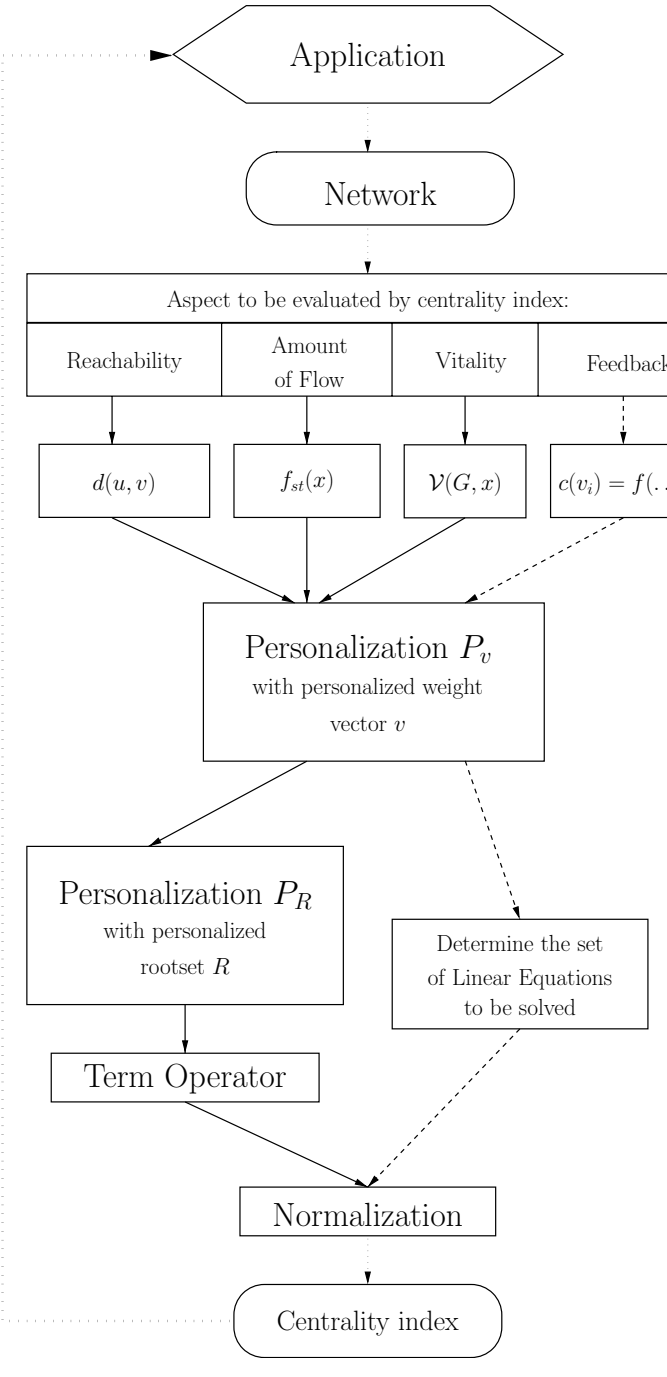


Fig. 5.1. A flow chart for choosing, adapting or designing an appropriate centrality measure for a given application

we consider several classical approaches which may also be used to characterize centrality measures.

5.4 Axiomatization

In Chapter 3, we saw that there are many different centrality indices fitting for many different applications. This section discusses the question whether there exist general properties a centrality should have.

We will first cover two axiomatizations of distance-based approaches of centrality indices and in a second subsection discuss two aximatisations for feedback centralities.

5.4.1 Axiomatization for Distance-Based Vertex Centralities

In the fundamental paper of Sabidussi [500], several axioms are defined for a vertex centrality of an undirected connected graph $G = (V, E)$. In the following we restate these in a slightly modified way. Sabidussi studied two operations on graphs:

Adding an edge (u, v)*:* Let u and v be distinct vertices of G where $(u, v) \notin E(G)$. The graph $H = (V, E \cup \{(u, v)\})$ is obtained from G by adding the edge (u, v).

Moving an edge (u, v)*:* Let u, v, w be three distinct vertices of G such that $(u, v) \in E(G)$ and $(u, w) \notin E(G)$. The graph $H = (V, (E \setminus \{(u, v)\}) \cup \{(u, w)\})$ is obtained by removing (u, v) and inserting (u, w). Moving an edge must be admissible, i.e., the resulting graph must still be connected.

Let $\mathcal{G}_n$ be the class of connected undirected graphs with n vertices. Furthermore, let $c\colon V \to \mathbb{R}_0^+$ be a function on the vertex set V of a graph $G = (V, E) \in \mathcal{G}_n$ which assigns a non-negative real value to each vertex $v \in V$. Recall, in Section 3.3.3 we denoted by $\mathcal{S}_c(G) = \{u \in V\colon \forall v \in V\ c(u) \geq c(v)\}$ the set of vertices of G of maximum centrality with respect to a given vertex centrality c.

Definition 5.4.1 (Vertex Centrality (Sabidussi [500])). *A function c is called a* vertex centrality on $G \in \mathcal{G}'_n \subseteq \mathcal{G}_n$*, and $\mathcal{G}'_n$ is called* c-admissible*, if and only if the following conditions are satisfied:*

1. *$\mathcal{G}'_n$ is closed under isomorphism, i.e., if $G \in \mathcal{G}'_n$ and H is isomorphic to G then also $H \in \mathcal{G}'_n$.*
2. *If $G = (V, E) \in \mathcal{G}'_n$, $u \in V(G)$, and H is obtained from G by moving an edge to u or by adding an edge to u, then $H \in \mathcal{G}'_n$, i.e., $\mathcal{G}'_n$ is closed under moving and adding an edge.*
3. *Let $G \simeq_\phi H$, then $c_G(u) = c_H(\phi(u))$ for each $u \in V(G)$.*[2]

[2] By $c_G(u)$ and $c_H(u)$ we denote the centrality value of vertex u in G and H, respectively.

