# A measure for the cohesion of weighted networks

Non Peer-reviewed author version

EGGHE, Leo & ROUSSEAU, Ronald (2003) A measure for the cohesion of weighted networks. In: Journal of the American Society for Information Science and Technology, 54(3). p. 193-202.

DOI: 10.1002/asi.10155
Handle: http://hdl.handle.net/1942/768

# A measure for the cohesion of weighted networks

**LEO EGGHE**
LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
and UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
E-mail: leo.egghe@luc.ac.be

**and**

**RONALD ROUSSEAU**
KHBO, Industrial Sciences and Technology,
Zeedijk 101, B-8400 Oostende, Belgium
& UIA, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium
& LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium
E-mail: ronald.rousseau@kh.khbo.be

ABSTRACT

A generalization of both the Botafogo-Rivlin-Shneiderman compactness measure and the Wiener index is presented. These new measures for the cohesion of networks can be used in case a dissimilarity value is given between nodes in a network or a graph. It is illustrated how a set of weights between connected nodes can be transformed into a set of dissimilarity measures for all nodes. The new compactness measure for the cohesion of weighted graphs has several desirable properties related to the disjoint union of two networks. Finally, an example is presented of the calculation of the new compactness measures for a co-citation and a bibliographic coupling network.

Keywords: cohesion, compactness, weighted networks, weighted graphs, generalized Wiener index, co-citation, bibliographic coupling

## 1. Introduction

Graph theory pervades both the theoretical and the practical domains of many sciences, including the information sciences. Indeed, pages and hyperlinks of the World Wide Web may be viewed as nodes and edges in a directed graph. Similarly, citations lead to another graph, namely a citation graph or network (Garner, 1967). The degree of interconnectedness of a hypertext or similar graph-like entities, such as citation networks, can be expressed using cohesion measures. One of these is compactness as introduced by Botafogo, Rivlin and Shneiderman (Botafogo et al., 1992). As the word 'compactness' has several meanings in mathematics and graph theory we will refer to the compactness notion as introduced by the fore-mentioned authors as BRS-compactness. An exact definition follows later.

BRS-compactness is a measure of cohesion which tries to capture how well-connected a hyperdocument or a network is. As a measure of cohesion its value can be used as a guideline for hypertext authoring systems (Johnson, 1995). It has been studied and discussed in many other works, see e.g. (De Vocht, 1994; Rivlin et al., 1994; Mendes et al., 1998; Fang & Rousseau, 2001; Egghe & Rousseau, 2001; Egghe, 2001). The cohesion of a hypermedia environment influences the retrieval efficiency of users, as was shown e.g. by Khan & Locatis (1998). Leazer & Furner (1999) study compactness in the context of textual identity networks, i.e. a set of documents that share a common semantic or linguistic form. In addition, they compare BRS-compactness with other so-called topological indices such as the Wiener index, stratum and Randić's index (Randić ,1975). The notion of stratum was generalized recently by Leo Egghe in a study on hierarchies (Egghe, 2001).

In informetric studies publications, citations, co-citations (Price, 1965; Garner 1967) as well as collaborations give rise to networks (Pritchard, 1984). A document citation network is not symmetric (if article A cites article B, B normally does not cite A), while a collaboration network definitely is: if author X has collaborated with author Y, automatically author Y has collaborated with X. Similarly, co-citation networks are clearly symmetric. Yet, links in author or country citation networks may or may not be bi-directional. Note that recently also other collaborations, such as actor collaborations have inspired fellow scientists (Barabási & Albert, 1999). These authors and others link their research to the so-called small-world phenomenon (Milgram, 1967; Kochen, 1989; Watts, 1999; Newman & Watts, 1999, Newman, 2000; Björneborn, 2001). Citation links have been inspirational to web search techniques such as those used by the Clever algorithm and by Google (Chakrabarti et al., 1999; Brin and Page, 1998; Henzinger, 2001; Ng et al. 2001). Moreover, the 'hubs' and 'authorities' approach is related to the Pinski-Narin influence weight citation measure (Pinski and Narin, 1976) and mimics the idea of 'highly cited documents' (authorities) and reviews (hubs), see e.g. Kleinberg (1999).

In this article, we study the compactness of a general network with general dissimilarity values between the nodes, and show how weights between nodes can be converted into such dissimilarity values. We present an example of such a calculation using a co-citation network and a network determined by bibliographic coupling. Applications of this measure abound. Indeed, De Bra (2000) observed that when studying the literature of a field large differences in the density of citations may be found. We sometimes see densely connected citation clusters with little or no links to other clusters, while occasionally a field seems rather loosely connected by citation

or collaboration links. De Bra suggests that the BRS-compactness measure can be used to identify research fields with a similar citation behavior. This, in turn, could be a factor in research evaluation exercises, where it is important to compare 'like with like'. For all these reasons we think it is necessary to have a closer look at the notion of BRS-compactness, to study its properties, to construct examples, and to propose appropriate generalizations.

## 2. Some notions from graph theory

A directed graph G, in short: digraph, consists of a set of nodes, denoted as N(G), and a set of links (also called arcs or edges), denoted as L(G). In this text the words 'network' and 'graph' are synonymous. A link e, is an ordered pair (a,b) representing a connection from node a to node b. Node a is called the initial node of link e, and node b is called the final node of the link. If there is an arc connecting nodes a and b, then we say that nodes a and b are directly connected. A path (or chain) from node a to node b is a sequence of distinct links $(a, u_1), (u_1,u_2), \ldots , (u_k,b)$. The length of this path is the number of links (here k+1). Note that, in general, a path from a to b does not necessarily imply a path from b to a. We assume in this paper that there exists at most one direct link between two nodes. Further, nodes will often receive an index number and will be identified through this number.

