

BRS-compactness in networks: Theoretical considerations related to cohesion in citation graphs, collaboration networks and the internet

Non Peer-reviewed author version

EGGHE, Leo & ROUSSEAU, Ronald (2003) BRS-compactness in networks: Theoretical considerations related to cohesion in citation graphs, collaboration networks and the internet. In: *Mathematical and Computer Modelling*, 37(7-8). p. 879-899.

DOI: 10.1016/S0895-7177(03)00091-8

Handle: <http://hdl.handle.net/1942/774>

**BRS-compactness in networks: theoretical considerations related to
cohesion in citation graphs, collaboration networks and the Internet**

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

Compactness as introduced by Botafogo, Rivlin and Shneiderman, in short: BRS-compactness, is studied in general, as it can be used to describe the cohesion of parts of the Internet or collaboration networks, and in the particular case of a unidirectional network, such as a citation graph. It is shown that the connection coefficient is an upper bound for the BRS-compactness value of a network. During our investigations we derive an upper bound for the generalized Wiener index of a directed graph. Several networks are constructed and their BRS-compactness values are calculated.

Keywords: BRS-compactness, networks, hyperlinks, Internet, citation networks, collaboration graphs, generalized Wiener index, sum of distances in a graph

1. Introduction

The pages and hyperlinks of the World Wide Web may be viewed as nodes and edges in a directed graph (Kleinberg et al., 1999; Broder et al., 2000). The degree of the interconnectedness of a hypertext or similar graph-like entities 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 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, 1994; Salton et al., 1994, Calvi & De Bra, 1997; Mendes et al., 1998). Indeed, the density and cohesion of links in a hypermedia environment influences the retrieval efficiency of users (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. They, moreover, compare BRS-compactness with other so-called topological indices such as the Wiener index, stratum and Randić's index (Randić, 1975).

In informetric studies publications, citations, co-citations (Price, 1965; Shepherd et al., 1990) as well as collaborations give rise to networks (Pritchard, 1984; Ding et al., 1998; Kretschmer, 1999). A citation network is clearly not symmetric (if article A cites

article B, then B normally does not cite A), while a collaboration network definitely is: if author X collaborates with author Y, then automatically author Y has collaborated with X. Note that recently also other collaborations, such as actor collaborations have inspired fellow scientists (Barabási & Albert, 1999). 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). Moreover, the 'hubs' and 'authorities' approach is related to the Pinski-Narin influence weight citation measure (1976) and mimics the idea of 'highly cited documents (authorities) and reviews (hubs). The exact relation between the older citation-based measures, such as the Pinski-Narin weights, including Geller's modification (Geller, 1978), and the newer hypertext and WWW-based approach is clearly described by Kleinberg (1999).

In this article, we study the compactness of a general network and show how this web metric may be used in citation analysis and the study of collaboration networks. Indeed, De Bra (2000) observed that when studying the literature of a field large differences in the density of citations may be found. Sometimes we see densely connected citation clusters with little or no links to other clusters. 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. For all these reasons we think it is necessary to have a closer look at the notion of BRS-compactness, to study its properties and to construct some more examples, besides those given by Botafogo, Rivlin and Shneiderman (1992) and De Vocht (1994).

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 , $a = \text{init}(e)$, and node b is called the final node of the link: $b = \text{fin}(e)$. The out-degree of a node b is the number of arcs leading out from it, i.e. the number of arcs e such that $\text{init}(e) = b$. Similarly the in-degree of a node, b , is the number of arcs e such that $\text{fin}(e) = b$ (Knuth, 1969, p.371). A path from node a to node b is a sequence of distinct links $(a, u_1), (u_1, u_2), \dots, (u_k, b)$. 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 . A cycle is a path of length > 1 , beginning and ending in the same node. A graph that does not contain any cycle is called an acyclic graph. In this paper we will always assume that edges are unweighted, or, equivalently, have a weight equal to one. 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.

Two graphs G and H are isomorphic if there exists a bijection f from G to H such that if $h_1 = f(g_1)$ and $h_2 = f(g_2)$, with $g_i \in N(G)$ and $h_i \in N(H)$, $i = 1, 2$, and if there exists a link in G between g_1 and g_2 , then there exists a corresponding link in H between $h_1 = f(g_1)$ and $h_2 = f(g_2)$ (in that order), and vice versa (Chen, 1971; Wilson, 1972).

A unidirectional graph is a graph in which a link between nodes a and b , implies that

there is not a (direct) link from b to a. In a unidirectional graph cycles may exist, but the smallest possible length is 3. If there are nodes a and b such that the whole graph consists of exactly one path of length $N-1$ from a to b we will refer to such a linear graph as a unidirectional N -chain. We will say that a graph T is a tree if it is unidirectional, acyclic and there exists exactly one point, called the root, from which each other point can be reached. The distance from the root to a node t in a tree is called the depth of t . If each node in the tree has the same number of children (at least those which have children), this number is called the branching factor of the tree. Nodes without children are called terminal nodes or leaves. The length of a longest path from the root to a leaf is called the tree-depth. A tree is balanced if at the same depth all nodes have the same number of children. Hence, in a balanced tree no leaf is further away from the root than any other leaf.