4. *Let $u \in V(G)$, and H be obtained from G by adding an edge to u, then $c_G(u) < c_H(u)$ and $c_G(v) \leq c_H(v)$ for each $v \in V$.*
5. *Let $u \in \mathcal{S}_c(G)$, and H be obtained from G either by moving an edge to u or by adding an edge to u, then $c_G(u) < c_H(u)$ and $u \in \mathcal{S}_c(H)$.*

The first two conditions provide a foundation for Condition 3 and 5. Note that certain classes of graphs fail to satisfy Condition 2, e.g., the class of all trees is closed under moving of edges, but not under addition of edges. Condition 3 describes the invariance under isomorphisms, also claimed in Definition 3.2.1. The idea behind Condition 4 is that adding an edge increases the degree of centralization of a network. Condition 5 is the most significant one. If an edge is moved or added to a vertex $u \in \mathcal{S}_c(G)$, then the centrality of u should be increased and it should contained in $\mathcal{S}_c(H)$, i.e., u must have maximal centrality in the new graph H.

For the degree centrality introduced in Section 3.3.1, it is easy to verify that the axioms are satisfied. Thus, the degree centrality is a vertex centrality in terms of Sabidussi's definition.

We shall now see that the vertex centrality $c_E(u)$ based on the eccentricity $e(u)$ introduced in Section 3.1 is not a vertex centrality according to Sabidussi's definition. In Figure 5.2 two graphs are shown where the eccentricity value for each vertex is indicated. The first graph is a simple path with one central vertex u_5. After adding the edge (u_5, u_9) the new central vertex is u_4. Thus, adding an edge according to Condition 5 does not preserve the center of a graph. Note, also Condition 4 is violated.

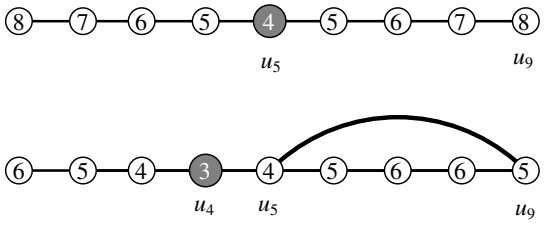


Fig. 5.2. The eccentricity $e(u)$ for each vertex $u \in V$ is shown. The example illustrates that the eccentricity centrality given by $c_E(u) = e(u)^{-1}$ is not a vertex centrality according to Sabidussi's definition (see Definition 5.4.1)

In Section 3.2 the closeness centrality of a vertex was defined by $c_C(u) = s(u)^{-1}$. Kishi [357] showed that this centrality is not a vertex centrality respecting Sabidussi's definition. An example is given in Figure 5.3, where the value of the total distance for each vertex is indicated. The median $\mathcal{M}(G) = \{u \in V : s(G) = s(u)\}$ of the left graph G consists of the vertices u, u', and u''. The insertion of edge (u, v) yields a graph H with $\mathcal{M}(H) \cap \mathcal{M}(G) = \emptyset$.

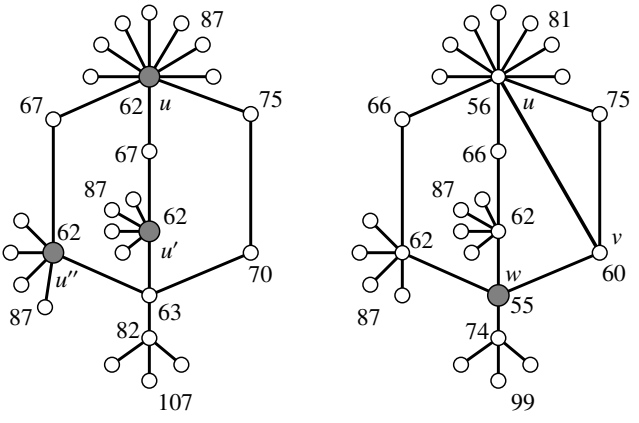


Fig. 5.3. The total distance $s(u)$ for each vertex $u \in V$ is shown. The example depicts that the closeness centrality defined by $c_C(u) = s(u)^{-1}$ is not a vertex centrality according to Sabidussi's definition (see Definition5.4.1)

Kishi [357] provides a definition for distance-based vertex centralities relying on Sabidussi's definition. Let c be a real valued function on the vertices of a connected undirected graph $G = (V, E)$, and let u and v be two distinct non-adjacent vertices of G. The insertion of (u, v) leads to a graph $H = (V, E \cup \{(u, v)\})$ where the difference of the centrality values is measured by $\Delta_{uv}(w) = c_H(w) - c_G(w)$ for each vertex $w \in G$.

Definition 5.4.2 (Vertex Centrality (Kishi [357])). *The function c is called a* vertex centrality *if and only if the following conditions are satisfied*

1. $\Delta_{uv}(u) > 0$, *i.e.,* $c_G(u) < c_H(u)$.
2. *For each $w \in V$ with $d(u, w) \leq d(v, w)$ it holds that $\Delta_{uv}(u) \geq \Delta_{uv}(w)$ for any pair of non-adjacent vertices u and v.*

The conditions of Definition 5.4.2 are quite similar to Condition 4 and 5 of Sabidussi's definition 5.4.1. Therefore, it is not surprising that the eccentricity is not a vertex centrality according to Kishi's definition. To see that, reconsider Figure 5.2 where vertex u_5 violates the Condition 2 of Kishi's definition. However, Kishi [357] showed that the closeness centrality is a vertex centrality with respect to Definition 5.4.2.

As these two examples show, it will still be a challenge to find minimal requirements which can be satisfied by a distance-based centrality index. In Section 3.2 we claimed that the centrality index only depends on the structure of the graph (cf. Def. 3.2.1). But as mentioned already, not every structural index will be accepted as a centrality index.