The distance from node a to node b (in an unweighted graph) is the smallest length of all the paths that join a to b. If such a path does not exist the length is infinity. A strongly connected component of a digraph is a set of nodes such that any two of them are joined by a path. Different strongly connected components in a network consist of disjoint sets of nodes. If a digraph consists of one strongly connected

component it is said to be strongly connected. If no two nodes are connected, the graph is said to be totally disconnected.

In many applications it is natural to assign a number to each arc of a graph. This number is called the weight of this arc, and the graph is then called a weighted graph. Weights can be distance expressed in km or miles (a road map), time needed to travel from point A to point B (a network of airplane connections, or a part of the Internet backbone), costs to move from situation S to situation T, the number of times author a cites author b (citation graph), number of times authors a and b are co-cited (a co-citation graph), etc. Note that weights can be subdivided into two categories: one related to graphs where high weights are considered unfavorable situations, and one related to graphs where high weights are rather favorable. The first category contains the distance, time and costs examples; the second one the (co)-citation examples. In this article we will only be concerned with the second category.

An undirected graph consists of a set of nodes and a set of edges, each of which is an unordered pair of nodes. A collaboration network is an example of such a graph: if author A co-authored an article with author B, then author B co-authored an article with A. Also co-citation networks and networks weighted by bibliographic coupling strengths are examples of undirected (often weighted) graphs.

For more information on graphs we refer the reader to (Gibbons, 1985; Harary, 1969; Trinajstić, 1992; Wilson, 1972).

## 3. Compactness

*3.1 Definition: the BRS network matrix in the unweighted case*

Any (finite) network can be described by a matrix D such that the element on the i-th row and j-th column, denoted as d(i,j), is equal to the length of a shortest path between the i-th and the j-th node of the network. If node j cannot be reached from node i then $d(i,,j) = \infty$. Note that, in general, d is not symmetric ($d(i,j) \neq d(j,i)$), hence d is not a proper metric or distance function. It is a metric in the case of undirected graphs. In their analysis of hypertexts and hyperlinks Botafogo et al. (1992) introduced the following convention: if node j cannot be reached from node i then d(i,,j) is not put equal to $\infty$, but takes as its value the number of nodes in the analyzed network. This representation will be called the BRS representation. Botafogo, Rivlin and Shneiderman (1992) refer to this matrix as the converted distance matrix.

3.2 Weighted networks

Considering weighted graphs means that we allow an arc e = (i,j) to be characterized by a number, called the weight of this arc. In this article we agree that weights are at least equal to one. In the case of a weighted network the new matrix (denoted as $D_w$, and defined further on), playing a similar role as the converted distance matrix, is obtained using a dissimilarity measure, denoted by the symbol d (as in the unweighted case). The number d(i,j) is a non-negative number, denoting the dissimilarity between the nodes i and j (in that order!). Indeed, the dissimilarity measures we are going to use are not necessarily symmetric. A metric, on the other hand, is an example of a (symmetric) dissimilarity measure. The dissimilarity between a node and itself is always put equal to zero, i.e. for all j: d(j,j) = 0. The dissimilarity between two unconnected nodes is put equal to a well-chosen large number (as in

the unweighted case). We will show that in our setting this number can be taken to be N, the number of nodes in the network. Hence, if nodes i and j are unconnected in a weighted network, then d(i,j) = N, the number of nodes in the network.

3.3 Determining a dissimilarity value between connected nodes

In this section we will give one - important - procedure of how to derive dissimilarity values for all nodes if weights between directly connected nodes are known.

Assume we have a weighted digraph, with weights w given between directly connected nodes. Recall that weights mean: number of citations (from author A to author B; or from journal J to journal K); number of collaborations (co-authorships) between authors, links from site S to site T, hyperlinks in a local hypertext document, and so on.

If two nodes i and j are connected through a chain, C, of length t, and with weights $w_k$, k = 1, ...,t, (where t can be one) then we propose as dissimilarity value $d_C(i,j)$ of this connection:

$$d_C(i,j) = \sum_{k=1}^{t} \frac{1}{w_k} \qquad (1)$$

The dissimilarity value between nodes i and j is then put equal to the minimum of the $d_C(i,j)$ values, where C denotes any chain connecting i with j:

$$d(i,j) = \min_C d_C(i,j) \qquad (2)$$

*Properties of this weighting and dissimilarity scheme*

- If all weights in the graph are equal to 1, then the dissimilarity value between any two connected nodes is just the length of a shortest path between these nodes.

- If one weight $w_k$ in a path increases, the corresponding $d_C(i,j)$ decreases.

- If the path increases with one link (leaving the already existing links unchanged) then the dissimilarity value increases.

Although many other schemes to convert weights to dissimilarity values were conceivable, we chose this particular procedure because of its interesting properties. In particular, as one is the smallest value for w, putting all weights equal to one (the associated unweighted network) yields the largest dissimilarity value for connected nodes. If d* denotes the dissimilarity value calculated for the associated unweighted graph, then clearly

$$d(i,j) \leq d^*(i,j) \quad (3)$$

A direct link is very important. Yet, if weights are high enough it is possible that an indirect connection leads to a smaller dissimilarity than the direct link. Examples will be provided in the last section where we study a real-world co-citation and bibliographic coupling network. For a chain with length t the requirement (assuming equal weights in every link) for an indirect link to be more important than a direct one with weight one, is:

$$\sum_{k=1}^{t} \frac{1}{w} = \frac{t}{w} < 1 \quad or \quad w > t \qquad (4)$$

Adopting this procedure gives a maximum dissimilarity value between any two connected nodes of N -1. Hence we will define the dissimilarity value between two unconnected nodes as N.

