3 Centrality Indices

Dirk Koschützki, Katharina Anna Lehmann,* Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl,* and Oliver Zlotowski**

Centrality indices are to quantify an intuitive feeling that in most networks some vertices or edges are more central than others. Many vertex centrality indices were introduced for the first time in the 1950s: e.g., the Bavelas index [50, 51], degree centrality [483] or a first feedback centrality, introduced by Seeley [510]. These early centralities raised a rush of research in which manifold applications were found. However, not every centrality index was suitable to every application, so with time, dozens of new centrality indices were published. This chapter will present some of the more influential, 'classic' centrality indices. We do not strive for completeness, but hope to give a catalog of basic centrality indices with some of their main applications.

In Section 3.1 we will begin with two simple examples to show how centrality indices can help to analyze networks and the situation these networks represent. In Section 3.2 we discuss the properties that are minimally required for a real-valued function on the set of vertices or edges of a graph to be a centrality index for vertices and edges, respectively.

In subsequent Sections 3.3–3.9, various families of vertex and edge centralities are presented. First, centrality indices based on distance and neighborhood are discussed in Section 3.3. Additionally, this section presents in detail some instances of facility location problems as a possible application for centrality indices. Next we discuss the centrality indices based on shortest paths in Section 3.4. These are naturally defined for both, vertices and edges. We decided to present both, vertex and edge centrality indices, in one chapter together since many families of centrality indices are naturally defined for both and many indices can be easily transformed from a vertex centrality to an edge centrality, and vice versa. Up to date there have been proposed many more centrality indices for vertices than for edges. Therefore, we discuss general methods to derive an edge centrality out of the definition of a vertex centrality in Section 3.5. The general approach of vitality measures is also applicable to edges and vertices. We will describe this family in Section 3.6. In Section 3.7, a family of centrality indices is presented that is derived from a certain analogy between information flow and current flow. In Section 3.8 centrality indices based on random processes are presented. In Section 3.9 we present some of the more prominent feedback centralities that evaluate the importance of a vertex by evaluating the centrality of its surrounding vertices.

* Lead authors

U. Brandes and T. Erlebach (Eds.): Network Analysis, LNCS 3418, pp. 16–61, 2005.

For many centrality indices it is required that the network at hand be connected. If this is not the case, computing these centralities might be a problem. As an example, shortest paths based centralities encounter the problem that certain vertices are not reachable from vertices in a different component of the network. This yields infinite distances for closeness centrality, and zero shortest-path counts for betweenness centrality. Section 3.10 of this chapter discusses how to deal with these problems in disconnected graphs.

Before we close the chapter we want to discuss a topic that spans the bridge between the analysis of networks on the level of elements and the level of the whole graph. In Section 3.11, we propose a very general method with which a structural index for vertices can be transformed into a structural index for graphs. This is helpful, e.g., in the design of new centrality indices which will be explained on a simple example. We close this chapter with some remarks on the history of centrality indices in Section 3.12.

3.1 Introductory Examples

Election of a leader is a frequent event in many social groups and intuitively, some persons in such an event are more important or 'central' than others, e.g. the candidates. The question is now how centrality indices can help to derive a measure of this intuitive observation. On this first example we want to illustrate that different kind of networks can be abstracted from such a social interaction and we want to show how network analysis with centrality indices may help to identify important vertices of these networks. A second example illustrates how the application of an edge centrality index may help to figure out important edges in a network. Both illustrations underline that there is no centrality index that fits all applications and that the same network may be meaningfully analyzed with different centrality indices depending on the question to be answered.

Before we begin the discussion on the examples, 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? In the course of time there have been different answers to this question. Each of them serves another intuition about the notion of centrality. Centrality can be interpreted as - among other things - 'influence', as 'prestige' or as 'control'. 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.

We will demonstrate the application of three different interpretations on the following example: A school class of 30 students has to elect a class representative and every student is allowed to vote for one other student. We can derive different graph abstractions from this situation that can later be analyzed with different centrality indices. We will first look at a network that represents the voting results directly. In this network vertices represent students and an edge from student A to student B is established if A has voted for B. In such a situation

a student could be said to be the more 'central' the more people have voted for him or her. This kind of centrality is directly represented by the number of edges pointing to the corresponding vertex. The so called 'in-degree centrality' is presented in Section 3.3.1.

Another view on the same situation results in another network: In this network an edge between A and B represents that student A has convinced student B to vote for his or her favorite candidate. We will call this network an 'influence network'. Let us assume that the class is mainly split into two big groups X and Y. Let some person have social relationships to members from both groups. If this person has a favorite candidate from group X and convinces a big part of group Y to vote for this candidate, he or she is 'central' because he or she mediates the most information between both groups. With this argument we can say that a vertex in the given influence network is the more central the more it is needed to transport the opinion of others. A family of centrality indices that tries to capture this intuition of 'being between groups' is the family of betweenness centrality indices, presented in Sections 3.4.2, 3.6.1 and 3.8.2.

In yet another perspective we could view the general social network of the class: Who is friends with whom? Someone who is a friend of an important person could be regarded as more important than someone having friends with low social prestige. The centrality of a vertex in this kind of network is therefore given by the centrality of adjacent vertices. This kind of 'feedback centrality' is captured by many centrality indices that are presented in Section 3.9.

In analogy to the centrality of vertices, some of the edges in a network can be viewed as being more important than others. We will illustrate this on a commonly used network, the Internet. Looking at the backbone of the Internet it is clear that the cables between servers on different continents are few and thus very important for the functionality of the system. This importance stems from the enormous data flow through the intercontinental cables that had to be redirected if one of these cables was out of service. There are mainly two different approaches to measure the centrality of an edge in a network: The first counts the number of substructures like traversal sets or the set of shortest paths in the graph on which an edge participates. An example for this approach is the betweenness centrality of edges, presented in Section 3.4.2. The second approach is based on the idea of measuring how much a certain network parameter is changed if the edge is removed. An example for this approach is the flow betweenness vitality, presented in Section 3.6.1.

We have shown for two examples that very different ideas of centrality can lead to centrality indices that help to analyze the situation represented by the given network. It is important to note that none of these measures is superior to the others. Every one is appropriate for some but not all questions in network analysis.

3.2 A Loose Definition

Before presenting any centrality indices, we first have to give a definition for centrality indices.[1] Historically there is no commonly accepted definition of what a centrality index is, and almost everybody introduced his or her centrality without giving a strict definition for centrality in general. Thus, here we will just state the least common ground for all centralities presented in the following sections. In Section 5.4 we will give some classes of centralities that follow much stricter definitions.

The intuition about a centrality is that it denotes an order of importance on the vertices or edges of a graph by assigning real values to them. As we have pointed out in the introduction to this chapter, the notion of 'importance' is by no means unambiguous. Nonetheless, as a minimal requirement we demand that the result of a centrality index is only depending on the structure of the graph. This demand is stated in the following definition of a structural index. Every of the centrality indices presented here is a structural index but it is important to note that not every structural index will be accepted as a centrality index. Section 5.4 will also show that to date there is no stricter definition that captures all of the introduced centrality indices.

Recall, that two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are isomorphic ($G_1 \simeq G_2$) if there exists a one-to-one mapping $\phi\colon V_1 \to V_2$ such that (u, v) is an edge in E_1 if and only if $(\phi(u), \phi(v))$ is an edge in E_2 (cf. Section 2.3).

Definition 3.2.1 (Structural Index). *Let $G = (V, E)$ be a weighted, directed or undirected multigraph and let X represent the set of vertices or edges of G, respectively. A real-valued function s is called a structural index if and only if the following condition is satisfied: $\forall x \in X\colon G \simeq H \implies s_G(x) = s_H(\phi(x))$, where $s_G(x)$ denotes the value of $s(x)$ in G.*

A centrality index c is required to be a structural index and thus induces at least a semi-order on the set of vertices or edges, respectively. By this order we can say that $x \in X$ is at least as central as $y \in X$ with respect to a given centrality c if $c(x) \geq c(y)$. Note that, in general, the difference or ratio of two centrality values cannot be interpreted as a quantification of how much more central one element is than the other.

The definition of a structural index expresses the natural requirement that a centrality measure must be invariant under isomorphisms. In particular, this condition implies that a centrality measure is also invariant under automorphisms.

3.3 Distances and Neighborhoods

In this section we will present centrality indices that evaluate the 'reachability' of a vertex. Given any network these measures rank the vertices according to the

[1] Centrality index will be used synonymously with centrality measure and, shortly, centrality.

number of neighbors or to the cost it takes to reach all other vertices from it. These centralities are directly based on the notion of distances within a graph, or on the notion of neighborhood, as in the case of the degree centrality. We start with this very basic index, the degree centrality. Other centralities, like eccentricity or closeness, will be presented in the light of a special application, the facility location problem.

3.3.1 Degree

The most simple centrality is the degree centrality $c_D(v)$ of a vertex v that is simply defined as the degree $d(v)$ of v if the considered graph is undirected. In directed networks two variants of the degree centrality may be appropriate: the in-degree centrality $c_{iD}(v) = d^-(v)$ and the out-degree centrality $c_{oD}(v) = d^+(v)$. The degree centrality is, e.g., applicable whenever the graph represents something like a voting result. These networks represent a static situation and we are interested in the vertex that has the most direct votes or that can reach most other vertices directly. The degree centrality is a local measure, because the centrality value of a vertex is only determined by the number of its neighbors. In the next section we investigate global centrality measures and consider their applications in a special set of problems, namely Facility Location Problems.

3.3.2 Facility Location Problems

Facility location analysis deals with the problem of finding optimal locations for one or more facilities in a given environment. Location problems are classical optimization problems with many applications in industry and economy. The spatial location of facilities often take place in the context of a given transportation, communication, or transmission system, which may be represented as a network for analytic purposes.

A first paradigm for location based on the minimization of transportation costs was introduced by Weber [575] in 1909. However, a significant progress was not made before 1960 when facility location emerged as a research field.

There exist several ways to classify location problems. According to Hakami [271] who considered two families of location problems we categorize them with respect to their objective function. 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. The second family of location problems considered by Hakimi 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 described for example in [524, 527] 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.

Our focus here is not to treat all facility location problems. The interested reader is referred to a bibliography devoted to facility location analysis [158]. The aim of this section is to introduce three important vertex centralities by examining location problems. In the subsequent section we investigate some structural properties of the sets of most central indices that are given by these centrality indices.

The definition of different objectives leads to different centrality measures. A common feature, however, is that each objective function depends on the distance between the vertices of a graph. In the following we assume that $G = (V, E)$ is a connected undirected graph with at least two vertices and we suppose that the distance $d(u, v)$ between two vertices u and v is defined as the length of the shortest path from u to v (cf. in Section 2.2.2). These assumptions ensure that the following centrality indices are well defined. Moreover, for reasons of simplicity we consider G to be an unweighted graph, i.e., all edge weights are equal to one. Of course, all indices presented here can equally well be applied to weighted graphs.

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 $e(u) = \max\{d(u, v) \colon v \in V\}$. The problem of finding an optimal location can be solved by determining the minimum over all $e(u)$ with $u \in V$. In graph theory, the set of vertices with minimal eccentricity is denoted as the center of G (cf. Section 3.3.3). The concept is illustrated in Figure 3.1. The eccentricity values are shown and the most central vertices are highlighted.

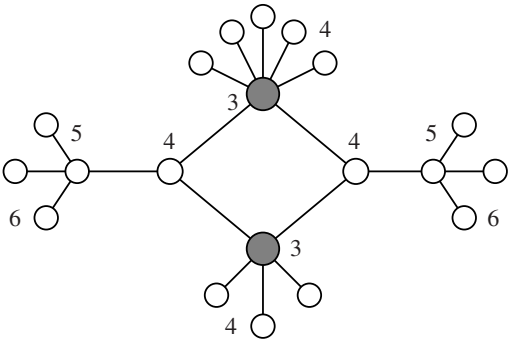


Fig. 3.1. Eccentricity values of a graph. Vertices in the center are colored in grey

Hage and Harary [278] proposed a centrality measure based on the eccentricity

$$c_E(u) = \frac{1}{e(u)} = \frac{1}{\max\{d(u,v)\colon v \in V\}}. \tag{3.1}$$

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 $v \in V$.

Closeness. Next 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[2] $\sum_{v \in V} d(u, v)$. The problem of finding an appropriate location can be solved by computing the set of vertices with minimum total distance. In Figure 3.2 the total distances for all vertices are shown and the vertices with minimal total distance are highlighted.

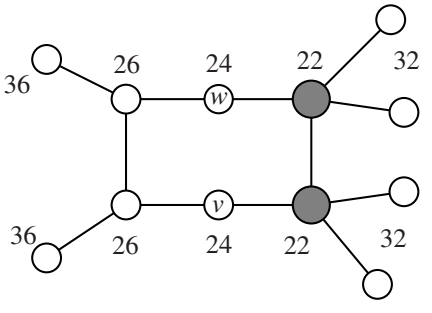


Fig. 3.2. Total distances of a graph. Lowest valued vertices are colored in grey. Note, the vertices v and w are more important with respect to the eccentricity

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. Various closeness-based measures have been developed, see for example [500, 51, 52, 433, 558, 451, 88]. In Section 3.10 we outline a measures developed for digraphs. The most commonly employed definition of closeness is the reciprocal of the total distance

[2] In [273], Harary used the term status to describe a status of a person in an organization or a group. In the context of communication networks this sum is also called transmission number.

$$c_C(u) = \frac{1}{\sum_{v \in V} d(u,v)}. \tag{3.2}$$

In our sense this definition is a vertex centrality, since $c_C(u)$ grows with decreasing total distance of u and it is clearly a structural index.