Finally, we want to note that there are also attempts to define requirements for a vertex centrality of an weakly connected directed graphs, see e.g. Nieminen [451].

5.4.2 Axiomatization for Feedback-Centralities

Up to now, we mainly discussed sets of axioms that defined and admitted centralities that are based either on shortest path distances or on the degree of the vertex. This section reviews axiomatizations that lead to feedback centralities or feedback-like centralities.

Far from being complete we want to give two examples of how an axiomatization could work. To our knowledge there are several approaches concerning axiomatization, but up to now there is a lack of structure and generality: Many properties a centrality should have are proposed in the literature, but those sets of properties in most cases depend very much on the application the authors have in mind and exclude known and well-established centralities.

We start with a paper by van den Brink and Gilles [563], which may serve as a bridge between degree-based and feedback-like centralities. This is continued by presenting results of Volij and his co-workers that axiomatically characterize special feedback-centralities.

From Degree to Feedback. In [563], van den Brink and Gilles consider directed graphs. In the main part of their paper the graphs are unweighted, but the axiomatic results are generalized to the weighted case. We only review the results for the unweighted case - the weighted case is strongly related but much more complicated with respect to notation.

The goal is to find an axiomatic characterization of centralities, or, to be more specific, of what they call *relational power measures* which assign to each directed network with n vertices an n-dimensional vector of reals such that the ith component of the vector is a measure of the relational power (or dominance) of vertex i.

The first measure is the *β-measure*, that was developed by the same authors [562] for hierarchical economic organizations. It measures the potential influence of agents on trade processes.

Let $\mathcal{G}_n$ be the set of unweighted directed graphs having n vertices. For a directed edge $(i, j) \in E$, vertex i is said to *dominate* vertex j.

Definition 5.4.3. *Given a set of vertices V with $|V| = n$, the* β-measure *on V is the function $\beta : \mathcal{G}_n \to \mathbb{R}^n$ given by*

$$\beta_G(i) = \sum_{j \in N_G^+(i)} \frac{1}{d_G^-(j)} \qquad \forall\ i \in V,\ G \in \mathcal{G}_n$$

(Remember that $d^-(j)$ is the in-degree of vertex j and $N^+(i)$ is the set of vertices j for which a directed edge (i, j) exists.)

This β-measure may be viewed as a feedback centrality, since the score for vertex i depends on properties of the vertices in its forward neighborhood.

A set of four axioms uniquely determines the β-measure. To state the four axioms, let $f : \mathcal{G}_n \to \mathbb{R}^n$ be a relational power measure on V. Moreover, we need the following definition:

Definition 5.4.4. *A* partition *of $G \in \mathcal{G}_n$ is a subset $\{G_1, \ldots, G_K\} \subseteq \mathcal{G}_n$ such that*

– $\bigcup_{k=1}^{K} E_k = E$ *and*
– $E_k \cap E_l = \emptyset \; \forall \; 1 \leq k, l \leq K, \; k \neq l$.

The partition is called independent *if in addition*

$$| \left\{ k \in \{1, \ldots, K\} : d^-_{G_k}(i) > 0 \right\} | \leq 1 \; \forall \; i \in V,$$

i.e., if each vertex is dominated in at most one directed graph of the partition.

Which properties should a centrality have in order to measure the relational power or dominance of a vertex?

First of all it would be good to normalize the measure in order to compare dominance values of different vertices - possibly in different networks. Due to the domination structure of their approach van den Brink and Gilles propose to take the number of dominated vertices as the total value that is distributed over the vertices according to their relational power:

Axiom 1: *Dominance normalization* For every $G \in \mathcal{G}_n$ it holds that

$$\sum_{i \in V_G} f_G(i) = | \left\{ j \in V_G : d^-_G(j) > 0 \right\} |.$$

The second axiom simply says that a vertex that does not dominate any other vertex has no relational power and hence gets the value zero:

Axiom 2: *Dummy vertex property* For every $G \in \mathcal{G}_n$ and $i \in V$ satisfying $N^+_G(i) = \emptyset$ it holds that $f_G(i) = 0$.

In the third axiom the authors formalize the fact that if two vertices have the same dominance structure, i.e. the same number of dominated vertices and the same number of dominating vertices, then they should get the same dominance-value:

Axiom 3: *Symmetry* For every $G \in \mathcal{G}_n$ and $i, j \in V$ satisfying $d^+_G(i) = d^+_G(j)$ and $d^-_G(i) = d^-_G(j)$ it holds that $f_G(i) = f_G(j)$.

Finally, the fourth axiom addresses the case of putting together directed graphs. It says that if several directed graphs are combined in such a way that a vertex is dominated in at most one directed graph (i.e. if the result of the combination may be viewed as an independent partition), then the total dominance value of a vertex should simply be the sum of its dominance values in the directed graphs.

Axiom 4: *Additivity over independent partitions* For every $G \in \mathcal{G}_n$ and every independent partition $\{G_1, \ldots, G_K\}$ of G it holds

$$\boldsymbol{f}_G = \sum_{k=1}^{K} \boldsymbol{f}_{G_k}.$$

Interestingly, these axioms are linked to the preceding sections on degree-based centralities: If the normalization axiom is changed in a specific way, then the unique centrality score that satisfies the set of axioms is the out-degree centrality. The authors call this *score-measure*. Note that an analogous result also holds for the weighted case.

In more detail, after substituting the dominance normalization by the score normalization (see Axiom 1b below), the following function is the unique relational power measure that satisfies Axioms 2 – 4 and 1b:

$$\sigma_G(i) = d_G^+(i) \; \forall \; i \in V, \; G \in \mathcal{G}_n$$

Instead of taking the number of dominated vertices as the total value that is distributed over the vertices according to their dominance, the total number of relations is now taken as a basis for normalization:

Axiom 1b: *Score normalization* For every $G \in \mathcal{G}_n$ it holds that

$$\sum_{i \in V} f_G(i) = |E|.$$

Above, we presented a set of axioms that describe a certain measure that has some aspects of feedback centralities but also links to the preceding section via its strong relation to the score measure. We now pass over to feedback centralities in the narrower sense.