In the applications studied here only the shortest path is of importance. In other applications it might be more natural to consider all paths (Rousseau, 1989; Egghe, 2001).

*3.4 Definition: BRS-compactness for weighted networks*

We define the general form of the BRS-compactness measure of an N-node network as:

$$C = \frac{MAX - \sum_{i,j=1}^{N} d(i,j)}{MAX - MIN} \qquad (5)$$

where d is a dissimilarity function. MIN and MAX denote the maximum and minimum values for the sum of dissimilarities.

Although we have outlined a specific procedure of deriving dissimilarity values, if weights are given, from now on we only assume the following properties for d:

1°) d is non-negative;

2°) for every two nodes, i and j, $d(i,j) \leq d^*(i,j)$, where $d^*$ denotes the dissimilarity of the associated unweighted network (this is property (2));

Note that when 2°) holds we have the following property:

if nodes i and j are connected then $d(i,j) < N$

As weights can be arbitrarily high, the corresponding dissimilarity values can be arbitrarily small, hence MIN = 0. This means that we actually use the infimum of all possible values, not the minimum, which does not exist. For MAX we use $N(N^2-N)$ as in the unweighted case. This leads to the following definition for the BRS-compactness for weighted graphs.

Definition: BRS-compactness for weighted graphs, denoted as $C_w$:

$$C_w = \frac{(N^2-N)N - \sum_{i,j=1}^{N} d(i,j)}{(N^2-N)N} = 1 - \frac{\sum_{i,j=1}^{N} d(i,j)}{(N^2-N)N} \qquad (6)$$

If N = 1, we leave $C_w$ undefined. Hence, we will always consider networks consisting of at least two nodes.

A totally disconnected, weighted graph has BRS-compactness $C_w$ equal to 0. We note that any unweighted graph may be considered a weighted one, with all weights equal to one, and with all dissimilarities of connected nodes equal to one. Conversely, every weighted graph has an associated unweighted graph (put all weights equal to 1). The BRS-compactness value of the associated unweighted

graph can be calculated using formula (6) for weighted graphs (this value will be denoted as $C_w^*$), as well as by the formula for unweighted graphs (then denoted as $C^{**}$) (Egghe-Rousseau, 2001). These formulae only differ in the value of the denominator. As the denominator for the weighted case is larger than that for the unweighted case (namely: $N^3$-$N^2$ versus $N^3$-$2N^2$ + N) the weighted BRS-compactness value of a graph with all weights equal to 1 is always smaller than the standard BRS-compactness value of this same graph. More precisely we have:

$$C_w^* = \frac{N-1}{N} C^{**} \qquad (7)$$

We further note that by (2)

$$C_w^* \leq C_w \qquad (8)$$

but clearly,

$$\lim_{N\to\infty} C_w^* = \lim_{N\to\infty} C^{**}$$

One element in formula (6) deserves special attention and will be defined as the generalized Wiener index.

Definition: the generalized Wiener index of a weighted graph

We define the generalized Wiener index of a weighted digraph, denoted by $W_w$, as the sum of all elements of the dissimilarity matrix. In the case of an undirected, strongly connected unweighted graph this sum divided by two is known as the Wiener index, after the chemist Harold Wiener (Wiener, 1947).

Introducing this notation in formula (6) yields:

$$C_w = 1 - \frac{W_w}{N^2(N-1)} \qquad (9)$$

or

$$W_w = N^2(N-1)(1-C_w) \qquad (10)$$

Definition: connection coefficient (Egghe-Rousseau, 2001)

Let now $\beta$, $\beta \in [0,1]$, be the fraction of all pairs (i,j) (with $i \neq j$) that are connected and let $A_\beta$ denote the set of those pairs (i,j) for which this happens, i.e. for which $d(i,j) < N$. The fraction $\beta$ is called the *connection coefficient* of the network. Following Pritchard (1984) we may say that in a communication network a high value of the connection coefficient improves the level of accessibility between nodes, and hence the transfer of information. The connection coefficient is either zero (and then $C_w = 0$) or it satisfies the following inequality:

$$\frac{1}{N(N-1)} \leq \beta \leq 1 \qquad (11)$$

The connection coefficient of a weighted graph is clearly equal to that of its associated unweighted one. A connection coefficient $\beta$ equal to one simply means that every two nodes have a dissimilarity value strictly smaller than N in the matrix $D_W$ (this implies that the graph is strongly connected).

## 4. Decomposition of the weighted BRS-compactness measure $C_w$

Consider a network with N nodes. Let $\beta$ be the connection coefficient. This implies that $(1-\beta)(N^2 - N)$ pairs of nodes are unconnected. Their dissimilarities are put equal to N. Let $A_\beta$ denote the set of all pairs of nodes that are connected:

$$A_\beta=\{(i,j)\,; i\neq j,\ i,j=1,\ldots,N, \text{ i and j are connected}\}$$

Then:

$$C_w = \frac{(N^2-N)N-(1-\beta)(N^2-N)N - \sum_{(i,j)\in A_\beta} d(i,j)}{(N^2-N)N}$$

or:

$$C_w = \beta - \frac{\sum_{(i,j)\in A_\beta} d(i,j)}{N^3-N^2} \qquad (12)$$

Consequently, the weighted BRS-compactness value consists of the connection coefficient minus a term involving the sum of the dissimilarity values of connected nodes.