If the existence of a link between nodes a and b necessarily implies the existence of a link from b to a we say that this network is a bi-directional graph. If a bi-directional graph consists of exactly one path of length $N - 1$ then we will refer to such a graph as a bi-directional N -chain. Fig. 1 a unidirectional N -chain, a bi-directional N -chain and a unidirectional N -loop.

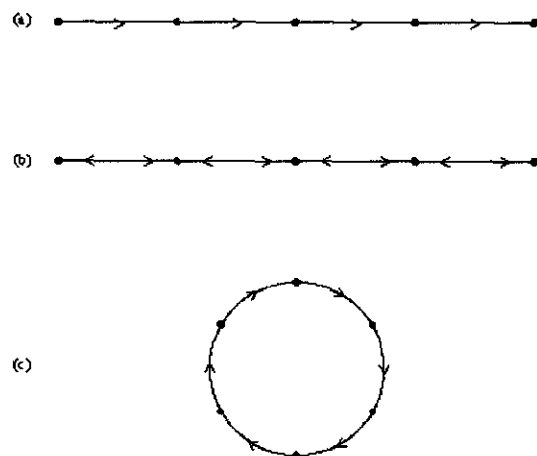


Fig 1 (a) Unidirectional chain; (b) bi-directional chain; (c) loop

The distance from node a to node b 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.

An undirected graph consists of a set of nodes and a set of edges, each of which is an unordered pair of nodes. Any bi-directional digraph can be considered as an undirected graph. 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 . Hence most of the results obtained in this paper can be applied to collaboration networks as studied e.g. in (Newman, 2001).

When applying our ideas to citation networks (a network where nodes are articles and a link from article a to article b means that article a refers to article b) we always assume that these are unidirectional, although this is in reality not always the case (due, e.g. to the existence of invisible colleges). All citation networks considered in this paper are moreover assumed to be acyclic.

For more information on graphs we refer the reader to (Knuth, 1969; Chen, 1971; Berge, 1967; Gibbons, 1985; Harary, 1969; Trinajstić, 1992).

3. Compactness

3.1 Definition The BRS network matrix (Botafogo et al., 1992)

Any (finite) network can be described by a matrix D such that its element on the i -th row and j -th column, denoted as $d(i,j)$, is equal to the shortest distance between the i -th and the j -th node of the network. If node j cannot be reached from node i then $d(i,j) = \infty$. 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 ∞ , but takes as its value the number of nodes in the analyzed network, see also (Leazer & Furner, 1999). This representation will be called the BRS representation and the associated matrix is denoted as D_B . Botafogo, Rivlin and Shneiderman (1992) refer to this matrix as the converted distance matrix. We define the generalized Wiener index of a general digraph, denoted by W , as the sum of all elements of the converted distance matrix. In the case of an undirected, strongly connected graph this sum divided by two is known as the Wiener index, after the chemist Harold Wiener (Wiener, 1947).

3.2 Definition: BRS-compactness

The BRS-compactness value, C , of a network consisting of $N \geq 2$ nodes, is calculated using a formula having the following general structure:

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

where $d(i,j)$ denotes an element of the network matrix under study while MAX and MIN denote the maximum and the minimum sum for the corresponding N -node network (Botafogo et al., 1992). We see that compactness is the normalized,

generalized Wiener index. If $N = 1$ (a network consisting of just one node), C is not defined.

3.3 The compactness formula for a general digraph and the connection coefficient

In the BRS-representation two unconnected nodes are attributed a distance value equal to N . There seems, however, to be no a priori reason why the value N must be used. Hence, we will just assume that this value is a function of the number of nodes in the citation network. This value is denoted as $\varphi(N)$ (Botafogo, Rivlin and Shneiderman denote this value by K). We will certainly put $\varphi(N) \geq N$, otherwise unconnected pairs could have a smaller distance than connected ones. This agreement leads to the following compactness formula for a general network.

The general BRS-compactness formula (Botafogo et al., 1992):

$$C = \frac{(N^2 - N)\varphi(N) - \sum_{i,j=1}^N d(i,j)}{(N^2 - N)(\varphi(N) - 1)} \quad (2)$$

MAX is here obtained in the case that no two pairs are connected. This gives $N^2 - N$ times the largest value, namely $\varphi(N)$. MIN is obtained when every two pairs of different nodes are connected. This gives a value of $N^2 - N$ multiplied by 1.

Definition: connection coefficient

Let now β , $\beta \in [0, 1]$, be the fraction of all pairs (i,j) (with $i \neq j$) that are connected and let A_β denote the set of those pairs (i,j) for which this happens, i.e. for which $d(i,j) < \varphi(N)$. The fraction β will be called the *connection coefficient* of the network. The

connection coefficient is either zero (and then $C = 0$) or it satisfies the following inequality:

$$\frac{1}{N(N-1)} \leq \beta \leq 1 \quad (3)$$

If the compactness value C is one (every two nodes have distance 1) then β is one too. The converse is not true: $\beta = 1$ simply means that every two nodes have a finite distance in the matrix D_B (this means that the graph is strongly connected). For a unidirectional chain $\beta = 1/2$, while for a bi-directional chain, and for a unidirectional loop the β -value is 1.