Before we discuss the competitive location problem, we want to mention the radiality measure and integration measure proposed by Valente and Foreman [558]. These measures can also be viewed as closeness-based indices. They were developed for digraphs but an undirected version is applicable to undirected connected graphs, too. This variant is defined as

$$c_R(u) = \frac{\sum_{v \in V} (\Delta_G + 1 - d(u,v))}{n-1} \tag{3.3}$$

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.

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. 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?

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 $\gamma_u(v) = |\{w \in V \colon d(u,w) < d(v,w)\}|$. If the salesman selects a vertex u and his competitor selects a vertex v, then he will have $\gamma_u(v) + \frac{1}{2}(n - \gamma_u(v) - \gamma_v(u)) = \frac{1}{2}n + \frac{1}{2}(\gamma_u(v) - \gamma_v(u))$ customers. 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

$$c_F(u) = \min\{f(u,v) \colon v \in V - u\}. \tag{3.4}$$

$c_F(u)$ 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.

In Figure 3.3 an example is shown where the centroid vertex is highlighted. Notice that for each vertex $u \in V$ in graph shown in Figure 3.4 $c_F(u) \leq -1$.

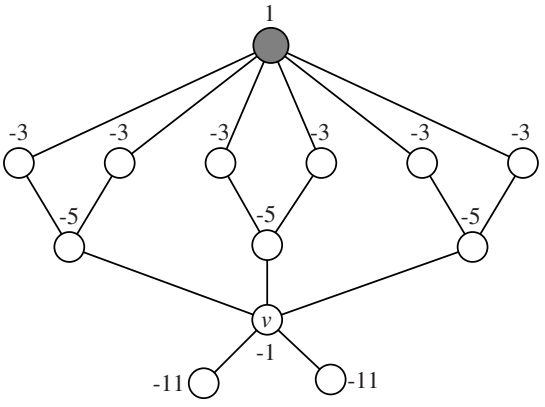


Fig. 3.3. A graph with one centroid vertex. Note that v is the vertex with maximal closeness centrality

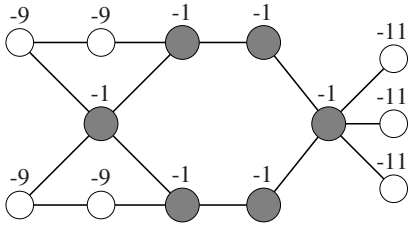


Fig. 3.4. All centroid values are negative. There is no profitable location for the salesman

Here, the salesman loses his advantage to choose as first. The strategy "choose after the leader has chosen" would be optimal.

Also the centroid measure is a structural index according to Definition 3.2.1. But in contrast to eccentricity and closeness, centroid values can be negative as well.

3.3.3 Structural Properties

In this section we will investigate several structural properties for the distance-based vertex centralities introduced in Section 3.3.2. Using Definition 3.2.1 the set of maximum centrality vertices $\mathcal{S}_c(G)$ of G with respect to a given vertex centrality c is given by

$$\mathcal{S}_c(G) = \{u \in V : \forall v \in V \; c(u) \geq c(v)\}. \tag{3.5}$$

Center of a Graph. In Section 3.3.2 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 $\mathcal{C}(G)$ of a graph G is

$$\mathcal{C}(G) = \{u \in V : r(G) = e(u)\}. \tag{3.6}$$

It is easy to show that $\mathcal{S}_{c_E}(G) = \mathcal{C}(G)$. Clearly, every undirected connected graph has a non-empty center. But where are the vertices of the center located? A basic result concerning the center of a tree is due to Jordan [336]

Theorem 3.3.1. *For any tree, the center of a tree consists of at most two adjacent vertices.*

Proof. The result is trivial if the tree consists of at most two vertices. We show that any other tree T has the same center as the tree T' which is obtained from T by removing all leaves. For each vertex u of T, only a leaf can be an eccentric vertex. Aa vertex u is an eccentric vertex of a vertex v if $d(u, v) = e(v)$. Since the eccentricity of each $u \in T'$ is one less than its eccentricity in T, T and T' have the same center. If the process of removing leaves is continued, we successively obtain trees having the same center as T. Finally, we obtain a subtree of T which consists of either one vertex or a pair of adjacent vertices. □

The proof shows that it is possible to determine the center without computing the vertex eccentricities. The following generalization of Theorem 3.3.1 due to Harary and Norman [281] deals with the location of the center in a connected separable graph, i.e., a graph which contains a cut-vertex. Recall, a cut-vertex of a graph is a vertex whose removal increases the number of components, i.e., if u is a cut-vertex of a connected graph G, then $G - u$ is disconnected. We call a graph 2-vertex-connected if G contains no cut-vertices (cf. Section 2.2.4). Note, each vertex of a graph distinct from a cut-vertex lies in exactly one 2-vertex-connected subgraph, and each cut-vertex lies in more than one.

Theorem 3.3.2. *Let G be a connected undirected graph. There exists a 2-vertex-connected subgraph in G containing all vertices of $\mathcal{C}(G)$.*

Proof. Suppose there is no 2-vertex-connected subgraph in G containing all the vertices of $\mathcal{C}(G)$. Then G has a cut-vertex u such that $G-u$ decomposes into the subgraphs G_1 and G_2 each of them containing at least one vertex of $\mathcal{C}(G)$. Let v be an eccentric vertex of u and P the corresponding shortest path between u and v of length $e(u)$. Then v must lie in G_1 or G_2, say G_2. Furthermore there exists at least one vertex w in G_1 which does not belong to P. Now, let $w \in \mathcal{C}(G)$ and let P' be a shortest path in G between w and u. Then $e(w) \geq d(w, u) + d(u, v) \geq 1 + e(u)$. So w does not belong to the center of G, a contradiction. Thus, there must be a 2-vertex-connected subgraph containing all vertices of center of G. □

Figure 3.1 in Section 3.3.2 shows a graph consisting of fourteen 2-vertex-connected subgraphs consisting of two vertices and one 2-vertex-connected subgraph in the middle containing the two central vertices.

Median of a Graph. The service facility problem presented in Sect. 3.3.2 was solved by determining the set of vertices with minimum total distance. If the minimum total distance of G is denoted by $s(G) = \min\{s(u)\colon u \in V\}$, the median $\mathcal{M}(G)$ of G is given by

$$\mathcal{M}(G) = \{u \in V\colon s(G) = s(u)\} \ . \tag{3.7}$$

Clearly $\mathcal{S}_{c_C}(G) = \mathcal{M}(G)$. Truszczyński [552] studied the location of the median in a connected undirected graph.

Theorem 3.3.3. *The median of a connected undirected graph G lies within a 2-vertex-connected subgraph of G.*

Similar to the center of a tree Theorem 3.3.3 implies the existence of at least one 2-vertex-connected subgraph containing the median of a tree.

Corollary 3.3.4. *The median of a tree consists of either a single vertex or a pair of adjacent vertices.*

The graph in Figure 3.2 contains a 2-vertex-connected subgraph of six vertices containing the median. Moreover, the example illustrates that $\mathcal{C}(G) \cap \mathcal{M}(G) = \emptyset$ is possible. Let $\langle \mathcal{M}(G) \rangle$ and $\langle \mathcal{C}(G) \rangle$ denote the subgraphs induced by $\mathcal{M}(G)$ and $\mathcal{C}(G)$, respectively. The results due to Hendry [293] and Holbert [300] show that the center and median can be arbitrarily far apart.

Theorem 3.3.5. *Let H_1 and H_2 be two connected undirected graphs. For any integer $k > 0$, there exists a connected undirected graph G, such that $\langle \mathcal{M}(G) \rangle \simeq H_1$, $\langle \mathcal{C}(G) \rangle \simeq H_2$, and the distance between $\mathcal{M}(G)$ and $\mathcal{C}(G)$ is at least k.*

This result is not surprising, because the center and the median represent solution sets of distinct objective functions.

Centroid of a Graph. The computation of the centroid of a graph is a maximin optimization problem. In Sect. 3.3.2 we have shown the relation to a competitive location problem. We defined the centroid value for a given vertex u by $c_F(u) = \min\{f(u, v)\colon v \in V - u\}$. In addition we call the objective function value $f(G) = \max\{c_F(u)\colon u \in V\}$ the centroid value of G and denote by

$$\mathcal{Z}(G) = \{u \in V\colon f(G) = c_F(u)\} \tag{3.8}$$

the set of vertices representing the centroid of G. With it the set $\mathcal{Z}(G)$ consists of all appropriate locations for the competitive location problem considered in Section 3.3.2.

We now focus on the location of the centroid in a graph. First we assume the graph is an undirected tree $T = (V, E)$. Let u be vertex of T. A branch of u is a maximal subtree containing u as a leaf. The number of branches at u is equal to the degree of u. The branch weight of u is the maximum number of edges among all branches of u. The vertex u is called a branch weight centroid vertex

if u has minimum branch weight and the branch weight centroid of T consists of all such vertices. Zenlinka [594] has shown that the branch weight centroid of T is identical with its median. Slater [524] used this result to show

Theorem 3.3.6. *For any tree the centroid and the median are identical.*

Theorem 3.3.6 and Corollary 3.3.4 together imply that the centroid of a tree consists of either a single vertex or a pair of adjacent vertices. Smart and Slater [527] also studied the relative location of the centroid in a connected undirected graph. The following Theorem is a generalization of Theorem 3.3.6.

Theorem 3.3.7. *For any connected undirected graph, the median and the centroid lie in the same 2-vertex-connected subgraph.*

Reconsider the graph in Fig. 3.3. The median and the centroid lie within the subgraph but $\mathcal{Z}(G) \cap \mathcal{M}(G) = \emptyset$. Let $\langle \mathcal{Z}(G) \rangle$ be the graph induced by $\mathcal{Z}(G)$. Smart and Slater [527] have shown the following.

Theorem 3.3.8. *Let H_1 and H_2 be to connected undirected graphs. For any integer $k \geq 4$, there exists a connected undirected graph G, such that $\langle \mathcal{Z}(G) \rangle \simeq H_1$, $\langle \mathcal{M}(G) \rangle \simeq H_2$, and the distance between $\mathcal{Z}(G)$ and $\mathcal{M}(G)$ is at least k.*

Furthermore, Smart and Slater [527] proved that the center, the median, and the centroid can be arbitrarily far apart in a connected undirected graph. In Fig. 3.5 an example is given where all sets are pairwise distinct. The following result summarizes Theorems 3.3.5 and 3.3.8.

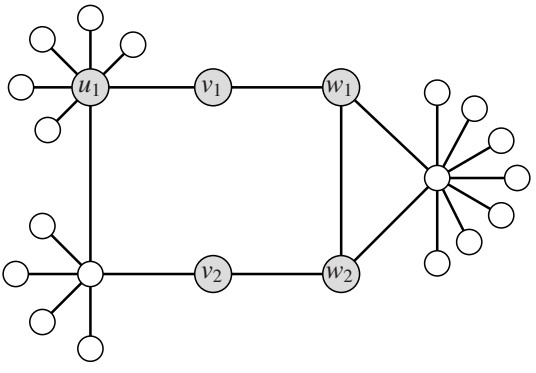


Fig. 3.5. $\mathcal{C}(G) = \{v_1, v_2\}$, $\mathcal{M}(G) = \{u_1\}$, and $\mathcal{Z}(G) = \{w_1, w_2\}$ are pairwise distinct

Theorem 3.3.9. *For three connected undirected graphs H_1, H_2, and H_3, and any integer $k \geq 4$, there exists an undirected connected graph G such that $\langle \mathcal{C}(G) \rangle \simeq H_1$, $\langle \mathcal{M}(G) \rangle \simeq H_2$, $\langle \mathcal{Z}(G) \rangle \simeq H_3$, and the distances between any two of them is at least k.*

Some of concepts presented here can be extended to digraphs. Chartrand et al. [115] showed that the result of Theorem 3.3.5 also holds for digraphs.

3.4 Shortest Paths

This section presents centrality indices that are based on the set of shortest paths in a graph. Shortest paths are defined on vertices as well as on edges and such, some centrality indices were first introduced as vertex centralities and later adapted as edge centralities. In the following, we will sometimes make a general statement regarding vertices and edges equally. We will call a vertex v or an edge e (graph) 'element' and denote the centrality of an element in general by x. The first two indices, stress and betweenness centrality of an element x, are based on the (relative) number of shortest paths that contain x. The last centrality index is only defined on edges and based on traversal sets. All three centrality indices can be defined on weighted or unweighted and directed or undirected and simple or multi graphs. For simplification we will discard any information about the underlying graph in the notation for a given centrality. Thus, c_X might denote the centrality indices of a weighted, undirected graph or any other combination of weights, direction and edge multiplicity. Note that the set of all shortest paths has to be computed in a preprocessing step with the appropriate algorithm, depending on the combination of these graph properties.