Feedback Centralities. In terms of citation networks, Palacios-Huerta and Volij [460] proposed a set of axioms for which a centrality with normalized influence proposed by Pinski and Narin [479] is the unique centrality that satisfies all of them. This Pinski-Narin-centrality is strongly related to the PageRank score in that it may be seen as the basis (of PageRank) that is augmented by the addition of a stochastic vector that allows for leaving the sinks.

To state the axioms properly we need some definitions. We are given a directed graph $G = (V, E)$ with weights ω on the edges and weights α on the vertices. In terms of citation networks V corresponds to the set of journals and $(i, j) \in E$ iff journal i is cited by journal j. The weight $\omega(i, j)$ is defined to be the number of citations to journal i by journal j if $(i, j) \in E$ and 0 otherwise, while the vertex weight $\alpha(i)$ corresponds to the number of articles published in journal i. The authors consider strongly connected subgraphs with the additional property that there is no path from a vertex outside the subgraph to a vertex contained in it. (Note that they allow loops and loop weights.) Palacios-Huerta and Volij call such subgraphs a *discipline*, where a discipline is a special *communication class* (a strongly connected subgraph) which itself is defined to be an equivalence class with respect to the equivalence relation of communication. Two journals i and j *communicate*, if either $i = j$ or if i and j *impact* each other, where i impacts j if there is a sequence of journals $i = i_0, i_1, \ldots, i_{K-1}, i_K = j$ such that i_{l-1} is cited by i_l, that is, if there is a path from i to j.

Define the $(|V| \times |V|)$-matrices

$$W = (\omega(i,j)),\ D_\omega = \mathrm{diag}(\omega(\cdot j)) \text{ with } \omega(\cdot j) = \sum_{i \in V} \omega(i,j),$$

and set WD_ω^{-1} to be the normalized weight matrix, and $D_\alpha = \mathrm{diag}\,(\alpha(i))$. Then the *ranking problem* $\langle V, \alpha, W\rangle$ is defined for the vertex set V of a discipline, the associated vertices weights α and the corresponding citation matrix W, and considers the ranking (a centrality vector) $\boldsymbol{c}_{\mathrm{PHV}} \geq 0$ that is normalized with respect to the l_1-norm: $\|\boldsymbol{c}_{\mathrm{PHV}}\|_1 = 1$.

The authors consider two special classes of ranking problems:

1. ranking problems with all vertex weights equal, $\alpha(i) = \alpha(j)\ \forall\ i,j \in V$ (*isoarticle problems*) and
2. ranking problems with all *reference intensities* equal, $\frac{\omega(\cdot i)}{\alpha(i)} = \frac{\omega(\cdot j)}{\alpha(j)}\ \forall\ i,j \in V$ (*homogeneous problems*).

To relate small and large problems, the *reduced ranking problem* R^k for a ranking problem $R = \langle V, \alpha, W\rangle$ with respect to a given vertex k is defined as $R^k = \langle V \setminus \{k\}, (\alpha(i))_{i \in V\setminus\{k\}}, (\omega_k(i,j))_{(i,j)\in V\setminus\{k\}\times V\setminus\{k\}}\rangle$, with

$$\omega_k(i,j) = \omega(i,j) + \omega(k,j)\frac{\omega(i,k)}{\sum_{l\in V\setminus\{k\}}\omega(l,k)}\ \forall\ i,j \in V \setminus \{k\}.$$

Finally, consider the problem of splitting a vertex j of a ranking problem $R = \langle V, \alpha, W\rangle$ into $|T_j|$ sets of identical vertices (j,t_j) for $t_j \in T_j$. For $V' = \{(j,t_j) : j \in V, t_j \in T_j\}$, the *ranking problem resulting from splitting* j is denoted by

$$R' = \langle V', (\alpha'((j,t_j)))_{j\in J, t_j\in T_j}, (\omega'((i,t_i)(j,t_j)))_{((i,t_i)(j,t_j))\in V'\times V'}\rangle,$$

with

$$\alpha'((j,t_j)) = \frac{\alpha(j)}{|T_j|},\ \omega((i,t_i)(j,t_j)) = \frac{\omega(i,j)}{|T_i||T_j|}.$$

Note that the latter two definitions of special ranking problems are needed to formulate the following axioms.

A *ranking method* Φ assigning to each ranking problem a centrality vector should then satisfy the following four axioms (at least the weak formulations):

Axiom 1: *invariance with respect to reference intensity*

Φ is *invariant with respect to reference intensity* if

$$\Phi(\langle V, \alpha, W\Gamma\rangle) = \Phi(\langle V\alpha, W\rangle)$$

for all ranking problems $\langle V, \alpha, W\rangle$ and every Matrix $\Gamma = \mathrm{diag}(\gamma_j)_{j\in V}$ with $\gamma_j > 0\ \forall\ j \in V$.

Axiom 2: *(weak) homogeneity*

a) Φ satisfies *weak homogeneity* if for all two-journal problems $R = \langle \{i, j\}, \alpha, W \rangle$ that are homogeneous and isoarticle, it holds that

$$\frac{\Phi_i(R)}{\Phi_j(R)} = \frac{\omega(i,j)}{\omega(j,i)}. \tag{5.7}$$

b) Φ satisfies *homogeneity* if (Equation 5.7) holds for all homogeneous problems.

Axiom 3: *(weak) consistency*

a) Φ satisfies *weak consistency* if for all homogeneous, isoarticle problems $R = \langle V, \alpha, W \rangle$ with $|V| > 2$ and for all $k \in V$

$$\frac{\Phi_i(R)}{\Phi_j(R)} = \frac{\Phi_i(R^k)}{\Phi_j(R^k)} \; \forall \; i, j \in V \setminus \{k\}. \tag{5.8}$$

b) Φ satisfies *consistency* if (Equation 5.8) holds for all homogeneous problems.