If a network has N nodes then, a priori, the largest possible distance between two connected nodes is $N - 1$. If, however, we know that its connection coefficient is β then the largest (possible) distance between two connected nodes is $L_\beta = \min(N - 1, \beta N(N - 1))$. 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.

3.4 A decomposition of the compactness measure

Using the connection coefficient the BRS-compactness formula can be rewritten as:

$$C = \frac{(N^2 - N)\phi(N) - (1 - \beta)(N^2 - N)\phi(N) - \sum_{(i,j) \in A_\beta} d(i,j)}{(N^2 - N)(\phi(N) - 1)} \quad (4)$$

This leads to the following decomposition of (2) in two parts. The first is determined by the upper limit for a network with a connection coefficient β ; the second part reduces this value further depending on the degree of connectedness.

$$C = \frac{\beta \phi(N)}{\phi(N)-1} - \frac{\sum_{(i,j) \in A_\beta} d(i,j)}{(N^2-N)(\phi(N)-1)} \quad (5)$$

Because there are $\beta(N^2-N)$ pairs $(i,j) \in A_\beta$ (pairs for which $d(i,j) < \phi(N)$), we immediately see that

$$\beta(N^2-N) \leq \sum_{(i,j) \in A_\beta} d(i,j) \quad (6)$$

Consequently, for fixed β :

$$C \in [0, \beta] \quad (7)$$

Note that the upper bound, β , can actually be reached, namely when all pairs $(i,j) \in A_\beta$ are at distance 1 (they are directly connected). We next derive a (much) better lower bound. Yet, relation (7) is all we need to study the limiting behavior of the following examples.

3.5 A limiting procedure for trees and disjoint unions of networks

We remind the reader that trees are important concepts in the information sciences. Distances between nodes in a tree, representing a hierarchical thesaurus, have been studied in the context of knowledge-based information retrieval (Kim & Kim, 1990). Let T be a balanced tree with branching factor $b > 1$ (Fig.2). The number of nodes in such a tree with depth d is:

$$N_d = 1 + b + b^2 + \dots + b^d = \frac{b^{d+1} - 1}{b - 1} \quad (8)$$

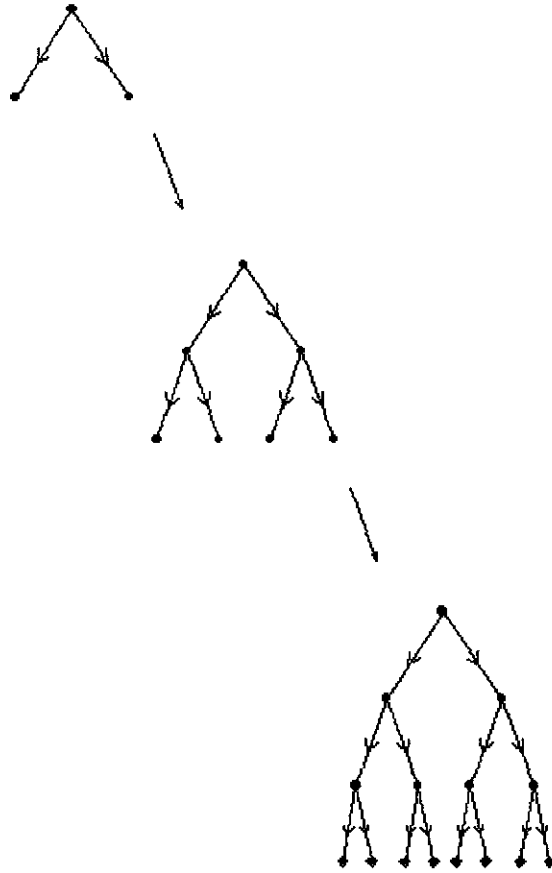


Fig.2 Construction of a tree with branching factor 2

In order to find the connection β_d of a tree at depth d we proceed step by step. At depth 1 there are b links of length 1. Expanding the tree and reaching depth 2 leads to b^2 links of length 2, plus b^2 new links of length 1. At the next expansion b^3 links of length 1, of length 2 and of length 3 are created. This yields at depth d a total number of links equal to:

$$\sum_{i=1}^d i b^i = \frac{d b^{d+2} - (d+1) b^{d+1} + b}{(b-1)^2} \quad (9)$$

Consequently,

$$\beta_d = \frac{db^{d+2} - (d+1)b^{d+1} + b}{(b-1)^2} = \frac{db^{d+1} - (d+1)b^d + 1}{\left(\frac{b^{d+1}-1}{b-1}\right)\left(\frac{b^{d+1}-1}{b-1}-1\right)} \quad (10)$$

The connection coefficient β_d is clearly smaller than 1. Moreover, $\lim_{d \rightarrow \infty} \beta_d = 0$ which proves, by (7), that the limiting compactness value of this balanced tree with fixed branching factor ($b > 1$) is zero.

Consider now a network consisting of N nodes. Starting from this network we consider the following construction of an infinite network. In a first step we construct a $2N$ -node network by adding, in a disconnected way, a copy of the first one, i.e. there are no connections between the first copy and the second. The resulting graph is called the disjoint union of this network with itself. Then we iterate this procedure, leading to a network with $4N$, $8N$, and in general $2^m N$ nodes. We will show that the limiting BRS-compactness value (m tending to infinity) of this network is zero.