3.4.1 Stress Centrality

The first centrality index based on enumeration of shortest paths is stress centrality $c_S(x)$, introduced in [519]. The author was concerned with the question how much 'work' is done by each vertex in a communication network. It is clear that communication or transport of goods will follow different kinds of paths in a social network. Nonetheless, the author of [519] models the set of paths used for communication as the set of shortest paths. The assumption is that counting 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. With this, an element is the more central the more shortest paths run through it. The formal definition is given by:

$$c_S(v) = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \sigma_{st}(v) \tag{3.9}$$

where $\sigma_{st}(v)$ denotes the number of shortest paths containing v. The definition given in [519] is not rigorous, but in analogy to the betweenness centrality all shortest paths that either start or end in v are not accounted for this centrality index. The calculation of this centrality index is given by a variant of a simple all-pairs shortest-paths algorithm that not only calculates one shortest path but all shortest paths between any pair of vertices. More about the algorithm for this centrality can be found in Section 4.2.1.

Although this centrality was designed to measure stress on vertices, the same definition can be applied for edges:

$$c_S(e) = \sum_{s \in V} \sum_{t \in V} \sigma_{st}(e) \tag{3.10}$$

where $\sigma_{st}(e)$ denotes the number of shortest paths containing edge e. In both cases stress centrality measures the amount of communication that passes an element in an all-to-all scenario. More precisely, it is not only an all-to-all scenario but every vertex sends as many goods or information units to every other vertex as there are shortest paths between them and stress centrality measures the according stress.

We next want to show how the stress centrality value of a vertex v is related to the stress centrality indices of the edges incident to v.

Lemma 3.4.1 (Relation between $c_S(v)$ and $c_S(e)$). *In a directed graph $G = (V, E)$, stress centrality on vertices and edges are related by*

$$c_S(v) = \frac{1}{2} \sum_{e \in \Gamma(v)} c_S(e) - \sum_{v \neq s \in V} \sigma_{sv} - \sum_{v \neq t \in V} \sigma_{vt} \tag{3.11}$$

for all $v \in V$.

Proof. Consider any shortest path connecting a pair $s \neq t \in V$. It contributes a value of 1 to the stress of each of its vertices and edges. Summing the contribution of a path over all edges that are incident to a vertex v thus yields twice its contribution to v itself if $v \in V \backslash \{s, t\}$, and 1 otherwise. The sum of contributions of all shortest paths to edges incident to a common vertex v hence satisfies the above relation, since v is $\sum_{v \neq s \in V} \sigma_{sv} + \sum_{v \neq t \in V} \sigma_{vt}$ times an endvertex of any shortest path. □

3.4.2 Shortest-Path Betweenness Centrality

Shortest-path betweenness centrality can be viewed as some kind of relative stress centrality. Here, we will first define it and then discuss the motivation behind this centrality index: Let $\delta_{st}(v)$ denote the fraction of shortest paths between s and t that contain vertex v:

$$\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.12}$$

where σ_{st} denotes the number of all shortest-path between s and t. Ratios $\delta_{st}(v)$ can be interpreted as the probability that vertex v is involved into any communication between s and t. Note, that the index implicitly assumes that all communication is conducted along shortest paths. Then the betweenness centrality $c_B(v)$ of a vertex v is given by:

$$c_B(v) = \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v) \tag{3.13}$$

As for stress centrality, the shortest paths ending or starting in v are explicitly excluded. The motivation for this is that the betweenness centrality of a vertex measures the control over communication between others.

The betweenness centrality index was introduced in [32, 226] and has found a wide field of applications. In [226] this new centrality index was introduced because it is problematic to apply the closeness centrality to a disconnected graph: the distance between two vertices in different components is usually set to infinity. With this, the closeness centrality (see subsection 3.2) in disconnected graphs will give no information because each vertex is assigned the same centrality value, namely $1/\infty$. We will discuss some resorts to this problem in Section 3.10.

The betweenness centrality does not suffer from this problem: Any pair of vertices s and t without any shortest path from s to t just will add zero to the betweenness centrality of every other vertex in the network.

Betweenness centrality is similar to stress centrality introduced in [519], but instead of counting the absolute number of shortest paths, the shortest-path betweenness centrality sums up the relative number of shortest paths for each pair of endvertices. These are interpreted as the extent to which a vertex v controls the communication between such pairs. Figure 3.6 gives an example why this might be more interesting than the absolute number of shortest paths. It shows two tripartite graphs in which the middle layer mediates all communication between the upper and the lower layer. The stress centrality of vertices in the middle layer is the same in both graphs but the removal of the middle vertex on the right would disconnect the whole system whereas in the right graph the removal of a single vertex would not. This is because the former has full responsibility for the communication in its graph whereas on the left side every vertex just bears one third of it.

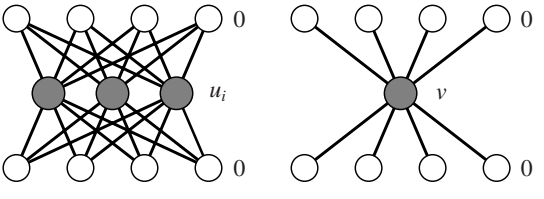


Fig. 3.6. $c_S(u_i) = 16$ and $c_B(u_i) = \frac{1}{3}$, $i = 1, 2, 3$ and $c_S(v) = 16$ but $c_B(v) = 1$. The graph shows on an example that stress centrality is not designed to evaluate how much communication control a vertex has

In [32] the shortest-path betweenness centrality – here called 'rush' – is viewed as a flow centrality: "The rush in an element is the total flow through that element, resulting from a flow between each pair of vertices". In this sense,

$\delta_{st}(v)$ is interpreted as the amount of flow that passes if one unit of flow is sent from s to t along shortest paths, and with a special division rule. In [32] the 'rush' is also defined for edges with $\delta_{st}(e)$ as the flow over edge e:

$$\delta_{st}(e) = \frac{\sigma_{st}(e)}{\sigma_{st}} \tag{3.14}$$

For reasons of consistency we will denote the resulting centrality not as 'rush on edges' but as the betweenness centrality $c_B(e)$ of edge e:

$$c_B(e) = \sum_{s \in V} \sum_{t \in V} \delta_{st}(e) \ . \tag{3.15}$$

Variants of Shortest-Path Betweenness Centrality. In [111, 580] some variants of the shortest-path betweenness centrality have been introduced. The authors generalize the approach of betweenness centrality by changing the set of paths $P(s,t)$ on which the betweenness centrality is evaluated. Instead of just using the set of all shortest paths between s and t any other set can be used for this variant. The general pattern is always the same: For each node pair s and t compute the fraction of paths in $P(s,t)$ that contain an element from the sum of all paths between s and t. To get the betweenness centrality $c_B(P(s,t))$ on a specified path set p sum over the terms for all node pairs. In [580], a number of possible path sets $P(s,t)$ was defined, as e.g. the set of k-shortest paths, i.e. the set of all paths not longer than $k \in \mathbb{N}$ or the set of k-shortest, node-disjoint paths. The according betweenness centralities did not get any special name but for reasons of consistency we will denote them as k-shortest paths and k-shortest vertex-disjoint paths betweenness centrality.

The authors in [111] were motivated by the fact that the betweenness centrality is not very stable in dynamic graphs (see also our discussion of the stability and sensitivity of centrality indices in Section 5.5). The removal or addition of an edge might cause great perturbations in the betweenness centrality values. To eliminate this, $P(s,t)$ was defined to contain all paths between a node pair s and t that are not longer than $(1+\epsilon)d(s,t)$. The resulting betweenness centrality for nodes and edges has been named ϵ-betweenness centrality. The idea behind this centrality index seems reasonable but analytical or empirical results on the stability of this index were not given.

Other variants of the general betweenness centrality concept are fundamentally different in their approach and calculation. We will discuss the flow betweenness centrality in Section 3.6.1 and the random-walk betweenness centrality in Section 3.8.2.

In the following theorem we state the relation between the edge and vertex betweenness centrality $c_B(e)$ and $c_B(v)$ of vertices and edges incident to each other:

Lemma 3.4.2 (Relation between $c_B(v)$ and $c_B(e)$). *In a directed graph $G = (V, E)$, shortest-path betweenness on vertices and edges are related by*

$$c_B(v) = \sum_{e \in \Gamma^+(v)} c_B(e) - (n-1) = \sum_{e \in \Gamma^-(v)} c_B(e) - (n-1) \tag{3.16}$$

for all $v \in V$.

Proof. Consider any shortest path connecting a pair $s \neq t \in V$. It contributes exactly $\frac{1}{\sigma_{st}}$ to the betweenness of its vertices and edges. Summing the contribution of a path over all incoming (or outgoing) edges of a vertex v thus equals its contribution to v itself if $v \in V \setminus \{s,t\}$, and $\frac{1}{\sigma_{st}}$ otherwise. The sum of contributions of all shortest paths to edges incident to a common vertex v hence satisfies the above relation, since v is $(n-1)$ times the first (last) vertex of paths to some vertex t (from some vertex s). □

3.4.3 Reach

In 2004, Ron Gutman [266] published a new approach to shortest path computation in hierarchical networks like road maps, for example. It is based on employing Dijkstras algorithm or the A* algorithm alternatively on a select subset of nodes. More specifically, only nodes having a high *reach* are considered. The concept is defined as follows:

Definition 3.4.3. *Given*

- *a directed graph* $G = (V, E)$ *with a nonnegative distance function* $m : E \to \mathbb{R}^+$, *which is called* reach metric
- *a path* P *in* G *starting at node* s *and ending at node* t
- *a node* v *on* P

the reach of v on P *is defined as* $r(v,P) := \min\{m(s,v,P), m(v,t,P)\}$, *the minimum of the distance from* s *to* v *and the distance from* v *to* t, *following path* P *according to the reach metric. The* reach of v in G, $r(v,G)$ *is the maximum value of* $r(v,Q)$ *over all least-cost paths* Q *in* G *containing* v.

When performing a Dijkstra-like shortest-path search towards a target t, nodes are only enqueued if they pass *test*(v), where *test*$(v) := r(v,G) \geq m(P) \vee r(v,G) \geq d(v,t)$. That is v is only disregarded if its reach is too small for it to lie on a least-cost path a distance $m(P)$ – denoting the length of the computed path from the origin s to v at the time v is to be inserted into the priority queue – from s and at a straight-line distance $d(v,t)$ from the destination. Note that this requires a distance function that is consistent with reach metric, such that on a path P from u to v, the path length $m(P) = m(u,v,P)$ must be at least $d(u,v)$.

At first, this reach centrality does not seem to make sense in order to simplify computation of shortest paths, since there is no obvious way of computing $r(v,G)$ for all nodes without solving an all pairs shortest path problem in the first place. However, Gutman goes on to show that in the above algorithm, even an upper bound for $r(v,G)$ suffices to preserve guaranteed shortest paths. Naturally, using an upper bound increases the number of nodes that need to be enqueued. The

author gives a sophisticated algorithm that yields practically useful bounds in a more feasible time complexity. Unfortunately, both quality and complexity are only empirically analyzed.

3.4.4 Traversal Sets

For $G = (V, E)$ and an edge $e \in E$ we call

$$T_e = \{(a, b) \in V \times V \mid \exists p.\ p \text{ is a shortest path from } a \text{ to } b \text{ and contains } e\}$$

the edge's *traversal set* – the set of source-destination pairs where for every pair some shortest path contains this edge. Now, the size of the traversal set would be an obvious measure for the importance of the edge. As claimed by Tangmunarunkit et al. [540], this simple method may not yield the desired result in some cases, so they propose the following different counting scheme.[3]

The traversal set T_e can be seen as a set of new edges, connecting those pairs of vertices that have shortest paths along e in the original graph. These edges (together with the vertices they connect) naturally constitute a graph, which is bipartite as we will now see.

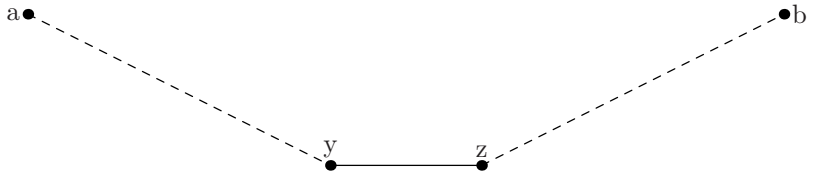


Fig. 3.7. The traversal set graph is bipartite

Let (a, b) be any edge in the traversal set graph T_e of edge $e = (y, z)$. This means that there is a shortest path p connecting a and b via e (cf. Figure 3.7). Without loss of generality, assume that p has the form $a - \cdots - y - z - \cdots - b$. Then, there cannot be an $a - z$ path shorter than the $a - y$ prefix of p, for else the resulting path along $a - \cdots - z - \cdots - b$ would be shorter than our shortest path p. In the other direction, no $y - b$ path may be shorter than our $z - b$ suffix of p. To summarize, a is closer to y, and b is closer to z. Let $\mathcal{Y}$ denote the set of all vertices closer to y than to z and let $\mathcal{Z}$ denote the set of all vertices closer to z. Thus, $\mathcal{Y}$ and $\mathcal{Z}$ form a partition of V. No two vertices belonging to the same set can be connected by an edge in this graph since the shortest path connecting them can never contain e. Thus, T_e is naturally bipartite with regard to $\mathcal{Y}$ and $\mathcal{Z}$.

[3] Both ways of counting yield values of different orders of magnitude for certain example graphs. However, we have not been able to identify a case where one scheme differentiates between two situations while the other does not. That is why we can only rely on the experience of Tangmunarunkit et al (ibid.).