From now on we will abbreviate the expression $\sum_{(i,j)\in A_\beta} d(i,j)$ by $\Sigma$. Note that $\Sigma$ is the generalized Wiener coefficient ($W_w$) only if the network is totally connected.

If $\Sigma^*$ denotes the value of $\Sigma$ for the associated unweighted network, then clearly (by our requirements on d) $\Sigma \leq \Sigma^*$.

Note. For an unweighted graph we have the following property (Egghe & Rousseau, 2001)

$$C \in \left[0,\ \beta \frac{N-1}{N}\right] \qquad (13)$$

This upper bound is not valid anymore, and neither is the inequality $\beta$ ($N^2$-N) $\leq \Sigma$. Indeed, consider a complete graph (every two nodes are connected, hence $\beta$ = 1) where every arc has the same weight w > 1. Putting d(i,j) = 1/w, makes $\Sigma$ equal to ($N^2$-N)/w, which is clearly smaller than ($N^2$-N), being here equal to $\beta$ ($N^2$-N). Further, $C_w$ is here equal to $1-\frac{N(N-1)}{w\,N^2(N-1)} = 1-\frac{1}{wN} > 1\bullet\frac{N-1}{N} = 1-\frac{1}{N}$ contradicting expression (13).

The following theorem gives a relation between the weighted compactness $C_w$ and the connection coefficient $\beta$. More precisely, it provides an upper bound for $\beta$ as a function of $C_w$.

Theorem

$$\beta \leq \frac{3N}{N+1} C_w \qquad (14)$$

Proof. We know already that

$$C_w^* = \frac{N-1}{N} C^{**} \qquad (7)$$

where $C_w^*$ denotes the BRS-compactness value of the associated unweighted graph, calculated using the formula for weighted graphs, and $C^{**}$ denotes the standard BRS-compactness value for the associated unweighted graph.

We also know that for unweighted digraphs (Egghe & Rousseau, 2001):

$$\beta \leq \frac{3(N-1)}{N+1} C^{**},$$

$$\text{hence}: \beta \leq \frac{3(N-1)}{N+1} \frac{N}{N-1} C_w^{\cdot} = \frac{3N}{N+1} C_w^{\cdot} \leq \frac{3N}{N+1} C_w \qquad (15)$$

where we have used inequality (8). This proves the theorem.

Note: As $\Sigma \leq \Sigma^*$ and as, for every unweighted network $\Sigma^* \leq \beta \frac{(N-1)N(2N-1)}{3}$ (Egghe & Rousseau, 2001), we also have:

$$\Sigma \leq \beta \frac{(N-1)N(2N-1)}{3} \qquad (16)$$

We next prove two results generalizing the corresponding results for unweighted graphs (and the corresponding BRS-compactness value). They show that also the weighted BRS-compactness measure behaves as expected.

Proposition

If we add one unconnected point to a weighted network then the weighted BRS-compactness value decreases.

Proof. If β denotes the connection coefficient before this unconnected point is added, and β' denotes the connection coefficient after adding one unconnected point, then

$$\beta' = \beta \frac{N-1}{N+1}$$

The weighted BRS-compactness of the new graph is:

$$C_w^{\cdot} = \beta' - \frac{\Sigma}{N(N+1)^2} = \beta \frac{N-1}{N+1} - \frac{\Sigma}{N(N+1)^2}$$

By definition (12) we see that

$$\Sigma = N^2(N-1)(\beta - C_w)$$

and hence:

$$\begin{aligned} C_w' &= \beta\frac{N-1}{N+1} - \frac{N^2(N-1)(\beta - C_w)}{N(N+1)^2} \\ &= \beta\left(\frac{N-1}{N+1} - \frac{N(N-1)}{(N+1)^2}\right) + \frac{N(N-1)}{(N+1)^2}C_w \\ &= \beta\frac{N-1}{(N+1)^2} + C_w\frac{N(N-1)}{(N+1)^2} \end{aligned}$$

From this expression we see that $C_w' < C_w$ if and only if

$$\beta\frac{N-1}{(N+1)^2} < C_w\left(\frac{(N+1)^2 - N(N-1)}{(N+1)^2}\right) = C_w\frac{3N+1}{(N+1)^2}$$

$$\Leftrightarrow$$

$$\beta < C_w\frac{3N+1}{N-1}$$

By the previous theorem this inequality is always satisfied.

What happens if we add a node that is 'strongly' connected to an existing network? We first observe that if the original network is totally disconnected, the addition of a new node that is at least connected to one node of the original one makes the compactness value strictly positive, and hence strictly larger than that of the original one. We now assume that the original network is not totally disconnected.

Proposition

If we add to an N-node network, a new node, whose dissimilarity to all other nodes is smaller than or equal to $\min_{i,j}\{d(i,j)\}, i \neq j$, then the weighted BRS-compactness value of this new network, denoted as $C_w'$, is strictly larger than $C_w$, the weighted BRS-compactness value of the original network.

Proof.