Let β be the connection coefficient of the original network. Then β_1 , the corresponding coefficient after the first iteration is equal to:

$$\beta_1 = \beta \frac{N-1}{2N-1} < \beta \frac{1}{2}$$

In general, when β_{m-1} is the connection coefficient for the network after $m-1$ iterations, then β_m , the connection coefficient after m iterations is:

$$\beta_m = \beta_{m-1} \frac{2^{m-1}N-1}{2^m N-1} < \beta \left(\frac{1}{2}\right)^m \quad (11)$$

As the connection coefficient is the upper bound for the compactness value of any network (7), this proves that the limiting compactness value of this infinite network is zero, and hence such a construction yields increasingly sparse networks.

4. Bounds for the sum of distances between connected nodes

4.1 Theorem

Given a network with N nodes and with connection coefficient β . If the length of the longest used path in the distance matrix is k_L then

$$\Sigma =: \sum_{(i,j) \in A_\beta} d(i,j) \leq \frac{k_L (3\beta(N^2 - N) - (k_L^2 - 1))}{3} \quad (12)$$

Proof.

If there exists a path of length k_L , then there also exist two paths of length $(k_L - 1)$, three paths of length $(k_L - 2)$, and so on, ending with k_L paths of length 1. Note, that all these paths are used in the distance matrix otherwise k_L would not be the longest one. This yields $k_L(k_L + 1)/2$ pairs (i, j) for which we know the exact distance $d(i, j)$. We obtain an upper bound for Σ by taking all $d(i, j)$ equal to k_L , except the $k_L(k_L + 1)/2$ ones mentioned above. This yields:

$$\begin{aligned} \sum_{(i,j) \in A_\beta} d(i,j) &\leq \left(\beta(N^2 - N) - \frac{k_L(k_L + 1)}{2} \right) k_L + \sum_{j=1}^{k_L} j(k_L + 1 - j) \\ &= \left(\beta(N^2 - N) - \frac{k_L(k_L + 1)}{2} \right) k_L + \frac{k_L(k_L^2 + 3k_L + 2)}{6} \\ &= \frac{k_L \left(6\beta(N^2 - N) - (3k_L^2 + 3k_L - k_L^2 - 3k_L - 2) \right)}{6} \\ &= \frac{k_L (3\beta(N^2 - N) - (k_L^2 - 1))}{3} \end{aligned}$$

This proves the theorem.

Corollary 1

If an N-node network with connection coefficient β has a maximum path length equal to k_L then its BRS-compactness value C satisfies the following inequality:

$$\frac{\beta \varphi(N)}{\varphi(N)-1} - \frac{k_L (3\beta(N^2-N) - k_L^2 + 1)}{3(N^2-N)(\varphi(N)-1)} \leq C \leq \beta \quad (13)$$

Note that if $k_L = 1$, the lower bound for C becomes equal to the upper bound β .

Corollary 2

A unidirectional N-chain is characterized by the following parameters:

$$\begin{aligned} \beta &= \frac{1}{2} \\ \Sigma &= \frac{N(N^2-1)}{6} \\ C &= \frac{3(N^2-N)\varphi(N) - N(N^2-1)}{6(N^2-N)(\varphi(N)-1)} \end{aligned}$$

Proof

Clearly, $\beta = \frac{1+2+\dots+(N-1)}{N^2-N} = \frac{1}{2}$. The fact that Σ is equal to $\frac{N(N^2-1)}{6}$ follows

from the proof of Theorem 4.1, noting that $k_L = N-1$ and hence the inequality in

Theorem 4.1 becomes an equality for a unidirectional chain. Finally,

$$\begin{aligned} C &= \beta \frac{\varphi(N)}{\varphi(N)-1} - \frac{\Sigma}{(N^2-N)(\varphi(N)-1)} \\ &= \frac{\varphi(N)}{2(\varphi(N)-1)} - \frac{N(N^2-1)}{6(N^2-N)(\varphi(N)-1)} \\ &= \frac{3(N^2-N)\varphi(N) - N(N^2-1)}{6(N^2-N)(\varphi(N)-1)} \end{aligned}$$

4.2 Proposition

Given a bi-directional network with N nodes and with connection coefficient β . If the length of the longest used path in the distance matrix is k_L then

$$\Sigma =: \sum_{(i,j) \in A_\beta} d(i,j) \leq \frac{k_L (3\beta(N^2 - N) - 2(k_L^2 - 1))}{3} \quad (14)$$

Proof

If there exists a path of length k_L , then there exists a second one, by the fact that the graph is bi-directional. There, similarly exist four paths of length $(k_L - 1)$, six paths of length $(k_L - 2)$, and so on, ending with $2k_L$ paths of length 1. This yields $k_L(k_L + 1)$ pairs (i, j) for which we know the exact distance $d(i, j)$. Again, we obtain an upper bound for Σ by taking all $d(i, j)$ equal to k_L , except the $k_L(k_L + 1)$ ones mentioned above. This yields:

$$\begin{aligned} \sum_{(i,j) \in A_\beta} d(i,j) &\leq (\beta(N^2 - N) - k_L(k_L + 1))k_L + 2 \sum_{j=1}^{k_L} j(k_L + 1 - j) \\ &= (\beta(N^2 - N) - k_L(k_L + 1))k_L + \frac{k_L(k_L^2 + 3k_L + 2)}{3} \\ &= \frac{k_L(3\beta(N^2 - N) - (3k_L^2 + 3k_L - k_L^2 - 3k_L - 2))}{3} \\ &= \frac{k_L(3\beta(N^2 - N) - 2(k_L^2 - 1))}{3} \end{aligned}$$

This proves the proposition.