An edge's *value* is then defined as the size of a minimum vertex cover on the bipartite graph formed by the traversal set:

$$C_{ts}(e) = \min\{|H| \mid H \text{ is a vertex cover for } T_e\}$$

Unlike the non-bipartite case, this is computable in polynomial time (less than $\Theta(n^3)$) using a theorem by Kőnig and Egerváry [366, 173], which states that the minimum size of a vertex cover equals the size of a maximum matching on bipartite graphs.

In [540] this centrality index has been used to characterize a graph with regard to its hierarchical organization. The authors determine the edge value pattern of sample paths in the original graph. If a high fraction of paths shows an up-down pattern of edge values, i.e., a paths begins with edges having a small value, the value raises along the path and then drops again to low values, the authors assume that this shows a high level of hierarchical organization of the underlying graph. An example on which this assumption is intuitively true is the graph of streets in a country: Some of them are only within cities, others are connecting smaller suburbs and some are high-speed freeways. Most paths from one location to another will follow streets that have low values at the beginning, then the driver will use a freeway and at last will use inner-city streets again at the end. This example shows that hierarchically organized networks may show an up-down pattern in the edge value distribution on many paths but the reverse will be hard to prove. This empirical finding should thus be treated with care.

3.5 Derived Edge Centralities

Historically, centrality indices were developed to analyze social networks. From this application, the emphasis lay on the analysis of the most central persons in social networks. This lead to a great number of different centrality indices for vertices. Most centrality indices for edges, e.g., the shortest path betweenness centrality, were only developed as a variant of the centrality index for vertices. Here, we want to discuss two methods with which every given centrality index for vertices can be transformed into a centrality index for edges.

3.5.1 Edge Centralities Derived from Vertex Centralities

One intuitive idea to derive an edge centrality from a vertex centrality is to apply the vertex centrality to the *edge graph* that is corresponding to the network to be analyzed:

Definition 3.5.1. *The* edge graph *of* $G = (V, E)$ *is* $G' = (E, K)$ *where* K *is the set of all edges* $e = ((x, y), (y, z))$ *where* $(x, y), (y, z) \in E$. *That is, two edges have a connection if they are adjacent to the same vertex* y *(with the first one in- and the second outbound for directed graphs).*

There are biased and unbiased centralities for vertices. Note that methods that incorporate previous knowledge usually do this by assuming that a subset

of 'root vertices' is especially important. For details on personalization see Section 5.2. Unlike the approaches described in there, an application on the edge graph then needs a description of central *edges*.

The size of the edge graph may be quadratic in the size of the original graph. For large graphs and computationally expensive methods this might well be a hindrance.

There is another caveat. Some of the more advanced techniques for vertices incorporate weighted edges, a feature that allows for more detailed models. However, in the edge graph these become weighted vertices, and there is no canonical way to use this data.

Finally, there is a philosophical point to be made against this approach: The vertex centralities described so far fall into the categories of degree, closeness and betweenness centrality. On the edge graph, these concepts translate into counting incident edges, closeness to other edges and position on paths between pairs of edges. However, when modeling phenomena using networks, we tend to have vertices representing entities, while edges describe relationships between these. Most of the time, these relationships are meaningless without the entities they connect. Therefore, none of the three mentioned categories seems to make a lot of sense as a centrality measure for edges.

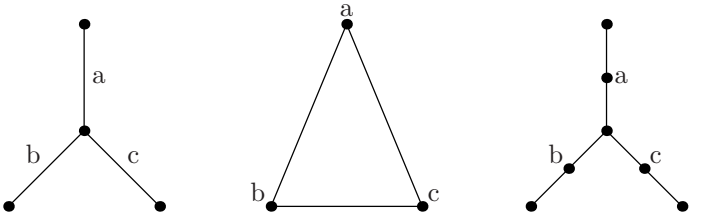


Fig. 3.8. Edge graph example

As an illustrative instance, look at the evaluation of the stress centrality on the left example graph in Figure 3.8. For a vertex x it is defined as the number of shortest paths that use x and do not end in x. The straightforward translation for an edge, say a, would be the number of shortest paths that use a, adding up to three in this example. In the middle, you find the corresponding edge graph. In contrast to the above, no shortest path (except those that end in a) crosses the vertex a. Obviously, the edge graph does not lead to the natural edge generalization of stress centrality. However, this natural generalization may be attained using a different graph translation. We will call this construction the *incidence graph*, and there is an illustrative instance on the right hand side of Figure 3.8: Each edge e is split by a new 'edge vertex' that receives the link's name.

Definition 3.5.2. *The* incidence graph *of* $G = (V, E)$ *is*

$$G'' = (V \cup E, \{(v, e) \mid \exists w : e = (v, w) \in E\} \cup \{(e, w) \mid \exists v : e = (v, w) \in E\}.$$

That is, a 'real vertex' and an 'edge vertex' become linked if they are incident in the original graph.

We can now use a biased version of stress vertex betweenness (see Section 5.2 for details on how to personalize measures), which only takes into account 'real vertex' pairs to measure the importance of 'edge vertices'. This way, most vertex measures may be translated into edge measures. As with the original centralities, it remains to check if the measure we achieve does have a sensible semantics with respect to the function of the network.

3.6 Vitality

Vitality measures are commonly used to determine the importance of vertices or edges in a graph. Given an arbitrary real-valued function on G a vitality measure quantifies the difference between the value on G with and without the vertex or edge. The main motivation behind this idea is that most networks have some quality that can be evaluated by a function on G: Imagine a transport network with different capacities on the edges in which the goal is to transport as much as possible of some good from some vertex s to another vertex t. The functionality of a network for this goal can be described by the maximal possible flow in it (see Section 2.2.3). The degree to which this quality is impaired by the loss of an edge or vertex can be viewed as the extent to which this edge or vertex is 'central' for the network. A second example is a graph representing a mobile communication network in which every vertex should be indirectly connected to all others over as few switching points as possible. The quality of this graph could be evaluated by its Wiener index, the sum over all distances in the graph (see Section 3.6.2). Then, the vitality of a vertex or edge x denotes the loss of this quality if x was removed from the network. More formally:

Definition 3.6.1 (Vitality Index). *Let $\mathcal{G}$ be the set of all simple, undirected and unweighted graphs $G = (V, E)$ and $f : \mathcal{G} \to \mathbb{R}$ be any real-valued function on $G \in \mathcal{G}$. A vitality index $\mathcal{V}(G, x)$ is then defined as the difference of the values of f on G and on G without element x: $\mathcal{V}(G, x) = f(G) - f(G \backslash \{x\})$.*

We will begin with a centrality index that is derived from the field of network flow problems. After that, a new centrality index, the closeness vitality, is presented that might be useful for some applications. The next subsection presents a new centrality index that is not a vitality index in the strict sense but the relationship to vitality indices is strong. The last subsection presents a discussion in how far the stress centrality presented in Section 3.4.1 can be interpreted as a vitality index.

3.6.1 Flow Betweenness Vitality

In this subsection we present a vertex centrality based on network flows. More precisely a measure for max-flow networks is presented which is similar to the

shortest-path betweenness described in Section 3.4.2 and makes the measure proposed in Freeman et al. [229] concrete.[4] As Stephenson and Zelen [533] observed, there is no reason to believe that information in a communication network between a pair of vertices takes place only on the shortest path. Obviously, there are applications where the centrality values computed by shortest path betweenness leads to misleading results. Thus other paths have to be considered instead.

Taking up the example of communication networks, Freeman et al. assumed information as flow and assigned with each edge a non-negative value representing the maximum of information that can be passed between its endpoints. In extending the betweenness model to flow networks, a vertex u will be seen as standing between other vertices. The goal is to measure the degree that the maximum flow between those vertices depends on u.

Based on this idea we provide a concise definition of a vertex centrality based on maximum flows. We call this centrality the max-flow betweenness vitality. Note that the maximum-flow problem between a source vertex s and a target vertex t was introduced in Section 2.2.3. For reasons of simplicity we further assume $G = (V, E)$ as a connected undirected network with non-negative edge capacities. By f_{st} we denote the objective function value of a maximum s-t-flow. The value f_{st} represents the maximal flow between s and t in G with respect to the capacity constraints and the balance conditions. As indicated above, we are now interested in the answer of the questions: How much flow must go over a vertex u in order to obtain the maximum flow value? And how does the objective function value change if we remove u from the network?

According to the betweenness centrality for shortest paths we define the max-flow betweenness for a vertex $u \in V$ by

$$c_{mf}(u) = \sum_{\substack{s,t \in V \\ u \neq s, u \neq t \\ f_{st} > 0}} \frac{f_{st}(u)}{f_{st}} \tag{3.17}$$

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 $\tilde{f}_{st}(u)$ is the maximal s-t-flow in $G \setminus u$. That is, $\tilde{f}_{st}(u)$ is determined by removing u form G and computing the maximal s-t-flow in the resulting network $G \setminus u$.

It is important to note, that this concept may also be applied to other network flow problems, e.g., the minimum-cost maximum-flow problem (MCMF) which may be viewed as a generalization of the max-flow problem. In a MCMF network each edge has a non-negative cost value and a non-negative upper capacity bound. The objective is to find a maximum flow of minimum cost between two designated vertices s and t. Applying the idea of measuring the vitality of each vertex to MCMF networks yields a new meaningful vitality measure. For further details relating to the MCMF problem see [6].

[4] Note that the original definition in [229] is ambiguous, because it neglects that a max-flow is not unique in general.

3.6.2 Closeness Vitality

In analogy to the closeness centrality index presented in Section 3.3.2, we will introduce a new centrality, based on the Wiener Index[5] [583]. The Wiener Index $I_W(G)$ of a graph G is defined as the sum over the distances of all vertex pairs:

$$I_W(G) = \sum_{v \in V} \sum_{w \in V} d(v, w) \tag{3.18}$$

It is easy to see that the Wiener Index can also be written as the sum of the closeness centrality values $c_C(v)$ (see Section 3.2) of all vertices v:

$$I_W(G) = \sum_{v \in V} \frac{1}{c_C(v)} \tag{3.19}$$

We will now define a new centrality called closeness vitality $c_{CV}(x)$, defined on both vertices and edges:

$$c_{CV}(x) = I_W(G) - I_W(G \setminus \{x\}) \tag{3.20}$$

Clearly, this new centrality is a vitality, with $f(G) = I_W(G)$. What does this centrality index measure? 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. With a small modification we can also calculate the average distance $d_{\varnothing}(G)$ between two vertices:

$$d_{\varnothing}(G) = \frac{I_W(G)}{n(n-1)} \tag{3.21}$$

This variant computes how much the costs are increased on average if the element x is removed from the graph.

There is one pitfall in the general idea of a closeness vitality: If x is a cut-vertex or a bridge, respectively, the graph will be disconnected after the removal. Then $c_{CV}(x)$ is $-\infty$ for this element. We will discuss some ideas to deal with the calculation of distance based centrality indices in Section 3.10.

3.6.3 Shortcut Values as a Vitality-Like Index

Although shortcut values are not a vitality index in the sense of Definition 3.6.1, they are nevertheless based on the concept of vitality. Thus, we present shortcut values here as a vitality-like index.

The shortcut value for edge e is defined by the maximum increase in distance between any two vertices if $e = (u, v)$ is removed from the graph. It is clear that this maximum increase can only be found between vertices that use e for all of

[5] Wiener itself named it 'path number' which is misleading. Subsequent articles quoted it as 'Wiener Index' [592]

their shortest paths. We claim that the increase in path length is maximized for the pair (u, v). This can easily be seen as follows. Clearly, the increase in distance for the pair (u, v) equals the difference between the length of e and the length of the shortest path p from u to v that does not use e. Further, other pair of vertices will either use their old path with e replaced by p, or use an alternative that is shorter than that.

Alternatively, the shortcut value can also be defined as the maximum relative increase in distance when all edge lengths are non-negative. In this case, the length of a shortest path using e is larger than the length of e, such that the relative increase is also maximized for the pair (u, v).

The shortcut values for all edges can be computed naïvely by $m = |E|$ many calls to a single-source shortest-path routine. Section 4.2.2 introduces a more efficient algorithm that is as efficient as computing $|V|$ single-source-shortest paths trees.

The notion of a shortcut value for an edge can be directly generalized to vertices, as the maximum increase in distance if the vertex is deleted.

3.6.4 Stress Centrality as a Vitality-Like Index

Stress centrality can be viewed as a vitality-like measure: Stress centrality (Section 3.4.1) counts the number of shortest paths containing a vertex or an edge and can thus be interpreted as the number of shortest paths that are lost if the vertex or edge is removed from the graph.

This sounds like a vitality measure but there is a crucial difference to the definition of vitality: The number of lost shortest paths has to be measured relatively to the number of shortest paths in the original graph. This is important, because a simple example shows that the total number of shortest paths can actually increase if a vertex or edge is removed from a graph (see Figure 3.9).