Axiom 4: *invariance with respect to the splitting of journals*

Φ is *invariant to splitting of journals*, i.e. for all ranking problems R and for all splittings R' of R holds

$$\frac{\Phi_i(R)}{\Phi_j(R)} = \frac{\Phi_{(i,t_i)}(R')}{\Phi_{(j,t_j)}(R')} \; \forall i, j \in V, \; \forall \; i \in T_i, \; \forall \; j \in T_j.$$

Palacios-Huerta and Volij show that the ranking method assigning the Pinski-Narin centrality $\boldsymbol{c}_{PN}$ given as the unique solution of

$$D_\alpha^{-1} W D_W^{-1} D_\alpha \boldsymbol{c} = \boldsymbol{c}$$

is the only ranking method that satisfies

- invariance to reference intensity (Axiom 1),
- weak homogeneity (Axiom 2a),
- weak consistency (Axiom 3a), and
- invariance to splitting of journals (Axiom 4).

Slutzki and Volij [526] also consider the axiomatization of ranking problems, which they call *scoring problems*. Although their main field of application is shifted from citation networks to (generalized) tournaments, it essentially considers the same definitions as above, excluding the vertex weights α. Further, they consider strongly connected subgraphs (not necessarily disciplines), and set $\omega(i, i) = 0$ for all $i \in V$, meaning that there are no self-references, i.e. no loops in the corresponding graph. For this case, the Pinski-Narin centrality may be characterized by an alternative set of axioms, and again it is the only centrality satisfying this set.

The Link to Normalization. Above, we saw that normalization is a question when dealing with axiomatizations. Either it is explicitly stated as an axiom (see the centralities of van den Brink and Gilles) or the normalization is implicitly assumed when talking about centralities (see the papers of Volij and his coworkers). The topic of normalization was already investigated in Section 5.1. Here, we report on investigations of Ruhnau [499] about normalizing centralities.

Her idea is based on an intuitive understanding of centrality, already formulated by Freeman in 1979 [227]:

"A person located in the center of a star is universally assumed to be structurally more central than any other person in any other position in any other network of similar size."

She formalizes this in the definition of a vertex-centrality for undirected connected graphs $G = (V, E)$.

Definition 5.4.5 (Ruhnau's vertex centrality axioms). *Let $G = (V, E)$ be an undirected and connected graph with $|V| = n$ and let $c_V : V \to \mathbb{R}$. c_V is called a* vertex-centrality *if*

1. *$c_V(i) \in [0, 1]$ for all $i \in V$ and*
2. *$c_V(i) = 1$ if and only if G is a star with n vertices and i the central vertex of it.*

(Note: Ruhnau calls this a node centrality*. For consistency with the rest of the chapter we used the equivalent term vertex centrality here.)*

The property of being a vertex-centrality may be very useful when comparing vertices of different graphs. To see this, compare the central vertex of a star of order n with any vertex in a complete graph of order n. Both have a degree of $n - 1$, but intuitively the central vertex of a star has a much more prominent role in the graph than any of the vertices in a complete graph.

Freeman [226] showed that the betweenness centrality satisfies the conditions of the above definition. Due to the fact that the eigenvector centrality normalized by the Euclidean norm has the property that the maximal attainable value is $1/\sqrt{2}$ (independent of n), and that it is attained exactly at the center of a star (see [465]), it is also a vertex-centrality (multiplied by $\sqrt{2}$). For more information about normalization, see Section 5.1.

5.5 Stability and Sensitivity

Assume that a network is modified slightly for example due to the addition of a new link or the inclusion of a new page in case of the Web graph. In this situation the 'stability' of the results are of interest: does the modification invalidate the computed centralities completely?

In the following subsection we will discuss the topic of stability for distance based centralities, i.e., eccentricity and closeness, introduce the concept of stable,

quasi-stable and unstable graphs and give some conditions for the existence of stable, quasi-stable and unstable graphs.

A second subsection will cover Web centralities and present results for the numerical stability and rank stability of the centralities discussed in Section 3.9.3.

5.5.1 Stability of Distance-Based Centralities

In Section5.4.1 we considered the axiomatization of connected undirected graphs $G = (V, E)$ and presented two definitions for distance-based vertex centralities. Moreover, we denoted by $\mathcal{S}_c(G) = \{u \in V : \forall v \in V\ c(u) \geq c(v)\}$ the set of maximum centrality vertices of G with respect to a centrality c and we studied the change of the centrality values if we add an edge (u, v) between two distinct non-adjacent vertices in G. In this section we focus on the stability of the center $\mathcal{S}_c(G)$ with respect to this graph operation (cf. Condition 5 of Definition 5.4.1).

Let $u \in \mathcal{S}_c(G)$ be a central vertex with respect to c, and $(u, v) \notin G$. Then the insertion of an edge (u, v) to G yields a graph $H = (V, E \cup (u, v))$. Regarding $\mathcal{S}_c(H)$ two cases can occur, either

$$\mathcal{S}_c(H) \subseteq \mathcal{S}_c(G) \cup \{v\} \tag{5.9}$$

or

$$\mathcal{S}_c(H) \not\subseteq \mathcal{S}_c(G) \cup \{v\} \tag{5.10}$$

for every vertex $v \in V$. Kishi [357] calls a graph for which the second case (Equation 5.10) occurs an *unstable graph* with respect to c. Figures 5.2 and 5.3 in Section 5.4.1 show unstable graphs with respect to the eccentricity and the closeness centrality. The first case (Equation 5.9) can be further classified into

$$\mathcal{S}_c(H) \subseteq \mathcal{S}_c(G) \quad \text{and} \quad u \in \mathcal{S}_c(H) \tag{5.11}$$

and

$$\mathcal{S}_c(H) \not\subseteq \mathcal{S}_c(G) \quad \text{or} \quad u \notin \mathcal{S}_c(H) \tag{5.12}$$

A graph G is called a *stable graph* if the first case (Equation 5.11) occurs, otherwise G is called a *quasi-stable graph*. The definition of stable graphs with respect to c encourages Sabidussi's claim [500] that an edge added to a central vertex $u \in \mathcal{S}_c(G)$ should strengthen its position.