Corollary (Plesnik, 1984)

A bi-directional N -chain is characterized by the following parameters:

$$\beta = 1$$

$$\Sigma = \frac{N(N^2 - 1)}{3}$$

$$C = \frac{3(N^2 - N)\varphi(N) - N(N^2 - 1)}{3(N^2 - N)(\varphi(N) - 1)}$$

Proof

This follows immediately from Proposition 4.2.

Best lower and upper bounds for Σ are known in graph theory (Entringer et al., 1976; Ng and Teh, 1966; Doyle and Graver, 1977; Plesník, 1984). The Σ -value for a bi-directional chain is a best upper bound for a graph with N vertices.

4.3 Acceptable functions for $\varphi(N)$ in the compactness formula

We explained already that we will always choose $\varphi(N) \geq N$. If now $\varphi(N) = N^\alpha$ ($\alpha \geq 1$) then a bi-directional N -chain has, by the previous corollary, a C -value

$$C = \frac{3N(N-1)N^\alpha - N(N^2-1)}{3N(N-1)(N^\alpha-1)} \quad (15)$$

If N tends to infinity this value tends to $2/3$, if $\alpha = 1$, and to 1 if $\alpha > 1$. This would mean that the compactness value of a network where some nodes may have arbitrarily large distances, can be as close to one as one likes. This is counterintuitive and provides a good argument for taking $\alpha = 1$. This example does not rule out the possibility of taking $\varphi(N) = cN$ ($c > 1$), yielding a compactness value for the bi-directional chain of $(3c-1)/3c$. Such a value is not a priori excluded or counterintuitive. Yet, following Botafogo, Rivlin and Shneiderman (1992) we will from now on take

$\varphi(N) = N$. This leads to the following formula for C:

$$C = \frac{\beta N}{N-1} - \frac{\sum_{(i,j) \in A_\beta} d(i,j)}{N(N-1)^2} \quad (16)$$

We note that even if $\varphi(N) = N$, the limiting value (for $N \rightarrow \infty$) of a bi-directional star (a single root, connected by bi-directional links to all other nodes) is one (Botafogo et al., 1992). In this graph the distance between two nodes (except if one of the nodes is the root) is equal to two. Such a bi-directional star is a model for a totally centralized network.

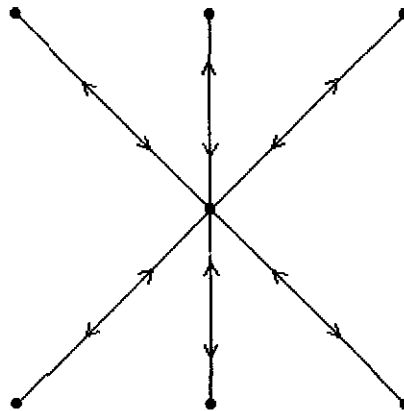


Fig. 3 A bi-directional star

Note

In a unidirectional, acyclic network the connection coefficient β is at most 1/2. Hence the BRS-compactness value of such a network is at most 0.5.

4.4 Proposition

Adding a new link between existing nodes in any network always increases the compactness value C.

Proof. This is trivial. Adding an extra link decreases the value of Σ , and as N stays constant, this means that C increases. The increase in C is at least equal to

$$\frac{1}{N(N-1)^2} \quad (17)$$

4.5 The meaning of cohesion and the importance of 'central' links

From Corollary 2 of Theorem 4.1 we know that Σ , the sum of the distances of all connected nodes of a unidirectional N -chain is $\frac{N(N-1)(N+1)}{6}$ and that its BRS-

compactness value is $\frac{2N-1}{6(N-1)}$ (take $\varphi(N) = N$ in Corollary 2). Note, in particular,

that the first (or the last) link in chain has a contribution in Σ equal to $N-1$. Indeed, this link participates once in the set of links of length 1, once in the set of links of length 2, and so on, ending with links of length $N-1$. Assuming N to be even we see, however, that the link in the middle, connecting node $\frac{N}{2}$ with node $\frac{N}{2}+1$, has a participation of 1 in the set of links of length 1, a participation of 2 in the set of links with length 2, increasing to a participation of $N/2$ in the set of links with length $N/2$, and then again a decreasing participation. Hence the middle link contributes $\frac{N^2}{4}$ to the Wiener index. This calculation illustrates the fact that a central link plays a more important role in the determination of the cohesion, as measured by BRS-compactness, than a more peripheral one. This is a desired property of a cohesion measure. If one were interested in replacing the BRS-compactness measure by another measure of cohesion, then this measure must have at least a similar property.