If we denote $\min_{i,j}\{d(i,j)\}, i \neq j$, simply by m, then we have to show that

$$C_w = 1 - \frac{W_w}{N^2(N-1)} \leq 1 - \frac{W_w + 2Nm}{(N+1)^2 N} \leq C_w'$$

This is equivalent with:

$$\frac{W_w}{N(N-1)} \geq \frac{W_w + 2Nm}{(N+1)^2}$$

*or*

$$W_w\left(\frac{3N+1}{N(N-1)(N+1)^2}\right) \geq \frac{2Nm}{(N+1)^2}$$

Because $W_w \geq m\,(N^2\text{-}N)$, this expression is satisfied if:

$$(3N+1)(N^2-N) \geq 2N^2(N-1)$$

*or*

$$3N+1 \geq 2N$$

which is always true. This proves this result.

*The compactness value of the disjoint union of two networks with known compactness.*

We assume that the first network contains M nodes, has connection coefficient $\beta_1$ and compactness value $C_1$. The second one contains N nodes, has connection coefficient $\beta_2$ and compactness value $C_2$. Hence the disconnected union has M+N nodes. The connection coefficient $\beta$ is:

$$\beta = \frac{\beta_1(M^2 - M) + \beta_2(N^2 - N)}{(M+N)^2 - (M+N)} \qquad (17)$$

and $\Sigma = \Sigma_1 + \Sigma_2$ (as the two parts are unconnected).

Now,

$$C_1 = \beta_1 - \frac{\Sigma_1}{M^2(M-1)}$$

$$C_2 = \beta_2 - \frac{\Sigma_2}{N^2(N-1)}$$

$$C = \beta - \frac{\Sigma}{(M+N)^2(M+N-1)}$$

Hence

$$C = \left(\beta_1 - \frac{\Sigma_1}{M^2(M-1)}\right) * \frac{M^2(M-1)}{(M+N)^2(M+N-1)}$$

$$+ \left(\beta_2 - \frac{\Sigma_1}{N^2(N-1)}\right) * \frac{N^2(N-1)}{(M+N)^2(M+N-1)} + Z$$

$$= C_1 \frac{M^2(M-1)}{(M+N)^2(M+N-1)} + C_2 \frac{N^2(N-1)}{(M+N)^2(M+N-1)} + Z$$

with

$$Z = \beta_1 \left( \frac{M^2 - M}{(M+N)(M+N-1)} - \frac{M^2(M-1)}{(M+N)^2(M+N-1)} \right)$$

$$+ \beta_2 \left( \frac{N^2 - N}{(M+N)(M+N-1)} - \frac{N^2(N-1)}{(M+N)^2(M+N-1)} \right)$$

$$= \beta_1 \frac{MN(M-1)}{(M+N)^2(M+N-1)} + \beta_2 \frac{MN(N-1)}{(M+N)^2(M+N-1)}$$

Hence:

$$C = C_1 \frac{M^2(M-1)}{(M+N)^2(M+N-1)} + C_2 \frac{N^2(N-1)}{(M+N)^2(M+N-1)}$$
$$+ \beta_1 \frac{MN(M-1)}{(M+N)^2(M+N-1)} + \beta_2 \frac{MN(N-1)}{(M+N)^2(M+N-1)}$$

Corollary 1

If M = N, $\beta_1 = \beta_2 = b$ and $C_1 = C_2 = c$, then the compactness, C, of the weighted graph consisting of these two unconnected graphs (e.g. two copies of the same graph) is equal to:

$$C = \frac{(N-1)(b+c)}{2(2N-1)} \qquad (18)$$

Corollary 2

The BRS-compactness value of a disjoint copy of two identical weighted networks is strictly smaller than the BRS-compactness value of one component. Using the notation of Corollary 1 this means:

$$C < c \qquad (19)$$

Proof.

By Corollary 1 we know that

$$C = \frac{(N-1)(b+c)}{2(2N-1)}$$

By Formula (15), this leads to:

$$C < \frac{N-1}{2(2N-1)} \left( c + \frac{3N}{N+1} c \right)$$
$$< \frac{1}{4} . (4c) = c$$

This result is, of course, not surprising, but it shows that the BRS-compactness value for weighted graphs has good properties.

## 5. An example of the calculation of the weighted compactness value using the procedure outlined in this paper

We consider the undirected network of co-cited papers in particle physics studied by Small (1973) to introduce the notion of co-citation. We will also consider the undirected network of bibliographic coupling strengths between these papers. Data are given in Small's original publication but presented here for the reader's convenience. Small's article also contains the corresponding co-citation network (but co-citation links with a weight smaller than 7 are not shown).

Table 1 Small's particle physics co-citation matrix

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | 1 | 9 | 7 | 13 | 5 | 2 | 2 | 2 | 11 |
| B | 1 | | 6 | 9 | 11 | 10 | 1 | 1 | 9 | 1 |
| C | 9 | 6 | | 19 | 32 | 18 | 3 | 0 | 6 | 7 |
| D | 7 | 9 | 19 | | 20 | 15 | 4 | 0 | 9 | 6 |
| E | 13 | 11 | 32 | 20 | | 50 | 7 | 7 | 19 | 18 |
| F | 5 | 10 | 18 | 15 | 50 | | 1 | 0 | 10 | 7 |
| G | 2 | 1 | 3 | 4 | 7 | 1 | | 68 | 21 | 1 |
| H | 2 | 1 | 0 | 0 | 7 | 0 | 68 | | 17 | 1 |
| I | 2 | 9 | 6 | 9 | 19 | 10 | 21 | 17 | | 0 |
| J | 11 | 1 | 7 | 6 | 18 | 7 | 1 | 1 | 0 | |

where letters denote the following articles.