In Figure 5.4 an example for a quasi-stable graph with respect to closeness centrality is shown. For each vertex the status value $s(u) = \sum_{v \in V} d(u, v)$ is indicated. Adding the edge (u, v) leads to a graph with a new central vertex v.

In [357] a more generalized form of closeness centrality is presented by Kishi: The centrality value $c_{GenC}(u)$ of a vertex $u \in V$ is

$$c_{GenC}(u) = \frac{1}{\sum_{k=1}^{\infty} a_k n_k(u)} \tag{5.13}$$

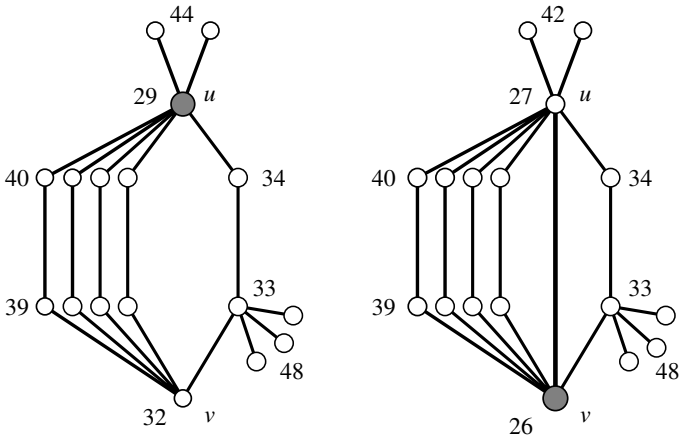


Fig. 5.4. A quasi-stable graph with respect to the closeness centrality. The values indicate the total distances $s(u)$. After inserting the edge (u, v) the new median is vertex v

where $n_k(u)$ is the number of vertices whose distance from u is k and each a_k is a real constant. With $a_k = k$ it is easy to see that

$$\frac{1}{\sum_{k=1}^{\infty} a_k n_k(u)} = \frac{1}{\sum_{v \in V} d(u, v)} = c_C(u).$$

Kishi and Takeuchi [358] have shown under which conditions there exists a stable, quasi-stable, and unstable graph for generalized centrality functions c_{GenC} of the form in Equation 5.13:

Theorem 5.5.1. *For any generalized vertex centrality c_{GenC} of the form in Equation 5.13 holds:*

1. *if $a_2 < a_3$, then there exists a quasi-stable graph, and*
2. *if $a_3 < a_4$, then there exists an unstable graph.*

Theorem 5.5.2. *Any connected undirected graph G is stable if and only if the generalized vertex centrality c_{GenC} given in Equation 5.13 satisfies $a_2 = a_3$. Moreover, G is not unstable if and only if c_{GenC} satisfies $a_3 = a_4$.*

Sabidussi has shown in [500] that the class of undirected trees are stable graphs with respect to the closeness centrality.

Theorem 5.5.3. *If an undirected graph G forms a tree, then G is stable with respect to the closeness centrality.*

5.5.2 Stability and Sensitivity of Web-Centralities

First, we consider stability with respect to the centrality *values*, later on we report on investigations on the centrality *rank*. We call the former *numerical stability* and the latter rank stability.

Numerical Stability. Langville and Meyer [378] remark that it is not reasonable to consider the linear system formulation of, e.g., the PageRank approach and the associated condition number[3], since it may be that the solution vector of the linear system changes considerable but the normalized solution vector stays almost the same. Hence, what we are looking for is to consider the stability of the eigenvector problem which is the basis for different Web centralities mentioned in Section 3.9.3.

Ng et al. [449] give a nice example showing that an eigenvector may vary considerably even if the underlying network changes only slightly. They considered a set of Web pages where 100 of them are linked to *algore.com* and the other 103 pages link to *georgewbush.com*. The first two eigenvectors (or, in more detail, the projection onto their nonzero components) are drawn in Figure 5.5(a). How the scene changes if five new Web pages linking to both *algore.com* and *georgewbush.com* enter the collection is then depicted in Figure 5.5(b).

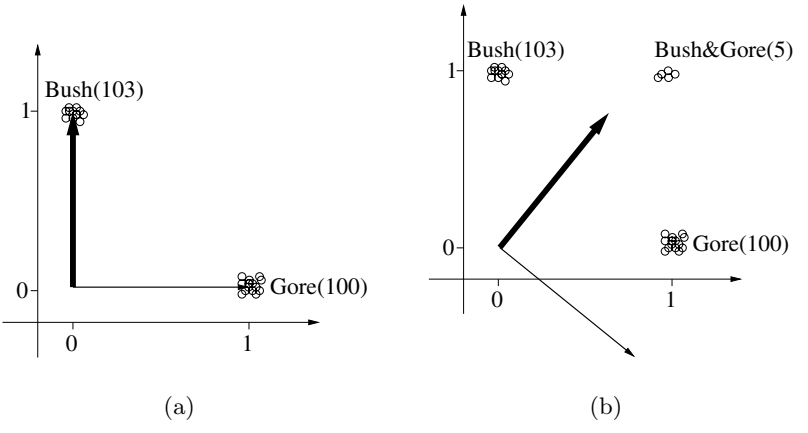
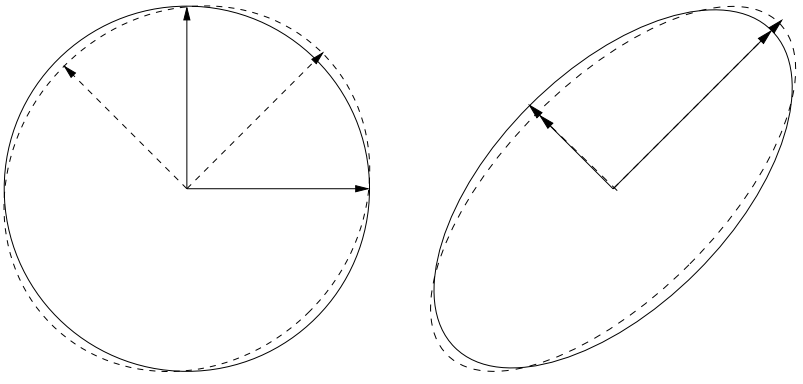