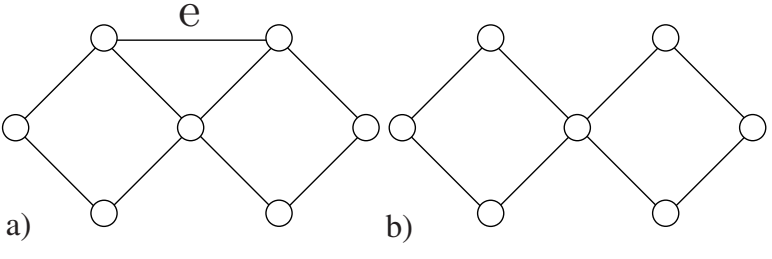


Fig. 3.9. The figure shows that the removal of an edge can actually increase the number of shortest paths in a graph

On the left side of Figure 3.9 (a) a small graph is shown with a total number of 54 shortest paths, 8 of them containing edge e. After the removal of e we find 64 shortest paths in the resulting graph. Of course, 18 of them are now longer

than before. When will the removal of an edge lead to an increase in the edge number? In this example, edge e is a shortcut for some of the paths from or to the two outermost vertices. As an example, we will take the path from the left outermost vertex to the right outermost vertex. As soon as e is removed, the distance between these nodes increases by one. Additionally, the number of shortest paths between them increases by three because now there are four paths with length 4 instead of only one with length 3 as before.

To interpret the stress centrality as a vitality measure we have to disregard shortest paths that have an increased length after the removal of an element. To formalize this idea we will give a definition of $f(G \setminus \{x\})$ that allows us to interpret the stress centrality of a vertex or an edge as vitality.

Let $f(G)$ be the number of all shortest paths in G and $f(G\backslash\{v\})$ be defined as following:

$$f(G\backslash\{v\}) = \sum_{s\in V}\sum_{t\in V} \sigma_{st}[d_G(s,t) = d_{G\backslash\{v\}}(s,t)] \tag{3.22}$$

The definition is given in Iverson-Notation, first described in [322], adapted by Knuth in [365]. The term inside the parentheses can be any logical statement. If the statement is true the term evaluates to 1, if it is false the term is 0. This notation makes the summation much easier to read than the classical notation in which logical statements are combined with the index of the sum. The definition of $f(G \setminus \{v\})$ is thus defined as the sum over the number of all those shortest paths that have the same length as the distance of s and t in G.

Analogously, let $f(G\backslash\{e\})$ be defined as following:

$$f(G\backslash\{e\}) = \sum_{s\in V}\sum_{t\in V} \sigma_{st}(e)[d_G(s,t) = d_{G\backslash\{e\}}(s,t)] \tag{3.23}$$

Defined in this way, the stress centrality $C_S(x)$ of an element x is exactly the difference between $f(G)$ and $f(G\backslash\{x\})$. It is important to note that the definition of $f(G \setminus \{x\})$ does not match the formal definition for a vitality measure. Nonetheless, the similarity of both is evident and thus we will denote the stress centrality as a vitality-like centrality index.

3.7 Current Flow

Shortest paths centralities rely on the crucial assumption that the flow of information, or the transport of goods in general, takes place along shortest paths. This section describes the current flow centralities, which are appropriate when the flow of information or transport does not adhere to this shortest paths assumption, but rather follows the behavior of an electrical current flowing through an electrical network.

3.7.1 Electrical Networks

Current flow centralities are based on the flow of electrical current in a network. We briefly describe currents in electrical networks below, and refer to [67]

for an extensive discussion. An electrical network is defined by an undirected, connected, and simple graph $G = (V, E)$, together with a conductance function $c : E \to \mathbb{R}$. External electrical current enters and leaves this network, which is specified by a supply function $b : V \to \mathbb{R}$. Positive values of b represent entering current, negative values represent current that leaves the network, and the amounts of entering and leaving currents are required to be equal: $\sum_{v \in V} b(v) = 0$. Since it is useful to talk about the direction of a current in the undirected graph, each edge $e \in E$ is arbitrarily oriented to obtain an oriented edge $\overrightarrow{e}$, which results in an oriented edge set $\overrightarrow{E}$.

A function $x : \overrightarrow{E} \to \mathbb{R}$ is called a (electrical) current in $N = (G = (V, E), c)$ if

$$\sum_{(v,w) \in \overrightarrow{E}} x(v, w) - \sum_{(w,v) \in \overrightarrow{E}} x(w, v) = b(v) \text{ for all } v \in V$$

and

$$\sum_{e \in C} x(\overrightarrow{e}) = 0$$

for every cycle $C \subseteq E$, that is, for every cycle in the undirected graph G. The former equation is known as Kirchoff's current law, and the latter as Kirchoff's potential law. Negative values of x are to be interpreted as current flowing against the direction of an oriented edge.

Alternatively to the current x, an electrical flow can also be represented by potentials. A function $p : V \to \mathbb{R}$ is a (electrical) potential if $p(v) - p(w) = x(v, w)/c(v, w)$ for all $(v, w) \in \overrightarrow{E}$. As an electrical network $N = (G, c)$ has a unique current x for any supply b, it also has a potential p that is unique up to an additive factor [67].

Define the Laplacian matrix $L = L(N)$ of the electrical network N to be

$$L_{vw} = \begin{cases} \sum_{e \ni v} c(e) & \text{if } v = w \\ -c(e) & \text{if } e = \{v, w\} \\ 0 & \text{otherwise} \end{cases}$$

for $v, w \in V$. Then, a potential p for an electrical network $N = (G, c)$ and a supply b can be found by solving the linear system $Lp = b$.

Finally, for the purpose of stating centralities based on electrical currents, define a unit s-t-supply b_{st} as a supply of one unit that enters the network at s and leaves it at t, that is, $b_{st}(s) = 1, b_{st}(t) = -1$, and $b_{st}(v) = 0$ for all $v \in V \setminus \{s, t\}$.

3.7.2 Current-Flow Betweenness Centrality

Newman [443] first considered centrality measures based on electrical currents. The current-flow betweenness of a vertex represents the fraction of unit s-t-supplies that passes through that vertex, just as shortest paths betweenness

counts the fraction of shortest s-t-paths through a vertex. For a fixed s-t pair, the so-called throughput of a vertex v forms the current-flow equivalent of the number of shortest paths $\sigma_{st}(v)$ through v. More precisely, the throughput of a vertex $v \in V$ with respect to a unit s-t-supply b_{st} is defined as

$$\tau_{st}(v) = \frac{1}{2}\left(-|b_{st}(v)| + \sum_{e \ni v} |x(\overrightarrow{e})|\right).$$

Here, the term $-|b_{st}(v)|$ sets the throughput of a vertex with non-zero supply equal to zero. For the current-flow betweenness, this ensures that a given unit s-t-supply does not count for the throughput of its source and sink nodes s and t. Further, the term $\frac{1}{2}$ adjusts for the fact that the summation counts both the current into and out of the vertex v.

Using the throughput definition, the current-flow betweenness centrality $c_{CB} : V \to \mathbb{R}$ for an electrical network $N = (G = (V, E), c)$ is defined as

$$c_{CB}(v) = \frac{1}{(n-1)(n-2)} \sum_{s,t \in V} \tau_{st}(v),$$

for all $v \in V$, where $1/(n-1)(n-2)$ is a normalizing constant. Thus, current-flow betweenness measures the fraction of throughput through vertex v, taken over all possible s-t pairs. Since an electrical network has a unique current for a given supply, current-flow betweenness is well defined.

3.7.3 Current-Flow Closeness Centrality

As with betweenness, the concept of closeness can also be extended from shortest paths to electrical current. For shortest paths, closeness is a measure of the shortest path distance from a certain vertex to all other vertices. For electrical current, Brandes and Fleischer [94] propose a closeness centrality that measures the distance between two vertices v and w as the difference of their potentials $p(v) - p(w)$. Their current-flow closeness centrality $c_{CC}(v) : V \to \mathbb{R}$ is defined as

$$c_{CC}(v) = \frac{n-1}{\sum_{t \neq v} p_{vt}(v) - p_{vt}(t)}$$

for all $v \in V$, where $(n-1)$ is again a normalizing factor. Here, the subscript vt on the potentials means that the potential stems from a unit v-t-supply b_{vt}.

Interestingly, Brandes and Fleischer [94] prove that current-flow closeness centrality is equal to information centrality. Stephenson and Zelen [533] introduced information centrality to account for information that flows along all paths in a network, rather than just along shortest paths. Information centrality also takes into account that certain paths carry a larger amount of information than others. Mathematically, information centrality $c_I : V \to \mathbb{R}$ is defined by

$$c_I(v)^{-1} = nM_{vv} + \text{trace}(M) - \frac{2}{n},$$

where the matrix M is defined as $(L+U)^{-1}$, with L being the Laplacian matrix, and U being a matrix of the same size with all entries equal to one.

3.8 Random Processes

Sometimes, it may not be possible for a vertex to compute shortest paths because of a lack of global knowledge. In such a case, shortest paths based centralities make no sense, and a random-walk model provides an alternative way of traversing the network. In a random walk something walks from vertex to vertex, following the edges of the network. Reaching some vertex v, it chooses one of the edges of v randomly to follow it to the next vertex.

The 'travel' of a bank note is a typical example for such a random walk. Somebody gets a brand new bill from her bank and gives it to someone else she encounters later on. Normally, nobody has any intention to give the bank note to someone special and the same bill may get to the same person more than once. For a marketing study, it could be of interest to find out the person or company who mediates most of these transactions. In the next section, we will have a closer look at the so-called random walk betweenness centrality that calculates the hot spots of mediation in such transactions.

3.8.1 Random Walks and Degree Centrality

In the case of undirected graphs, an observation can be made that relates the random-walk centrality with its complex definition to the most basic of all centralities, degree.

In the following theorem we prove that the stationary probabilities in the canonical random walk on a graph are proportional to the degree of the vertex.

Theorem 3.8.1. $p_{ij} = \frac{a_{ij}}{d(i)} \implies \pi_i = \frac{d(i)}{\sum_{v \in V} d(v)}$

Proof.

$$(\pi P)_j = \sum_{i \in V} \pi_i p_{ij} = \frac{\sum_{i \in V} d(i) p_{ij}}{\sum_{v \in V} d(v)} = \frac{\sum_{i \in V} a_{ij}}{\sum_{v \in V} d(v)} = \frac{d(j)}{\sum_{v \in V} d(v)} = \pi_j$$

□

3.8.2 Random-Walk Betweenness Centrality

The random-walk betweenness centrality introduced in [443] is based on the following idea. Suppose that vertex s has a message for vertex t but neither s nor any other vertex knows how to send it to t on a shortest path. Each vertex that gets the message for vertex t will just send it to any of its adjacent vertices at random. We assume that the graph is unweighted, undirected and connected.

This so-called random walk is modeled by a discrete-time stochastic process. At time 0, vertex s sends a message to one of its neighbors. If the message reaches vertex t at any time it will not be forwarded any further and such be absorbed by t. More formally, let m_{ij} describe the probability that vertex j sends the message to vertex i in time $k+1$ if it had it at time k:

$$m_{ij} = \begin{cases} \frac{a_{ij}}{d(j)} & \text{if } j \neq t \\ 0 & \text{else} \end{cases} \tag{3.24}$$

where a_{ij} denotes the ij-th element of the adjacency matrix A (see Section 2.3) and $d(j)$ is the degree of vertex j. The resulting matrix is denoted by M. Let D be the degree matrix of the graph:

$$d_{ij} = \begin{cases} d(i) & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{3.25}$$

The inverse D^{-1} of this matrix has the inverted vertex degrees on its diagonal, and is zero elsewhere. Because of the special behavior of vertex t the matrix notation $M = A \cdot D^{-1}$ is not correct. Removing the t-th row and column of all matrices yields a correct relation between the three matrices:

$$M_t = A_t \cdot D_t^{-1}, \tag{3.26}$$

where the index denotes the missing row and column, respectively.

Random-walk betweenness centrality considers all paths that a random walk can use, as well as the probabilities that such paths are used. Thus, the question arises how to compute the set of used paths, and how to compute the probability of using a single one of these paths. To guide the reader on his way, we first discuss how many different $i - j$ paths of length r exist in a given graph, where i and j are arbitrarily chosen vertices. It can easily be seen that the answer is $(A^r)_{ij}$, where A^r denotes the rth power of A. However, we are not interested in the number of random walks, but in the probability that a random walk of r steps, that starts at s, ends in vertex j. This is given by the r-th power of M_t at row j, column s, denoted by $(M_t^r)_{js}$. With this, the probability that the message is sent to vertex i in step $r + 1$ is given by:

$$\left(M_t^{r+1}\right)_{js} = m_{ij}^{-1} \left(M_t^r\right)_{js} \tag{3.27}$$

Now, we are interested in the probability that vertex j is sending a message that is starting at s to vertex i, summing over all paths, beginning at length 0 to ∞.

Note that all entries in any matrix M_t^r are values between 0 and 1, and thus the sum over all paths is convergent (see Theorem 3.9.2):

$$\sum_{r=0}^{\infty} m_{ij}^{-1} \left(M_t^r\right)_{js} = m_{ij}^{-1} [(I_{n-1} - M_t)^{-1}]_{js} \tag{3.28}$$

where I_{n-1} is the identity matrix of dimension $n - 1$.

Let $\boldsymbol{s}$ be a vector with dimension $n - 1$ that is 1 at vertex s and 0 else. Writing equation 3.28 in matrix notation we get:

$$\boldsymbol{v^{st}} = D_t^{-1} \cdot (I - M_t)^{-1} \cdot \boldsymbol{s} \tag{3.29}$$

$$= (D_t - A_t)^{-1} \cdot \boldsymbol{s} \tag{3.30}$$

The vector $\boldsymbol{v^{st}}$ describes the probability to find the message at vertex i while it is on its random walk from vertex s to vertex t. Of course, some of the random walks will have redundant parts, going from vertex a to vertex b and back again to vertex a. It does not seem reasonable to give a vertex a high centrality if most of the random walks containing it follow this pattern. Since the network is undirected every cycle will be accounted for in both directions, thus extinguishing each other. It is important to note that $\boldsymbol{v^{st}}$ contains only the net probability that disregards these cycles.