**A**. Bjorken, J.D. (1966) Applications of the chiral U(6) ⊗ U(6) algebra of densities. *Phys. Rev.* 148, 1467

**B**. Gasiorowicz, S. and Geffen, D.A. (1969). Effective Lagrangians and field algebras with chiral symmetry. *Rev. Mod. Phys.* 41, 531

**C**. Gell-Mann, M. (1962). Symmetries of baryons and mesons. *Phys. Rev.* 125, 1067

**D.** Gell-Mann, M. (1964). The symmetry group of vector and axial vector currents. *Physics* 1, 63

**E**. Gell-Mann, M., Oakes, R.J. and Renner, B. (1968). Behavior of current divergences under $SU_3$ x $SU_3$. *Phys. Rev.* 175, 2195

**F**. Glashow, S.L. and Weinberg S. (1968). Breaking chiral symmetry. *Phys. Rev. Lett.* 20, 224

**G**. Lovelace, C. (1968). A novel application of Regge trajectories. *Phys. Letters B*, 28, 264

**H.** Veneziano, G. (1968). Construction of a crossing-symmetric Regge-behaved amplitude for linearly rising trajectories. *Nuovo Cimento* 57, 190

**I**. Weinberg, S. (1966). Pion scattering lengths. *Phys. Rev. Lett.* 17, 616

**J.** Wilson, K.G. (1969). Non-Lagrangian models of current algebra. *Phys. Rev.* 179, 1499

Next, dissimilarity values were calculated following the procedure outlined in this article. The results are shown in Table 2. Table 3 shows whether the dissimilarity value was obtained from a direct link (DIR) or not. In the latter case Table 3 shows the node (or nodes) used in the calculation. Observe that, if there is no direct link (as e.g. for C and H) then the dissimilarity value must be calculated using a connecting node. However, it also happens that an indirect link yields a smaller dissimilarity value than the direct link. This is, for instance, the case for A and B, which yields the smallest dissimilarity value if they are connected via E.

Table 2 Dissimilarity values of Table 1

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 0.1678 | 0.1082 | 0.1269 | 0.0769 | 0.0969 | 0.1772 | 0.1884 | 0.1296 | 0.0909 |
| B | 0.1678 | 0 | 0.1222 | 0.1111 | 0.0909 | 0.1000 | 0.1587 | 0.1699 | 0.1111 | 0.1465 |
| C | 0.1082 | 0.1222 | 0 | 0.0526 | 0.0313 | 0.0513 | 0.1315 | 0.1427 | 0.0839 | 0.0868 |
| D | 0.1269 | 0.1111 | 0.0526 | 0 | 0.05 | 0.0667 | 0.1503 | 0.1615 | 0.1026 | 0.1056 |
| E | 0.0769 | 0.0909 | 0.0313 | 0.05 | 0 | 0.02 | 0.1003 | 0.1115 | 0.0526 | 0.0556 |
| F | 0.0969 | 0.1000 | 0.0513 | 0.0667 | 0.02 | 0 | 0.1203 | 0.1315 | 0.0726 | 0.0756 |
| G | 0.1772 | 0.1587 | 0.1315 | 0.1503 | 0.1003 | 0.1203 | 0 | 0.0147 | 0.0476 | 0.1558 |
| H | 0.1884 | 0.1699 | 0.1427 | 0.1615 | 0.1115 | 0.11315 | 0.0147 | 0 | 0.0588 | 0.1670 |
| I | 0.1296 | 0.1111 | 0.0839 | 0.1026 | 0.0526 | 0.0726 | 0.0476 | 0.0588 | 0 | 0.1082 |
| J | 0.0909 | 0.1465 | 0.0868 | 0.1056 | 0.0556 | 0.0756 | 0.1984 | 0.1670 | 0.1082 | 0 |

Table 3 Shortest paths in the particle physics co-citation network

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | E | E | E | DIR | E | E-I | E-I | E | DIR |
| B | E | | E | DIR | DIR | DIR | I | I | DIR | E |
| C | E | E | | DIR | DIR | E | E-I | E-I | E | E |
| D | E | DIR | DIR | | DIR | DIR | E-I | E-I | E | E |
| E | DIR | DIR | DIR | DIR | | DIR | I | I | DIR | DIR |
| F | E | DIR | E | DIR | DIR | | E-I | E-I | E | E |
| G | E-I | I | E-I | E-I | I | E-I | | DIR | DIR | E-I |
| H | E-I | I | E-I | E-I | I | E-I | DIR | | DIR | E_I |
| I | E | DIR | E | E | DIR | E | DIR | DIR | | E |
| J | DIR | E | E | E | DIR | E | E-I | E-I | E | |

The generalized Wiener index of this weighted co-citation network is 9.388. As N = 10, and the connection coefficient is 1, the weighted compactness value is 0.9896.

For the associated unweighted network we find a generalized Wiener index of 98, and hence a weighted compactness value of 0.891. This compactness is smaller than that for the weighted network (as it should).

Similarly, Tables 4, 5 and 6 show the bibliographic coupling strengths for the nodes of the bibliographic coupling network (see Fig.1 and Table 4), the dissimilarity values (Table 5) and the connecting nodes used to calculate these dissimilarity values

(Table 6). Note that here we sometimes need four connecting nodes for a path leading to a smallest dissimilarity.