Fig. 5.5. A small example showing instability resulting from perturbations of the graph. The projection of the eigenvector is shown and the perturbation is visible as a strong shift of the eigenvector

Regarding the Hubs & Authorities approach Ng et al. the authors give a second example, cf. Figs 5.6(a) and 5.6(b). In the Hubs & Authorities algorithm the largest eigenvector for $S = A^T A$ is computed. The solid lines in the figures represent the contours of the quadratic form $x^T S_i x$ for two matrices S_1, S_2 as well as the contours of the slightly (but equally) perturbed matrices. In both figures the associated eigenvectors are depicted. The difference (strong shift in the eigenvectors in the first case, almost no change in the eigenvectors in the

[3] $cond(A) = \|A\|\|A^{-1}\|$ (for A regular)

second case) between the two figures consists of the fact that S_1 has an eigengap[4] $\delta_1 \sim 0$ whereas S_2 has eigengap $\delta_2 = 2$. Hence in the case that the eigengap is almost zero, the algorithm may be very sensitive about small changes in the matrix whereas in case the eigengap is large the sensitivity is small.



(a) (b)

Fig. 5.6. A simple example showing the instability resulting from different eigengaps. The position of the eigenvectors changes dramatically in the case of a small eigengap (a)

Ng et al. also show this behavior theoretically

Theorem 5.5.4. *Given $S = A^T A$, let $\boldsymbol{c}_{HA\text{-}A}$ be the principal eigenvector and δ the eigengap of S. Assume $d^+(i) \leq d$ for every $i \in V$ and let $\varepsilon > 0$. If the Web graph is perturbed by adding or deleting at most k links from one page, $k < \left(\sqrt{d+\alpha} - \sqrt{d}\right)^2$, $\alpha = \frac{\varepsilon\delta}{4+\sqrt{2}\varepsilon}$ then the perturbed principal eigenvector $\tilde{\boldsymbol{c}}_{HA\text{-}A}$ of the perturbed matrix $\tilde{S}$ satisfies $\|\boldsymbol{c}_{HA\text{-}A} - \tilde{\boldsymbol{c}}_{HA\text{-}A}\|_2 \leq \varepsilon$.*

Theorem 5.5.5. *If S is a symmetric matrix with eigengap δ, then there exists a perturbed version $\tilde{S}$ of S with $\|S - \tilde{S}\|_F = \mathcal{O}(\delta)$ that causes a large $(\Omega(1))$ change in the principal eigenvector.*
(Note that $\|X\|_F = \left(\sum_i \sum_j (x_{ij}^2)\right)^{1/2}$ denotes the Frobenius norm.)

If we consider the PageRank algorithm, then the first fact that we have to note is that for a Markov chain having transition matrix P the sensitivity of the principal eigenvector is determined by the difference of the second eigenvalue to 1. As shown by Haveliwala and Kamvar [290] the second eigenvalue for the PageRank-matrix with P having at least two irreducible closed subsets satisfies

[4] Difference between the first and the second largest eigenvalue.

$\lambda_2 = d$. This is true even in the case that in Formula 3.43 the vector $\mathbf{1}_n$ is substituted by any stochastic vector $\boldsymbol{v}$, the so-called *personalization vector*, cf. Section 5.2 for more information about the personalization vector.

Therefore a damping factor of $d = 0.85$ (this is the value proposed by the founders of Google) yields in general much more stable results than $d = 0.99$ which would be desirable if the similarity of the original Web graph with its perturbed graph should be as large as possible.

Ng et al. [449] proved

Theorem 5.5.6. *Let $U \subseteq V$ be the set of pages where the outlinks are changed, $\boldsymbol{c}_{PR}$ be the old PageRank score and $\boldsymbol{c}_{PR}^U$ be the new PageRank score corresponding to the perturbed situation. Then*

$$\|\boldsymbol{c}_{PR} - \boldsymbol{c}_{PR}^U\|_1 \leq \frac{2}{1-d} \sum_{i \in U} c_{PR}(i).$$

Bianchini, Gori and Scarselli [61] were able to strengthen this bound. They showed

Theorem 5.5.7. *Under the same conditions as given in Theorem 5.5.6 it holds*

$$\|\boldsymbol{c}_{PR} - \boldsymbol{c}_{PR}^U\|_1 \leq \frac{2d}{1-d} \sum_{i \in U} c_{PR}(i).$$

(Note that $d < 1$.)

Rank Stability. When considering Web centralities, the results are in general returned as a list of Web pages matching the search-query. The scores attained by the Web pages are in most cases not displayed and hence the questions that occurs is whether numeric stability also implies stability with respect to the rank in the list (called *rank-stability*). Lempel and Moran [388] investigated the three main representatives of Web centrality approaches with respect to rank-stability.