At this point, it becomes clear that random walks are closely related to current flows in electrical networks, see Section 3.7. Indeed, consider an electrical network $N = (G, c)$ with unit edge weights $c(e) = 1$ for all $e \in E$. The unit edge weights yield a Laplacian matrix $L(N) = D - A$, where D is the degree matrix and A the adjacency matrix of the graph G. So, a potential p_{st} in N for a unit s-t-supply b_{st} is a solution to the system $Lp_{st} = b_{st}$. The matrix L is not of full rank, but this problem can be circumvented by fixing one potential, say for vertex v, since potentials are unique up to an additive factor. Removing the rows and columns corresponding to the fixed vertex v yields the matrices L_v, D_v, and A_v, where L_v has full rank and is thus invertible. We conclude that a potential p_{st} for the unit s-t-supply b_{st} is given by $p_{st} = L_v^{-1} b_{st} = (D_v - A_v)^{-1} b_{st}$. The latter is equivalent to Equation (3.29) above, which shows the relation between electrical currents and potentials and random walks. For a more in-depth discussion of this relation, we refer to [67].

Thus, the random-walk betweenness centrality $c_{RWB} : V \to \mathbb{R}$ that we are looking for is equivalent to current-flow betweenness, that is, $c_{RWB}(v) = c_{CB}(v)$ for all $v \in V$. Newman [443] and Brandes and Fleischer [94] describe this betweenness equivalence in more detail.

3.8.3 Random-Walk Closeness Centrality

The same approach gives a kind of random-walk closeness centrality, where we look for the mean first passage time (MFPT). A centrality based on MFPT is introduced as Markov centrality in [580]. The mean first passage time m_{st} is defined as the expected number of nodes a particle or message starting at vertex s has encountered until it encounters vertex t for the first time. It is given by the following series:

$$m_{st} = \sum_{n=1}^{\infty} n \cdot f_{st}^{(n)} \tag{3.31}$$

where $f_{st}^{(n)}$ denotes the probability that t is arrived for the first time after exactly n steps. Let M denote the MFPT matrix in which m_{st} is given for all pairs s, t. M can be computed by the following equation:

$$M = (I - EZ_{dg})\, D \tag{3.32}$$

where I denotes the identity matrix, E is a matrix containing all ones, and S is a diagonal matrix with:

$$s_{ij} = \begin{cases} \frac{1}{\pi(v)} & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{3.33}$$

$\boldsymbol{\pi}$ denotes the stationary distribution of the random walk in the given graph (see Section 2.4), i.e., the expected relative time a particle will be on vertex v during the random walk. (This model assumes that the transport of the message or particle to another nodes takes virtually no time.) The matrix Z_{dg} agrees with the so called fundamental matrix Z on the diagonal but is 0 everywhere else. Matrix Z itself is given by:

$$Z = \left(I - A - 1_n \boldsymbol{\pi}^T\right)^{-1} \tag{3.34}$$

where 1_n is a column vector of all ones. The Markov centrality $c_M(v)$ is now defined as the inverse of the average MFPT for all random walks starting in any node s with target v (or vice versa):

$$c_M(v) = \frac{n}{\sum_{s \in V} m_{sv}} \tag{3.35}$$

This centrality can be defined for both directed and undirected networks. In directed networks the centrality is meaningfully defined for both, the average MFPT for random walks ending in v or leaving v. The expected number of steps from v to all other vertices or from all other vertices to v might be interpreted as a distance from v to all other vertices if a particle or information uses a random walk. Thus, the Markov centrality of a vertex is a kind of a (averaged) random-walk closeness centrality.

3.9 Feedback

This section presents centralities in which a node is the more central the more central its neighbors are. Some of these measures like Katzs status index belong to the oldest centralities presented in this chapter, others have their roots in the analysis of social networks. A third group belongs to the big class of analysis methods for the Web graph that is defined as the set of pages in the WWW connected by Web links.

Note, that in the following subsections centrality indices will be denoted as vectors. All feedback centralities are calculated by solving linear systems, such that the notation as a vector is much more convenient than using a function expressing the same. We just want to state here that all centrality indices presented here are fulfilling the definition of a structural index in Definition 3.2.1 if $c_X(i)$ is defined as $(c_X)_i$.

3.9.1 Counting All Paths – The Status Index of Katz

One of the first ideas with respect to feedback centralities was presented by Leo Katz [352] in 1953. It is based on the following observation: 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.

The idea of Katz is therefore to count additionally all indirect votes where the number of intermediate individuals may be arbitrarily large.

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. 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

$$\boldsymbol{c}_K(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$$

if the infinite sum converges.

In matrix notation we have

$$\boldsymbol{c}_K = \sum_{k=1}^{\infty} \alpha^k (A^T)^k \mathbf{1}_n. \tag{3.36}$$

(Note that $\mathbf{1}_n$ is the n-dimensional vector where every entry is 1, cf. also Chapter 2.)

To guarantee convergence we have to restrict α.

Theorem 3.9.1. *If A is the adjacency matrix of a graph G, $\alpha > 0$, and λ_1 the largest eigenvalue of A, then*

$$\lambda_1 < \frac{1}{\alpha} \Longleftrightarrow \sum_{k=1}^{\infty} \alpha^k A^k \text{ converges.}$$

For the proof see, e.g., [208].

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

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

or, in another form

$$(I - \alpha A^T) \boldsymbol{c}_K = \mathbf{1}_n,$$

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

3.9.2 General Feedback Centralities

In this subsection three centralities that are well known in the area of social network analysis are described.

Bonacich's Eigenvector Centrality. In 1972 Phillip Bonacich introduced a centrality measure based on the eigenvectors of adjacency matrices [71]. 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. 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,
b. the convergence of an infinite sequence, and
c. the solving of a linear equation system

In the following we describe all three approaches and call the results $\boldsymbol{s}^a$, $\boldsymbol{s}^b$, and $\boldsymbol{s}^c$.

First, we explain the factor analysis approach. For a better understanding think of the graph as a friendship network, where an edge denotes friendship between the persons that are modeled as vertices. We want to define a centrality that measures the ability to 'find friends'. Thus, we are interested in a vector $\boldsymbol{s}^a \in \mathbb{R}^n$, such that the i-th entry $\boldsymbol{s}^a_i$ should hold the interaction or 'friendship' potential of the vertex i. We declare that $\boldsymbol{s}^a_i \boldsymbol{S}^a_j$ 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 $\boldsymbol{s}^a$ that minimizes the following expression:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{s}^a_i \boldsymbol{s}^a_j - a_{ij})^2 \tag{3.37}$$

A second approach presented by Bonacich is based on an infinite sequence. For a given $\lambda_1 \neq 0$ we define

$$\boldsymbol{s}^{b_0} = \mathbf{1}_n \quad \text{and} \quad \boldsymbol{s}^{b_k} = A\frac{\boldsymbol{s}^{b_{k-1}}}{\lambda_1} = A^k \frac{\boldsymbol{s}^{b_0}}{\lambda_1^k}\,.$$

According to Theorem 3.9.2, the sequence

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

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

Theorem 3.9.2. *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and λ_1 the largest eigenvalue of A, then*

$$\lim_{k\to\infty} A^k \frac{\boldsymbol{s}^{b_0}}{\lambda_1^k}$$

converges towards an eigenvector of A with eigenvalue λ_1.

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:

$$\boldsymbol{s}_i^c = \sum_{j=1}^{n} a_{ij}\boldsymbol{s}_j^c \quad \text{resp.} \quad \boldsymbol{s}^c = A * \boldsymbol{s}^c \tag{3.38}$$

This equation system has a solution only if $\det(A-I) = 0$. We solve $\lambda\boldsymbol{s} = A\boldsymbol{s}$, the eigenvalue problem for A, instead. According to Theorem 3.9.3, under the given conditions for the graph defined above, exactly one eigenvector contains entries that are either all positive or all negative. Therefore, we use the absolute value of the entries of this eigenvector as the solution.

Theorem 3.9.3. *Let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix of an undirected and connected graph. Then:*

- *The largest eigenvalue λ_1 of A is simple.*
- *All entries of the eigenvector for λ_1 are of the same sign and not equal to zero.*

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

$$\boldsymbol{c}_{\text{EV}} = \frac{|\boldsymbol{s}^c|}{||\boldsymbol{s}^c||} \tag{3.39}$$

In general, whenever one has a graph with multiple, poorly spanned dense clusters, no single eigenvector will do a satisfactory job of characterizing walk-based centrality. This is because each eigenvector will tend to correspond to loadings on a given cluster: Everett and Borgatti [194] 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, $c_i \in [0, 1]$.

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, see 3.37, 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.

According to [71], the eigenvector centrality can be applied to disconnected graphs. In this case several eigenvectors have to be taken into account, one for every component of the graph.

Hubbell Index. Even earlier than Bonacich, Charles Hubbell [319] suggested in 1965 a centrality measure based on the solution of a system of linear equations. His approach uses directed weighted graphs where the weights of the edges may be real numbers. A graph may contain loops but has to be simple, too. Please note that the adjacency matrix $W(G)$ of a graph G is asymmetric and contains real numbers instead of zeros and ones.

The general assumption of Hubbell's centrality measure is similar to the idea of Bonacich: the value of a vertex v depends on the sum of the values of each adjacent vertex w multiplied with the weight of the incident edge $e = (v, w)$. Therefore, the following equation should hold: $\boldsymbol{e} = W\boldsymbol{e}$. To make the equation system solvable an additional parameter called the exogenous input or the boundary condition $\boldsymbol{E}$ has to be added. This is a column vector containing external information for every vertex. Hubbell suggested that if this boundary condition is unknown $\boldsymbol{E} = \mathbf{1}$ may be used.

The final equation is

$$\boldsymbol{s} = \boldsymbol{E} + W\boldsymbol{s} \tag{3.40}$$

Through a simple transformation this equation can be rewritten into $\boldsymbol{s} = (I - W)^{-1}\boldsymbol{E}$. This system has a solution if the matrix $(I - W)$ is invertible. Since $\frac{I}{(I-W)} = \sum_{k=1}^{\infty} W^k$ holds, this is identical to the problem of the convergence of the geometric series. According to Theorem 3.9.1, the series converges against $\frac{I}{(I-W)}$ if and only if the largest eigenvalue λ_1 of W is less than one.

The solution $\boldsymbol{S}$ of the equation system 3.40 is called Hubbell centrality $\boldsymbol{c}_{\mathrm{HBL}}$ or Hubbell Index.

Bonacich's Bargaining Centrality. Both feedback centralities presented so far follow the idea of positive feedback: the centrality of a vertex is higher if it is connected to other high-valued vertices. In 1987 Phillip Bonacich [72] suggested a centrality which is not restricted to this concept. His idea supports both, the positive influence as seen for example in communication networks, and the negative influence as seen in bargaining situations. In bargaining situations a participant is strong if he is connected to individuals having no other options and are therefore weak.

Bonacich's bargaining centrality is defined for unweighted and directed graphs $G = (V, E)$ without loops. Therefore the adjacency matrix is not necessarily symmetric and contains only zeros and ones. The definition is

$$c_{\alpha,\beta}(i) = \sum_{j=1}^{n} (\alpha + \beta * c_{\alpha,\beta}(j)) a_{ij}$$

or, in matrix notation,

$$c_{\alpha,\beta} = \alpha(I - \beta A)^{-1} A\mathbf{1} \tag{3.41}$$

As can easily be seen from the matrix notation, the parameter α is just a scaling factor. Bonacich suggests a value such that $\sum_{i=1}^{n} c_{\alpha,\beta}(i)^2 = n$ holds. Therefore only the second parameter β is of interest. This parameter may be chosen either positive or negative, covering positive or negative influence, respectively. The choice $\beta = 0$ leads to a trivial solution where the centrality correlates with the degree of the vertices. A negative value for β may lead to negative values for the centralities of the vertices. Additionally it follows from the equation that the larger $|\beta|$ the higher the impact of the structure of the network on the centrality index is.

Equation 3.41 is solvable if the inverse of $(I - \beta A)$ exists. According to Theorem 3.9.4, this inverse exists if no eigenvalue of A is equal to 1.

Theorem 3.9.4. *Let $M \in \mathbb{R}^{n \times x}$ be a matrix and $\lambda_1, \ldots, \lambda_n$ the eigenvalues of M.*

$$(I - M) \text{ is invertible} \iff \forall i \in \{1 \ldots n\}\ \lambda_i \neq 1\,.$$

We call $c_{\alpha,\beta}$ the bargaining centrality $\boldsymbol{c}_{\mathrm{BRG}}$.

In this subsection three different approaches to measure feedback centrality values where presented. They seem very similar but differences are for example the coverage of weighted versus unweighted edges or positive versus negative influence networks.

3.9.3 Web Centralities

Many people use the World Wide Web to search for information about interesting topics. Due to the immense size of the network consisting of Web pages that are connected by hyperlinks powerful search engines are required. But how does a search engine decide which Web pages are appropriate for a certain search query? For this, it is necessary to score the Web pages according to their relevance or importance. This is partly done by a pure text search within the content of the pages. Additionally, search engines use the structure of the network to rank pages and this is where centrality indices come into play.[6]

In this section we discuss three main representatives of Web-scoring algorithms:

[6] Many concepts used for the 'Web centralities' are not new, especially the idea of eigenvectors as a centrality was known long before the Web was established. We decided to use this headline due to the interest of the last years into this topic.