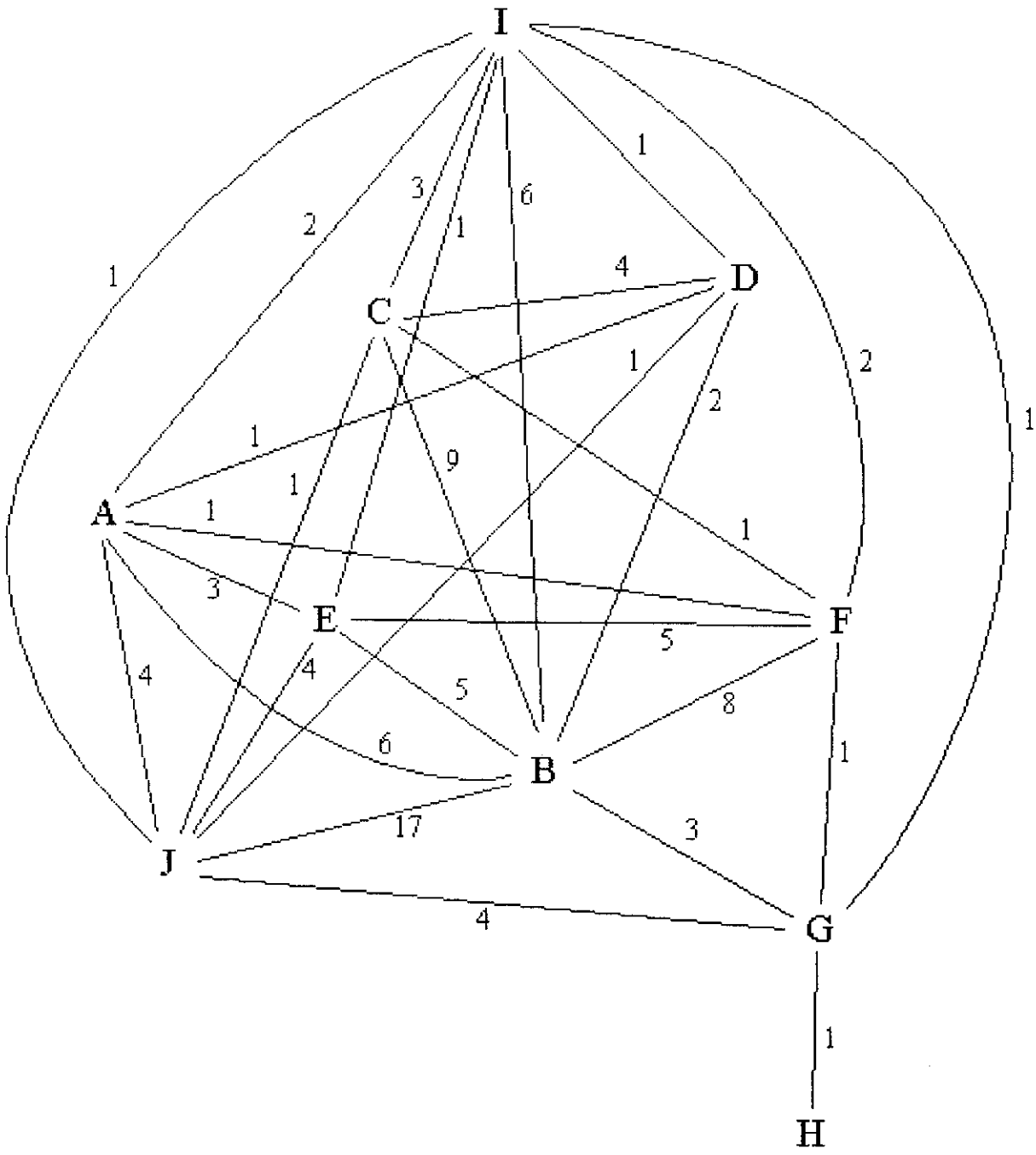


Fig.1 The weighted bibliographic coupling values of Small's particle physics network

Table 4. Network of bibliographic coupling strengths of Small's particle physics network

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | 6 | 0 | 1 | 3 | 1 | 0 | 0 | 2 | 4 |
| B | 6 | | 9 | 2 | 5 | 8 | 3 | 0 | 6 | 17 |
| C | 0 | 9 | | 4 | 0 | 1 | 0 | 0 | 3 | 1 |
| D | 1 | 2 | 4 | | 0 | 0 | 0 | 0 | 1 | 1 |
| E | 3 | 5 | 0 | 0 | | 5 | 0 | 0 | 1 | 4 |
| F | 1 | 8 | 1 | 0 | 5 | | 1 | 0 | 2 | 0 |
| G | 0 | 3 | 0 | 0 | 0 | 1 | | 1 | 1 | 4 |
| H | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | 0 | 0 |
| I | 2 | 6 | 3 | 1 | 1 | 2 | 1 | 0 | | 1 |
| J | 4 | 17 | 1 | 1 | 4 | 0 | 4 | 0 | 1 | |