4.6 Weaker inequalities for the sum of all distances that are strictly smaller than N

In this section we will derive weaker inequalities than inequality (12). Although weaker they have the advantage that they depend on less parameters and, hence, can be used when certain data (such as k_L) are not known. Recall that with $\sum_{(i,j) \in A_\beta} d(i,j)$ denoted as Σ :

$$\Sigma \leq \frac{k_L (3\beta(N^2 - N) - (k_L^2 - 1))}{3} \quad (12)$$

As $\beta \leq 1$, we always have:

$$\Sigma \leq \beta \frac{k_L (3(N^2 - N) - (k_L^2 - 1))}{3} \quad (18)$$

Considering the second factor on the left-hand side as a function of k_L (with N fixed), we see that it is increasing for $k_L < k_0 = \sqrt{N(N-1) + \frac{1}{3}}$. Because $k_0 > N - 1$, and $1 \leq k_L \leq N - 1$, we may replace k_L by $N - 1$. This leads to the following (weaker) inequality:

$$\Sigma \leq \beta \frac{(N-1)N(2N-1)}{3} \quad (19)$$

Inequality (19) leads to the following inequality between β and C .

4.7 Theorem

Given a network with N nodes, connection coefficient β and compactness C , then:

$$C \leq \beta \leq \min\left(\frac{3(N-1)}{N+1}C, 1\right) \leq \min(3C, 1) \quad (20)$$

Proof.

The first inequality in formula (20) follows from (13). The last one is trivial, and so is the fact that β and C are always smaller than or equal to 1. So, we only have to show

that $\beta \leq \frac{3(N-1)}{N+1} C$. This inequality follows from this chain of relations:

$$\begin{aligned}
C &= \frac{\beta N}{N-1} - \frac{\Sigma}{N(N-1)^2} && \text{by (16)} \\
&\geq \frac{\beta N}{N-1} - \frac{\beta(N-1)N(2N-1)}{3N(N-1)^2} && \text{by (19)} \\
&= \beta \left(\frac{3N-2N+1}{3(N-1)} \right) = \beta \frac{N+1}{3(N-1)}
\end{aligned}$$

Consequently: $\beta \leq \frac{3(N-1)}{N+1} C$.

5. Adding one node: its influence on the compactness value

In this section we show that adding one node, disconnected from all others, lowers the compactness value of the network. Adding, however, a node that is connected to all others increases the compactness value. This shows that the compactness measure as proposed by Botafogo, Rivlin and Shneiderman (1992) has nice (and expected) properties.

5.1 Adding one node disconnected from all other ones.

The reader will notice that the proof of this result is surprisingly difficult (or at least more complicated than the authors expected). If the compactness value was zero before the expansion it stays zero, and if the compactness value was 1 it certainly decreases. We next consider compactness values that lie strictly between 0 and 1, hence with β -values strictly between 0 and 1.

If the compactness value before the expansion was

$$C = \frac{\beta N}{N-1} - \frac{\Sigma}{N(N-1)^2} \quad (16)$$

with Σ the sum of all $d(i,j)$ not equal to N , then its compactness value after the expansion is:

$$C' = \frac{\beta'(N+1)}{N} - \frac{\Sigma}{(N+1)N^2} \quad (21)$$

where β' denotes the connection coefficient of the new, expanded network. Now, $\beta' = \beta \frac{N-1}{N+1}$, so that we have to show that:

$$\begin{aligned} \frac{\beta N}{N-1} - \frac{\Sigma}{N(N-1)^2} &> \frac{\beta(N-1)}{N} - \frac{\Sigma}{(N+1)N^2} \\ &\text{or} \\ \frac{\beta N}{N-1} - \frac{\beta(N-1)}{N} &> \frac{\Sigma}{N(N-1)^2} - \frac{\Sigma}{(N+1)N^2} \end{aligned}$$

This inequality reduces, after some simple algebra to:

$$\Sigma(3N-1) < \beta(2N^4 - N^3 - 2N^2 + N) = \beta N(N^2-1)(2N-1) \quad (22)$$

Applying inequality (19) gives that it is sufficient to prove:

$$\beta \frac{(N-1)N(2N-1)}{3} (3N-1) \leq \beta N(N-1)(N+1)(2N-1)$$

or, equivalently:

$$3N-1 \leq 3(N+1)$$

This inequality is clearly true, proving that adding one node, disconnected from all others decreases the compactness value of the network.

We observe that eliminating β' and Σ from (16) and (21) leads to the following formula expressing C' as a function of C , N and β :

$$C' = C \frac{(N-1)^2}{N(N+1)} + \beta \frac{N-1}{N(N+1)} \quad (23)$$

5.2 Adding one node connected to all other ones

If the compactness value before the expansion was

$$\frac{N^3 - N^2 - S}{N^3 - 2N^2 + N} \quad (24)$$

where S denotes the sum of all $d(i,j)$, then its compactness value after the expansion is:

$$\frac{(N+1)^3 - (N+1)^2 - (S+2N)}{(N+1)^3 - 2(N+1)^2 + (N+1)}$$

Hence, we have to show that

$$\frac{N^3 - N^2 - S}{N^3 - 2N^2 + N} \leq \frac{N^3 + 2N^2 - N - S}{N^3 + N^2}$$

or

$$(N^3 - N^2 - S)(N^3 + N^2) \leq (N^3 + 2N^2 - N - S)(N^3 - 2N^2 + N)$$

After some calculations this leads to:

$$N^2(N-1)(3N-1) \leq SN(3N-1)$$

or

$$N(N-1) \leq S$$

Because $N(N-1)$ is the smallest possible value for S, this proves 5.2.