– PageRank
– Hubs & Authorities
– SALSA

Whereas PageRank only takes the topological structure into account, the latter two algorithms combine the 'textual importance' of the Web page with its 'topological importance'. Moreover, Hubs & Authorities (sometimes also called HITS algorithm) assigns two score values to each Web page, called hub and authority. The third approach, SALSA, discussed at the end of this section, is in some sense a combination of the others.

In the following we assume that the Web is represented by a digraph $G = (V, E)$ with a one-to-one-correspondence between the Web pages and the vertices $v \in V$ as well as between the links and the directed edges $(v, w) \in E$.

The Model of a Random-Surfer. Before defining centrality indices suitable for the analysis of the Web graph it might be useful to model the behavior of a Web surfer. The most common model simulates the navigation of a user through the Web as as a random walk within the Web graph.

In Section 2.4 the concept of random walks in graphs was introduced. The Web graph $G = (V, E)$ is formally defined as V the set of all Web pages p_i where an edge $e = (p_i, p_j) \in E$ is drawn between two pages if and only if page p_i displays a link to page p_j. As the Web graph is usually not strongly connected the underlying transition matrix T is not irreducible and may not even be stochastic as 'sinks' (vertices without outgoing links) may exist. Therefore, the transition matrix T of the Web graph has to be modified such that the corresponding Markov chain converges to a stationary distribution.

To make the matrix T stochastic we assume that the surfer jumps to a random page after he arrived at a sink, and therefore we set all entries of all rows for sinks to $\frac{1}{n}$. The definition of the modified transition matrix T' is

$$t'_{ij} = \begin{cases} \frac{1}{d^+(i)}, & \text{if } (i,j) \in E \\ \frac{1}{n}, & \text{if } d^+(i) = 0 \end{cases}$$

This matrix is stochastic but not necessarily irreducible and the computation of the stationary distribution π' may not be possible. We therefore modify the matrix again to get an irreducible version T''. Let $E = \frac{1}{n}\mathbf{1}_n^T\mathbf{1}_n$ be the matrix with all entries $\frac{1}{n}$. This matrix can be interpreted as a 'random jump' matrix. Every page is directly reachable from every page by the same probability. To make the transition matrix irreducible we simply add this new matrix E to the existing matrix T':

$$T'' = \alpha T' + (1 - \alpha)E$$

Factor α is chosen from the range 0 to 1 and can be interpreted as the probability of either following a link on the page by using T' or performing a jump to a random page by using E. The matrix T'' is by construction stochastic

and irreducible and the stationary distribution π'' may be computed for example with the power method (see Section 4.1.5).

By modifying E, the concept of a random jump may be adjusted for example more towards a biased surfer. Such modifications leads directly to a personalized version of the Web centrality indices presented here. For more details on this topic, see Section 5.2.

PageRank. PageRank is one of the main ingredients of the search engine Google [101] and was presented by Page et al. in 1998 [458]. The main idea is to score a Web page with respect to its topological properties, i.e., its location in the network, but independent of its content. PageRank is a feedback centrality since the score or centrality of a Web page depends on the number and centrality of Web pages linking to it

$$\boldsymbol{c}_{\mathrm{PR}}(p) = d \sum_{q \in \Gamma_p^-} \frac{\boldsymbol{c}_{\mathrm{PR}}(q)}{d^+(q)} + (1-d)\,, \tag{3.42}$$

where $\boldsymbol{c}_{\mathrm{PR}}(q)$ is the PageRank of page q and d is a damping factor.

The corresponding matrix notation is

$$\boldsymbol{c}_{\mathrm{PR}} = dP\boldsymbol{c}_{\mathrm{PR}} + (1-d)\mathbf{1}_n\,, \tag{3.43}$$

where the *transition matrix* P is defined by

$$p_{ij} = \begin{cases} \frac{1}{d^+(j)}, \text{ if } (j,i) \in E \\ 0, \quad \text{otherwise} \end{cases}$$

This is equivalent to $p_{ij} = \frac{1}{d^+(j)} a_{ji}$ or $P = D^+ A$ in matrix notation, where D^+ denotes the diagonal matrix where the i-th diagonal entry contains the out degree $d^+(i)$ of vertex i.

Mostly, the linear system 3.43 is solved by a simple power (or Jacobi) iteration:

$$\boldsymbol{c}_{\mathrm{PR}}^k = dP\boldsymbol{c}_{\mathrm{PR}}^{k-1} + (1-d)\mathbf{1}_n. \tag{3.44}$$

The following theorem guarantees the convergence and a unique solution of this iteration if $d < 1$.

Theorem 3.9.5. *If $0 \leq d < 1$ then Equ. 3.43 has a unique solution $\boldsymbol{c}_{PR}^* = (1-d)\left(I_n - dP\right)^{-1}\mathbf{1}_n$ and the solutions of the dynamic system 3.44 satisfy $\lim_{k\to\infty} \boldsymbol{c}_{PR}^k = \boldsymbol{c}_{PR}^*$ for any initial state-vector $\boldsymbol{c}_{PR}^0$.*

A slightly different approach is to solve the following dynamic system

$$\boldsymbol{c}_{\mathrm{PR}}^k = dP\boldsymbol{c}_{\mathrm{PR}}^{k-1} + \frac{\alpha^{k-1}}{n}\mathbf{1}_n\,, \tag{3.45}$$

where $\alpha^{k-1} = \|\boldsymbol{c}_{\mathrm{PR}}^{k-1}\| - \|dP\boldsymbol{c}_{\mathrm{PR}}^{k-1}\|$. The solutions of this system converge to $\frac{\boldsymbol{c}_{\mathrm{PR}}^*}{\|\boldsymbol{c}_{\mathrm{PR}}^*\|}$, the normalized solution of 3.44.

Hubs & Authorities. Shortly after the presentation of PageRank, Kleinberg introduced the idea of scoring Web pages with respect to two different 'scales' [359], called hub and authority, where

"A good hub is a page that points to many good authorities"

and

"A good authority is a page that is pointed to by many good hubs".

In contrast to PageRank, Kleinberg proposed to include also the content of a Web page into the scoring process. The corresponding algorithm for determining the hub and authority values of a Web page consists of two phases, where the first phase depends on the search query and the second phase deals only with the link structure of the associated network.

Given the search query σ, in the first phase of the algorithm an appropriate subgraph $G[V_\sigma]$ induced by a set of Web pages $V_\sigma \subseteq V$ is extracted, where

- V_σ should be comparably small,
- V_σ should contain many pages relevant for the search query σ, and
- V_σ should contain many important authorities.

This goal is achieved by using algorithm 1 to calculate V_σ, the set of relevant Web pages.

Algorithm 1: Hubs & Authorities, 1st Phase

Output: V_σ, the set of relevant pages

Use a text based search engine for search query σ
Let W_σ be the list of results
Choose $t \in \mathbb{N}$
Let $W_\sigma^t \subset W_\sigma$ contain the t pages ranked highest
$V_\sigma := W_\sigma^t$
forall $i \in W_\sigma^t$ **do**
  $V_\sigma := V_\sigma \cup \Gamma^+(i)$
  if $|\Gamma^-(i)| \leq r$ *(r is a user-specified bound)* **then**
    $V_\sigma := V_\sigma \cup \Gamma^-(i)$
  else
    choose $\Gamma_r^-(i) \subseteq \Gamma^-(i)$ such that $|\Gamma_r^-(i)| = r$
    $V_\sigma := V_\sigma \cup \Gamma_r^-(i)$
return V_σ

The second phase of the Hubs & Authorities algorithm consists of computing the hub and authority scores for the Web pages in $G[V_\sigma]$ which is done by taking into account the mutual dependence between hubs and authorities. This mutual dependence can be expressed by

$$\boldsymbol{c}_{\text{HA-H}} = A_\sigma \boldsymbol{c}_{\text{HA-A}} \text{ assuming } \boldsymbol{c}_{\text{HA-A}} \text{ is known and} \tag{3.46}$$

$$\boldsymbol{c}_{\text{HA-A}} = A_\sigma^T \boldsymbol{c}_{\text{HA-H}} \text{ assuming } \boldsymbol{c}_{\text{HA-H}} \text{ is known,} \tag{3.47}$$

where A_σ is the adjacency matrix of $G[V_\sigma]$.

Algorithm 2: Hubs & Authorities Iteration

Output: Approximations for $\boldsymbol{c}_{\text{HA-H}}$ and $\boldsymbol{c}_{\text{HA-A}}$

$\boldsymbol{c}^0_{\text{HA-A}} := \mathbf{1}_n$
for $k = 1 \dots$ **do**
  $\boldsymbol{c}^k_{\text{HA-H}} := A_\sigma \boldsymbol{c}^{k-1}_{\text{HA-A}}$
  $\boldsymbol{c}^k_{\text{HA-A}} := A^T_\sigma \boldsymbol{c}^k_{\text{HA-H}}$
  $\boldsymbol{c}^k_{\text{HA-H}} := \frac{\boldsymbol{c}^k_{\text{HA-H}}}{\|\boldsymbol{c}^k_{\text{HA-H}}\|}$
  $\boldsymbol{c}^k_{\text{HA-A}} := \frac{\boldsymbol{c}^k_{\text{HA-A}}}{\|\boldsymbol{c}^k_{\text{HA-A}}\|}$

Since neither $\boldsymbol{c}_{\text{HA-H}}$ nor $\boldsymbol{c}_{\text{HA-A}}$ are known, Kleinberg proposes an iterative procedure including a normalization step shown in algorithm 2. He shows

Theorem 3.9.6. *If A_σ is the adjacency matrix of $G[V_\sigma]$ then $\lim_{k\to\infty} \boldsymbol{c}^k_{HA\text{-}A} = \boldsymbol{c}_{HA\text{-}A}$ and $\lim_{k\to\infty} \boldsymbol{c}^k_{HA\text{-}H} = \boldsymbol{c}_{HA\text{-}H}$, where $\boldsymbol{c}_{HA\text{-}A}$ ($\boldsymbol{c}_{HA\text{-}H}$) is the first eigenvector of $A^T_\sigma A_\sigma$ ($A_\sigma A^T_\sigma$)*

Therefore, the given iterative procedure is nothing but solving the eigenvector-equations

$$\lambda \boldsymbol{c}_{\text{HA-A}} = (A^T_\sigma A_\sigma)\boldsymbol{c}_{\text{HA-A}}$$
$$\lambda \boldsymbol{c}_{\text{HA-H}} = (A_\sigma A^T_\sigma)\boldsymbol{c}_{\text{HA-H}}$$

for the largest eigenvalue by a power iteration, see Section 4.1.5. The vector $\boldsymbol{c}_{\text{HA-A}}$ then contains the scores for the vertices with respect to their authority, whereas $\boldsymbol{c}_{\text{HA-H}}$ is the vector of hub scores.

SALSA. In 2000, Lempel and Moran developed the SALSA (Stochastic Approach for Link Structure Analysis) algorithm [387]. The authors introduced this new Web-scoring approach to retain on the one hand the intuitive and appealing idea of hubs and authorities and to provide the index on the other hand with a higher robustness against the so called 'TKC effect'. TKC stands for *Tightly-Knit Community*, a small set of highly connected Web pages that in some cases may cause the Hubs & Authorities algorithm to rank the corresponding Web pages high even if they cover only a small (or no) aspect of the query. To this end Lempel and Moran combined the ideas of PageRank with those of Hubs & Authorities.

SALSA is a 3-phase algorithm where the first phase is identical to the first phase of the Hubs & Authorities algorithm: it constructs the graph $G[V_\sigma]$ for a certain search query σ (see algorithm 1). In the second phase an artificial bipartite undirected graph $\bar{G}_\sigma = (V^h_\sigma \dot\cup V^a_\sigma, \bar{E})$ according to the algorithm 3 is

built. For the third phase of SALSA recall that the PageRank algorithm works with the transition matrix P which is the transposed adjacency matrix of the underlying graph with the non-zero columns weighted by their column sums. The Hubs & Authorities algorithm uses the product of the adjacency matrix A_σ of $G[V_\sigma]$ with its transpose. For SALSA the following matrices are defined:

P_σ: A_σ with each non-zero column weighted by its column sum
R_σ: A_σ with each non-zero row weighted by its row sum

Algorithm 3: SALSA, 2^{nd} phase

Output: The bipartite undirected graph $\bar{G}_\sigma$

forall $i \in V_\sigma$ **do**
  if $d^+(i) > 0$ **then**
    create a copy i^h of i in V_σ^h
  if $d^-(i) > 0$ **then**
    create a copy i^a of i in V_σ^a
forall $e = (i, j) \in E(G[V_\sigma])$ **do**
  create an undirected edge $\bar{e} = \{i^h, j^a\}$ in $\bar{E}$

Then the indices of the non-zero columns (rows) of $R_\sigma P_\sigma^T$ correspond to the elements in V_σ^h and those of $P_\sigma^T R_\sigma$ to V_σ^a. Define

A_σ^h: non-zero rows and columns of $R_\sigma P_\sigma^T$
A_σ^a: non-zero rows and columns of $P_\sigma^T R_\sigma$

and use power iteration (see Section 4.1.5) to compute the SALSA authority scores $\boldsymbol{c}_{\text{S-A}}$ and the SALSA hub scores $\boldsymbol{c}_{\text{S-H}}$.