Table 5 Dissimilarity values of Table 4

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 0.1667 | 0.2778 | 0.5278 | 0.3333 | 0.2917 | 0.4755 | 1.4755 | 0.3333 | 0.2255 |
| B | 0.1667 | 0 | 0.1111 | 0.3611 | 0.2000 | 0.125 | 0.3088 | 1.3088 | 0.1667 | 0.0588 |
| C | 0.2778 | 0.1111 | 0 | 0.2500 | 0.3111 | 0.2361 | 0.4199 | 1.4199 | 0.2778 | 0.1699 |
| D | 0.5278 | 0.3611 | 0.2500 | 0 | 0.5611 | 0.4861 | 0.6699 | 1.6699 | 0.5278 | 0.4199 |
| E | 0.3333 | 0.2000 | 0.3111 | 0.5611 | 0 | 0.2000 | 0.5000 | 1.5000 | 0.3667 | 0.2500 |
| F | 0.2917 | 0.1250 | 0.2361 | 0.4861 | 0.2000 | 0 | 0.4338 | 1.4338 | 0.2917 | 0.1838 |
| G | 0.4755 | 0.3088 | 0.4199 | 0.6699 | 0.5000 | 0.4338 | 0 | 1 | 0.4755 | 0.2500 |
| H | 1.4755 | 1.3088 | 1.4199 | 1.6699 | 1.5000 | 1.4338 | 1 | 0 | 1.4755 | 1.2500 |
| I | 0.3333 | 0.1667 | 0.2778 | 0.5278 | 0.3667 | 0.2917 | 0.4755 | 1.4755 | 0 | 0.2255 |
| J | 0.2555 | 0.0588 | 0.1699 | 0.4199 | 0.2500 | 0.1838 | 0.2500 | 1.25 | 0.2255 | 0 |

Table 6 Shortest paths in the bibliographic coupling network

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | DIR | B | B-C | DIR | B | B-J | B-J-G | B | B |
| B | DIR | | DIR | C | DIR | DIR | J | J-G | DIR | DIR |
| C | B | DIR | | DIR | B | B | B-J | B-J-G | B | B |
| D | B-C | C | DIR | | C-B | C-B | C-B -J | C-B - J-G | C-B | C-B |
| E | DIR | DIR | B | C-B | | DIR | J | J-G | B | DIR |
| F | B | DIR | B | C-B | DIR | | B-J | B-J-G | B | B |
| G | B-J | J | B-J | C-B-J | J | B-J | | DIR | J-B | DIR |
| H | B-J-G | J-G | B-J-G | C-B-J-G | J-G | B-J-G | DIR | | G-J-B | G |
| I | B | DIR | B | C-B | B | B | J-B | G-J-B | | B |
| J | B | DIR | B | C-B | DIR | B | DIR | G | B | |

The generalized Wiener index of this weighted bibliographic coupling network is 48.0362. As N = 10, and the connection coefficient is 1, the weighted compactness value is 0.9466.

For the associated unweighted network we find a generalized Wiener index of 132, and hence a weighted compactness value of 0.8533.

It is intuitively clear that the co-citation network is more compact than the bibliographic coupling network. The obtained compactness values reflect this fact. On the other hand, both networks are completely connected and are densely connected (in the intuitive sense of the word). This fact is reflected by the high compactness values we obtained.

## 6. Conclusion

The new measure for the cohesion of weighted networks introduced here generalizes the well-known BRS-compactness measure. It can be used in the case a dissimilarity value (possibly derived from weights) is given between nodes in a network or a graph. A procedure is outlined illustrating how a set of weights between connected nodes can be transformed into a set of dissimilarity measures for all nodes. The new measure for the cohesion of weighted graphs has several desirable properties, related to adding a new node or to the disjoint union of two networks. An example is presented, involving Small's co-citation and bibliographic coupling network, of the calculation of the new compactness measure.

We suggest to test, using the new measure, De Bra's suggestion about research fields with a similar structure. It would also be interesting to test whether the new

measure can distinguish between (real-world) random and so-called small-world networks.

**References**


Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science, 286,* 509-512.

Björneborn, L. (2001). Small-world linkage and co-linkage. Proceedings of the 12th ACM Conference on Hypertext (Aarhus, Denmark).

Botafogo, R.A., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems, 10*, 142-180.

Brin S. and Page L. (1998). Anatomy of a large-scale hypertextual web-search engine. *Proceedings of the 7th International World Wide Web Conference* (Brisbane, Australia, April 14-18), (p. 107-117).

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S.R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Hypersearching the Web. *Scientific American 280(6),* 54-60.

De Bra, P. (2000). Using hypertext metrics to measure research output levels, *Scientometrics, 47*, 227-236.

De Vocht, J. (1994). *Experiments for the characterization of hypertext structures.* Masters Thesis, Eindhoven University of Technology.

Egghe, L. (2001) Development of hierarchy theory for digraphs using concentration theory based on a new type of Lorenz curve. *Mathematical and Computer Modeling* (to appear).

Egghe, L. and Rousseau, R. (2001). BRS-compactness in networks: theoretical considerations related to cohesion in citation graphs, collaboration networks and the Internet. *Mathematical and Computer Modeling* (to appear).

Fang, Y.,and Rousseau, R. (2001). Lattices in citation networks: an investigation into the structure of citation graphs. *Scientometrics* 50, 273-287.

Garner, R. (1967). A computer oriented, graph theoretic analysis of Citation Index structures. Drexel University Press, Philadelphia.

Gibbons, A. (1985). *Algorithmic graph theory*. Cambridge (UK): Cambridge University Press.

Harary, F. (1969). *Graph theory*. Reading(MA): Addison-Wesley.

Henzinger, M.R. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing, 5(1),* 45-50.

Johnson, S. (1995). Control for hypertext construction. *Communications of the ACM. 38(8),* p.87.

Khan, K., and Locatis, C. (1998). Searching through cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web. *Journal of the American Society for Information Science, 49*, 176-182.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46 (5),* 604-632.

Kochen, M. (1989). *The Small world*. Norwood (NJ): Ablex.

Leazer, G.H. and Furner, J. (1999).Topological indices of textual identity networks. In: Woods, L. (Ed.), *Knowledge: creation, organization and use: Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, Medford (NJ): Information Today, 345-358.

Mendes, E., Hall, W. and Harrison, R. (1998). Applying metrics to the evaluation of