6. Another formula for BRS-compactness

Consider a network with N nodes and with connection coefficient β . Then we introduce the following definition.

6.1 Definition

Let δ_k , $k = 1, 2, \dots, L_\beta$, (recall that L_β denotes the largest possible distance between two nodes, given that the connection coefficient β) is be the fraction of the nodes in A_β for which $d(i,j) = k$. As the δ_k s are fractions, we have:

$$\sum_{k=1}^{L_\beta} \delta_k = 1 \quad (25)$$

Consequently, this leads to the following new formula for C:

$$\begin{aligned} C &= \frac{\beta N}{N-1} - \frac{\sum_{k=1}^D k \delta_k \beta (N^2 - N)}{N(N-1)^2} \\ &= \frac{\beta}{N-1} \left(N - \sum_{k=1}^{L_\beta} k \delta_k \right) \end{aligned} \quad (26)$$

We know that $\beta = 0.5$ for a unidirectional N-chain. In this case we further have:

$$\delta_1 = \frac{2}{N} > \delta_2 = \frac{2(N-2)}{N(N-1)} > \delta_3 \cdots > \delta_{N-1} = \frac{2}{N(N-1)} \quad (27)$$

For a bi-directional N-chain $\beta = 1$, but the δ_k 's are the same as for a unidirectional one. For a unidirectional N-loop, $\beta = 1$, and all δ_k 's are equal to $1/(N-1)$.

This leads to the following research problem: for which networks is

$$\delta_D \leq \delta_{D-1} \leq \cdots \leq \delta_1 \quad (28)$$

Note that it is easy to find networks, unidirectional as well as bi-directional ones where (28) is not satisfied. Indeed for the following network with 7 nodes (Fig.4) , we have:

$$\beta = \frac{5}{14}, \delta_1 = \frac{2}{5}, \delta_2 = \frac{3}{5} \text{ and } C = \frac{9}{28}$$

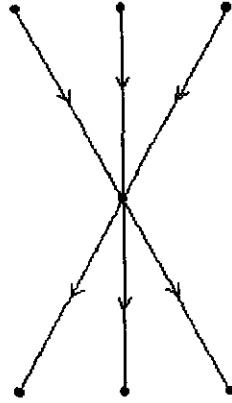


Fig.4 A unidirectional network where inequality (28) is not satisfied

7. Disjoint unions of arbitrary networks

Let G_1 and G_2 be two disjoint networks, the first having N_1 (> 1) nodes, the second one having N_2 (> 1) nodes. We next consider their (disjoint) union. The aim of this section is to obtain the compactness value C and connection coefficient β of this union, as a function of the compactness values of G_1 and G_2 (denoted respectively as C_1 and C_2), their connection coefficients β_1 and β_2 and the number of nodes N_1 and N_2 .

7.1 Lemma

With the notations introduced above, we have:

$$\beta = \frac{\beta_1 N_1 (N_1 - 1) + \beta_2 N_2 (N_2 - 1)}{(N_1 + N_2)(N_1 + N_2 - 1)} \quad (29)$$

Proof.

Graph G_1 contains, by the definition of the connection coefficient, $\beta_1(N_1^2 - N_1)$

connected pairs of nodes. Similarly, G_2 contains $\beta_2(N_2^2 - N_2)$ connected pairs of nodes. Then β , the connection coefficient of G , the disjoint union of G_1 and G_2 , is:

$$\beta = \frac{\beta_1 N_1(N_1 - 1) + \beta_2 N_2(N_2 - 1)}{(N_1 + N_2)(N_1 + N_2 - 1)}$$

Corollary

The connection coefficient $\beta = \lambda_1 \beta_1 + \lambda_2 \beta_2$ with $\lambda_1, \lambda_2 \in [0, 1]$, and $\lambda_1 + \lambda_2 < 1$.

Proof. Clearly $\lambda_j = \frac{N_j^2 - N_j}{N_1^2 + 2N_1N_2 + N_2^2 - N_1 - N_2}$, with $j = 1, 2$. Now,

$$\lambda_1 + \lambda_2 = \frac{N_1^2 + N_2^2 - N_1 - N_2}{N_1^2 + N_2^2 - N_1 - N_2 + 2N_1N_2} < 1 \quad (30)$$

This result is not unexpected: β is not a convex combination of β_1 and β_2 ($\lambda_1 + \lambda_2 \neq 1$) as there is a disjoint union involved. Intuitively: there is a loss in cohesion. This corresponds to a decrease in compactness (at least if G_1 and G_2 are 'similar') as will be shown shortly.

7.2 Theorem

Using the notation introduced above we find for the value C , of the compactness of a disjoint union:

$$C = \frac{C_1 N_1(N_1 - 1)^2 + C_2 N_2(N_2 - 1)^2 + N_1 N_2 (\beta_1(N_1 - 1) + \beta_2(N_2 - 1))}{(N_1 + N_2)(N_1 + N_2 - 1)^2} \quad (31)$$

Proof.