3.10 Dealing with Insufficient Connectivity

Most of the centrality-measures presented so far assume that the underlying network is connected. If this is not the case, computing these centralities might be a problem. For local centrality indices, such as degree centrality, this connectivity assumption has no implications. However, this is not the case in general. In this section, we investigate how to deal with disconnected undirected graphs and weakly connected digraphs.

Consider, for example, the centralities based on shortest paths, such as the measures based on eccentricity or closeness. Both centralities depend on the knowledge of the shortest paths length $d(u, v)$ between all pairs of vertices u and v. For a disconnected undirected graph or a weakly connected digraph there are pairs of vertices for which this length is not defined, and it is not clear how to deal with them. A very naive approach would be to restrict the computation of centrality values to subgraphs where the measure is well defined, i.e., to compute

the centrality measure for a vertex with respect to its component or strong components in the case of digraphs. This approach is not very reasonable in most applications. Consider, for example, a (directed) network consisting of two (strong) components, where one is the complete graph of two vertices, and the other one is the complete graph with $n-2$ vertices, where n is large. Then the above approach yields a closeness value of 1 for all vertices, but it seems obvious that the vertices in the large component are much more central than the two other vertices.

3.10.1 Intuitive Approaches

A common way to deal with this problem is to simply multiply the centrality values with the size of the component, following the intuition that the vertices in large components are more important. This seems to be reasonable, but it is not proper unless the centrality measure behaves proportional to the size of the network. Computational experiments of Poulin, Boily and Mâsse [481] indicate that this is not the case for closeness and eccentricity.

Two other repair mechanisms use inverse path lengths, and arbitrary fixed values for the distance between unconnected vertices. The latter possibility yields an approximation of the desired centrality values. However, Botafogo et al. [88] have shown that the result strongly depends on the fixed value k for the unconnected vertex pairs. They defined a closeness-based measure for digraphs

$$c_{C'}(u) = \frac{\sum_{v \in V} \sum_{w \in V} d(v,w)}{\sum_{v \in V} d(u,v)} \tag{3.48}$$

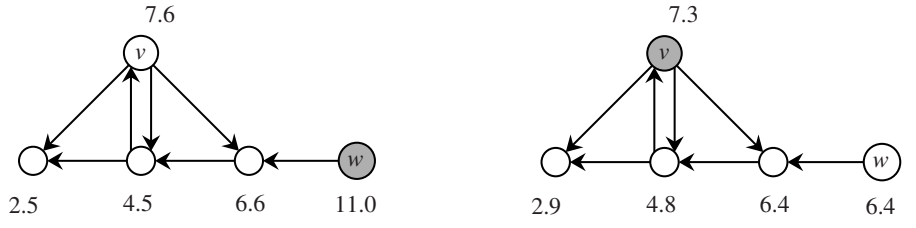
where the distance $d(u,v)$ between any unconnected vertex pair u and v is set to k. Clearly, an appropriate value for k is the number of vertices n, since the maximum distance between any two vertices is at most $n-1$. In the digraph of Fig. 3.10 the vertex reaching all other vertices is w. For $k = 2n$ w becomes the vertex with highest centrality value but for $k = n$ the vertex v which does not reach w has highest value. This example shows that the choice of k will crucially influence the order of centrality index values assigned to the vertices.

Moreover, the centrality based on the eccentricity does not make sense anymore in non-connected graphs or in non-strongly connected digraphs. If the fixed value is large enough, then it dominates all other distances in the graph and yields centrality values that differ only in a very small range.

The usage of inverse path lengths makes it more difficult to interpret and compare centrality values. By substituting the path lengths in the closeness centrality by their inverses, and multiplying the sum of the inverse path length by $(n-1)$, we do not obtain the closeness centrality but an entirely different centrality measure.

3.10.2 Cumulative Nominations

A more sophisticated approach was presented by Poulin, Boily and Mâsse [481]. Their starting point is a measure that is very similar to Bonacich's eigenvector



(a) $k = 2n$ (b) $k = n$

Fig. 3.10. The centralities with respect to the measure due to Botafogo et al. are shown. In each subfigure the vertex with the maximum value is colored grey

centrality. The *cumulative number of nominations* centrality $c_{\mathrm{CNN}}(i)$ of vertex i is defined to be the ith component of the l_1-normalized eigenvector corresponding to the largest eigenvalue of $A+I$, where A is the adjacency matrix. In other words, c_{CNN} is the solution of $(A+I-\lambda_1 I)p = 0$ under the restriction $\sum_i p_i = 1$. Therefore, Bonacich's centrality and the cumulative number of nominations only differ by a constant. Poulin, Boily and Mâsse claim that their measure when computed by an iterative algorithm converges faster and is more stable. Moreover, their centrality may be applied to bipartite graphs as the graph corresponding to $(A + I)$ is not bipartite, even if the graph for A is.

Due to the normalization, c_{CNN} is not independent of the size of the connected component. The more vertices the component contains, the smaller the absolute centrality values become. But, using the approach of iteratively solving

$$c^{k+1}_{\mathrm{CNN}} = (A + I)c^{k}_{\mathrm{CNN}},$$

the authors obtain the *cumulative nominations index of centrality*

$$c_{\mathrm{CN}}(i) = c_{\mathrm{CS}}(i) \lim_{k\to\infty} c^{k}_{\mathrm{CNN}}(i),$$

where $c_{\mathrm{CS}}(i)$ is the size of the component containing vertex i. This cumulative nominations index assigns a value of 1 to a vertex having an average structural position in a connected component.

In addition, the *cumulated nominations growth rate centrality index* of a vertex is defined as

$$c_{\mathrm{CNG}}(i) = \lim_{k\to\infty} \left[\left(\sum_j a_{ij} c^{k}_{\mathrm{CNN}}(j) + c^{k}_{\mathrm{CNN}}(i) \right) \frac{1}{c^{k}_{\mathrm{CNN}}(i)} \right],$$

and is the same for each vertex in a connected component.

This growth rate allows a comparison between different connected components. To this end, the *multi-component cumulated nominations centrality index* c_{MCN} is defined by

$$c_{\mathrm{MCN}}(i) = c_{\mathrm{CN}}(i) c_{\mathrm{CNG}}(i),$$

and, to take into account the (relative) size of the components (vertices in larger components should get a larger centrality score), we get the *corrected multi-component cumulated nominations centrality index*

$$c_{\mathrm{CMCN}}(i) = c_{\mathrm{MCN}}(i) c_{\mathrm{CS}}(i).$$

The authors report on computational experiments which indicate that neither c_{MCN} nor c_{CMCN} depends on n, hence both are centrality measures well suited for networks consisting of more than one component.

3.11 Graph- vs. Vertex-Level Indices

This section makes a connection between the analysis of a network on the level of vertices and on the level of the whole graph: Intuitively, it is clear that some graphs are more centralized than others, i.e., some graphs are more depending on the most central nodes than others. The star topology in which only one vertex v is connected to all others but all other vertices are only connected to v is a very centralized graph. A clique where every vertex is connected to every other vertex is not centralized.

Freeman [226] has proposed a very general approach with which the centralization $c_X(G)$ of a graph G can be calculated in relation to the values of any vertex centrality index c_X :

$$c_X(G) = \frac{\sum_{i \in V} c_X(j)^* - c_X(i)}{n-1} \tag{3.49}$$

where $c_X(j)^*$ denotes the largest centrality value associated with any vertex in the graph under investigation. This approach measures the average difference in centrality between the most central point and all others. If normalized centralities in the range of $[0,1]$ are used, the centralization value will also be in the range $[0,1]$ (for further details to the normalization of centrality indices see Section 5.1). Other obvious possibilities to generate a graph index from the distribution of centrality indices are to compute the variance of the values or the maximal difference between centrality values or any other statistics on these values.

On the other hand, also a structural index for graphs like the Wiener Index (see Section 3.6.2) can be transformed into a structural index for vertices. We want to formalize this idea by first defining a structural index for graphs.

Definition 3.11.1 (Structural Index for Graphs). *Let $G = (V, E)$ be a weighted, directed or undirected multigraph. A function $C\colon G \to \mathbb{R}$ is called a structural index for graphs if and only if the following condition is satisfied:* $\forall G' \simeq G : \Longrightarrow\ C(G') = C(G))$.

Let $f : V \to \mathbb{R}$ be any structural index on the vertices of a graph and let $\bigodot$ be an operator on the set of all vertices V, like the summation over

$f(v)$, the average of all terms $f(v)$, the calculation of the variance of all $f(v)$ or the maximum/minimum operator. Then $\bigodot V =: f(G)$ defines a graph measure because all structural indices on vertices are stable under isomorphism. On the other hand, let $f : G \leftarrow \mathbb{R}$ be a structural index on the whole graph. Let $G(v, d)$ be the induced subgraph in which all vertices are contained with a hopping distance to v of no more than d. I.e. $G(v, d) = (V', E')$ is a subset of $G = (V, E)$ with $V' = \{w \in V | d(w, v) \leq d\}$ and $E' = \{(x, y) \in V' \times V' | (x, y) \in E\}$. Then $f(G(d, v))$ defines at least a structural index on the vertices of this graph, and in most cases also a reasonable vertex centrality index.

With this we can for example derive a centrality index from the Wiener Index by constraining the calculation of it to subgraphs with a small diameter. Such an approach might be useful in networks, where a message will not be transported more than k steps before it dies, as it is the case in some peer-to-peer network protocols. The new centrality index would then measure how well connected a node is within the subgraph of diameter k. It should be noted, however, that these subgraphs will be of different sizes in most cases. How centrality index values can be compared with each other in this case is discussed in the section about applying centrality indices to disconnected graphs (see Section 3.10).

3.12 Chapter Notes

Many interesting facts and a good overview of centrality indices used in social network analysis are given in [569]. Hage and Harary carried some of these ideas to a graph theoretic notation [269].

The notion of 'centrality' is very graphic and can be supported by adequate visualization. An approach to visualizing centrality measures in an intuitive way is [96] (see also Figure 1.2).

Closeness Centrality. Closeness centrality is often cited in the version of Sabidussi [500]. Nonetheless, it was also mentioned by Shimbel [519] but not as a centrality index. He defined the *dispersion* as the sum of all distances in a graph. Thus, it is a synonym for the Wiener Index [583] (see also Section 3.6.2). For directed graphs he defined the accessibility $A(i, G)$ of G from vertex i as $A(i, G) = \sum_{j \in V} d(i, j)$ and the accessibility $A^{-1}(i, G)$ of vertex i from G as $A^{-1}(i, G) = \sum_{j \in V} d(j, i)$. These two values are easily recognized as directed version of the closeness centrality.

Betweenness Centrality. Betweenness centrality was introduced by Freeman [226] and, independently, Anthonisse [32]. He was inspired by ideas of Bavelas [50]. Bavelas was the first who tried to map psychological situations to graphs. His main interest was the notion of centers (called 'innermost regions'), but he additionally discussed the following example: A group of Italian speaking women is employed in a large garment factory. Only one of them speaks English. Bavelas states: "It is difficult to imagine that the English speaking member would

be other than central with respect to communication which had of necessity to pass through her (...) It is interesting in passing to point out the importance of the English speaking member with respect to the group's perception of the 'outside'. (...)To the extent that policy decisions are based upon information, as to the state of affairs 'outside', withholding information, coloring or distroting it in transmission, or in other ways misrepresenting the state of the outside will fundamentally affect these decisions."

Both edge and vertex betweenness have found many applications in the analysis of social networks (for example [457]), sexual intercourse networks (see [81]), or terrorist networks (for example [111]). Another interesting application is a graph clustering algorithm based on edge betweenness centrality [445]. Modern techniques try to approximate the expected congestion in a communication network using vertex betweenness [522]. According to this, the probability for congestion can be decreased by scaling the bandwidth proportional to betweenness centrality of a vertex. Nonetheless, betweenness centrality does not always scale with the expected congestion, as indicated in [304] (see also the introduction to Chapter 4).

The algorithmic complexity of this index is $\mathcal{O}(nm)$ for unweighted networks and $\mathcal{O}(nm + n^2 \log n)$ for weighted networks (for details see Section 4.2. Since this runtime makes it very hard to compute the betweenness centrality for graphs bigger than approximately 10,000 vertices, one should consider alternatives. In Section 4.3.1 we will discuss a way to approximate betweenness centrality. In Section 5.2.1 a personalized variant of the betweenness centrality is presented. A directed version of shortest-path betweenness centrality was first discussed in [32] and reinvented in [578].

Feedback Centralities. As far as we know, the first paper that defined a feedback centrality (without actually naming it in this way) was published by Seeley [510]. The status index of Katz was presented shortly afterwards in 1953 [352]. The index defined by Hubbell [319] and the approach presented by Bonacich [71] focus on the idea of propagating strength, where a high value vertex influences all vertices in his vicinity. All of these approaches solely focus on positive feedback relations. The first centrality index that covered negative feedback relation was presented by Bonacich [72].

Web Centralities. We covered three Web centralities: PageRank ([101, 458]), Hubs & Authorities ([359]) and SALSA ([387]). Especially for PageRank a whole bunch of papers is available and therefore we just give three references ([61, 378, 379]) which are a good starting point for further investigations of the topic.