To show that numeric stability does not necessarily imply rank-stability they used the graph $G = (V, E)$ depicted in Figure 5.7. Note that in the graph any undirected edge $[u, v]$ represents two directed edges (u, v) and (v, u). From G two different graphs $G_a = (V, E \cup \{(y, h_a)\})$ and $G_b = (V, E \cup \{(y, h_b)\})$ are derived (they are not displayed). It is clear that the PageRank vector $\boldsymbol{c}_{\mathrm{PR}}^a$ corresponding to G_a satisfies

$$0 < c_{\mathrm{PR}}^a(x_a) = c_{\mathrm{PR}}^a(y) = c_{\mathrm{PR}}^a(x_b),$$

and therefore $c_{\mathrm{PR}}^a(h_a) > c_{\mathrm{PR}}^a(h_b)$.

Analogously in G_b we have

$$0 < c_{\mathrm{PR}}^b(x_a) = c_{\mathrm{PR}}^b(y) = c_{\mathrm{PR}}^b(x_b),$$

hence $c_{\mathrm{PR}}^b(h_a) < c_{\mathrm{PR}}^b(h_b)$.

Concluding we see that by shifting one single outlink from a very low-ranking vertex y induces a complete change in the ranking:

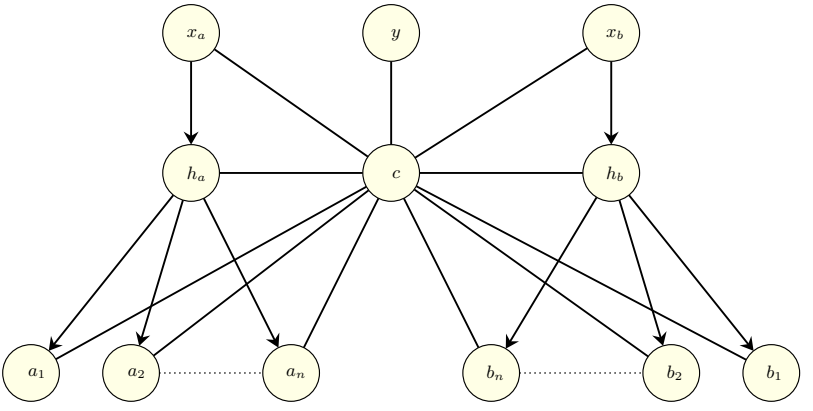


Fig. 5.7. The graph G used for the explanation of the rank stability effect of PageRank. Please note that for G_a a directed edge from y to h_a is added and in the case of G_b from y to h_b

$$c^a_{\mathrm{PR}}(a_i) > c^a_{\mathrm{PR}}(b_i) \text{ and } c^b_{\mathrm{PR}}(a_i) < c^b_{\mathrm{PR}}(b_i) \; \forall \; i.$$

To decide whether an algorithm is rank-stable or not we have to define the term *rank-stability* precisely. Here we follow the lines of Borodin et al. [87] and [388]. Let $\mathcal{G}$ be a set of directed graphs and $\mathcal{G}_n$ the subset of $\mathcal{G}$ where all directed graphs have n vertices.

Definition 5.5.8. *1. Given two ranking vectors $\boldsymbol{r}^1$ and $\boldsymbol{r}^2$, associated to a vertex-set of order n, the ranking-distance between them is defined by*

$$d_r(\boldsymbol{r}^1, \boldsymbol{r}^2) = \frac{1}{n^2} \sum_{i,j=1}^{n} l_{ij}^{\boldsymbol{r}^1, \boldsymbol{r}^2}$$

$$\text{where } l_{ij}^{\boldsymbol{r}^1, \boldsymbol{r}^2} = \begin{cases} 1, & r_i^1 < r_j^1 \text{ and } r_i^2 > r_j^2 \\ 0, & \text{otherwise} \end{cases}$$

2. An algorithm $\mathcal{A}$ is called rank-stable *on $\mathcal{G}$ if for each k fixed we have*

$$\lim_{n \to \infty} \max_{\substack{G_1, G_2 \in \mathcal{G}_n \\ d_e(G_1, G_2) \leq k}} d_r\left(\mathcal{A}(G_1), \mathcal{A}(G_2)\right) \longrightarrow 0,$$

where $d_e(G_1, G_2) = |(E_1 \cup E_2) \setminus (E_1 \cap E_2)|$.

Hence an algorithm $\mathcal{A}$ is rank-stable on $\mathcal{G}$ if for each k the effect on the ranking of the nodes of changing k edges vanishes if the size of the node-set of a graph tends to infinity.

Borodin et al. were able to show that neither the Hubs & Authorities algorithm nor the SALSA method are rank-stable on the set of all directed graphs $\bar{\mathcal{G}}$.

However, they obtained a positive result by considering a special subset of $\bar{\mathcal{G}}$, the set of authority connected directed graphs $\mathcal{G}^{ac}$:

Definition 5.5.9. *1. Two vertices $p, q \in V$ are called* co-cited*, if there is a vertex $r \in V$ satisfying $(r, p), (r, q) \in E$.*
2. p, q are connected by a co-citation path *if there exist vertices $p = v_0, v_1, \ldots, v_{k-1}, v_k = q$ such that (v_{i-1}, v_i) are co-cited for all $i = 1, \ldots, k$.*
3. A directed graph $G = (V, E)$ is authority connected *if for all p, q satisfying $d^-(p), d^-(q) > 0$ there is a co-citation path.*

Lempel and Moran argue that it is reasonable to restrict the stability investigations to this subset of directed graphs due to the following observation:

- if p, q are co-cited then they cover the same subject,
- the relevance of p and q should be measured with respect to the same bar, and
- there is no interest in answering questions like "is p a better geography resource that q is an authority on sports?"

For authority connected subgraphs it holds that

- SALSA is rank-stable on $\mathcal{G}^{ac}$,
- The Hubs & Authorities algorithm is not rank-stable on $\mathcal{G}^{ac}$, and
- PageRank is not rank-stable on $\mathcal{G}^{ac}$.

Note that the latter two results were obtained by Lempel and Moran [388].

With this result we finish the discussion of sensitivity and stability of Web centralities. Interested readers are directed to the original papers shortly mentioned in this section.