Denoting in the first graph $\sum_{\substack{d(i,j) < N_1 \\ i \neq j}} d(i,j)$ by Σ_1 , $\sum_{\substack{d(i,j) < N_2 \\ i \neq j}} d(i,j)$ by Σ_2 in the second

one, and $\sum_{\substack{d(i,j) < N_1+N_2 \\ i \neq j}} d(i,j)$ by Σ in the union, we have:

$$\begin{aligned} C_1 &= \frac{\beta_1 N_1}{N_1 - 1} - \frac{\Sigma_1}{N_1(N_1 - 1)^2} \\ C_2 &= \frac{\beta_2 N_2}{N_2 - 1} - \frac{\Sigma_2}{N_2(N_2 - 1)^2} \\ C &= \frac{\beta(N_1 + N_2)}{N_1 + N_2 - 1} - \frac{\Sigma}{(N_1 + N_2)(N_1 + N_2 - 1)^2} \end{aligned} \quad (32)$$

It is clear that $\Sigma = \Sigma_1 + \Sigma_2$ because we have a disjoint union. Substituting the value of β (29) and this sum in expression (32) yields:

$$C = \frac{(N_1 + N_2) \beta_1 N_1 (N_1 - 1) + \beta_2 N_2 (N_2 - 1)}{N_1 + N_2 - 1} - \frac{\Sigma_1 + \Sigma_2}{(N_1 + N_2)(N_1 + N_2 - 1)^2}$$

Rearranging terms gives:

$$\begin{aligned} C &= \frac{\beta_1 N_1}{N_1 - 1} \left(\frac{N_1 - 1}{N_1 + N_2 - 1} \right)^2 + \frac{\beta_2 N_2}{N_2 - 1} \left(\frac{N_2 - 1}{N_1 + N_2 - 1} \right)^2 \\ &\quad - \frac{\Sigma_1}{N_1(N_1 - 1)^2} \frac{N_1(N_1 - 1)^2}{(N_1 + N_2)(N_1 + N_2 - 1)} \\ &\quad - \frac{\Sigma_2}{N_2(N_2 - 1)^2} \frac{N_2(N_2 - 1)^2}{(N_1 + N_2)(N_1 + N_2 - 1)} \end{aligned}$$

As $\frac{\Sigma_j}{N_j(N_j - 1)^2} = \frac{\beta_j N_j}{N_j - 1} - C_j$, for $j = 1, 2$, we obtain:

$$\begin{aligned} C &= C_1 \frac{N_1}{N_1 + N_2} \left(\frac{N_1 - 1}{N_1 + N_2 - 1} \right)^2 + C_2 \frac{N_2}{N_1 + N_2} \left(\frac{N_2 - 1}{N_1 + N_2 - 1} \right)^2 \\ &\quad + \frac{\beta_1 N_1}{N_1 - 1} \left[\left(\frac{N_1 - 1}{N_1 + N_2 - 1} \right)^2 - \frac{N_1(N_1 - 1)^2}{(N_1 + N_2)(N_1 + N_2 - 1)^2} \right] \\ &\quad + \frac{\beta_2 N_2}{N_2 - 1} \left[\left(\frac{N_2 - 1}{N_1 + N_2 - 1} \right)^2 - \frac{N_2(N_2 - 1)^2}{(N_1 + N_2)(N_1 + N_2 - 1)^2} \right] \end{aligned}$$

Simplifying this expression leads to:

$$C = C_1 \frac{N_1}{N_1 + N_2} \left(\frac{N_1 - 1}{N_1 + N_2 - 1} \right)^2 + C_2 \frac{N_2}{N_1 + N_2} \left(\frac{N_2 - 1}{N_1 + N_2 - 1} \right)^2 \\ + \frac{\beta_1 N_1}{N_1 - 1} \frac{(N_1 - 1)^2 N_2}{(N_1 + N_2)(N_1 + N_2 - 1)^2} + \frac{\beta_2 N_2}{N_2 - 1} \frac{(N_2 - 1)^2 N_1}{(N_1 + N_2)(N_1 + N_2 - 1)^2}$$

or:

$$C = \frac{C_1 N_1 (N_1 - 1)^2 + C_2 N_2 (N_2 - 1)^2 + N_1 N_2 (\beta_1 (N_1 - 1) + \beta_2 (N_2 - 1))}{(N_1 + N_2)(N_1 + N_2 - 1)^2} \quad (31)$$

This proves the theorem.

7.3 Some special cases

1°) Taking $N_1 = N_2 = n$, $C_1 = C_2 = c$ and $\beta_1 = \beta_2 = b$ gives:

$$C = \frac{c(n-1)^2 + bn(n-1)}{(2n-1)^2} \quad (33)$$

2°) Taking $N_2 = 1$ (and leaving β_2 and C_2 unspecified, but finite) gives:

$$C = \frac{C_1 (N_1 - 1)^2 + \beta_1 (N_1 - 1)}{(N_1 + 1)N_1}$$

which is exactly formula (23). This shows that, although formula (23) does not follow from the proof of theorem 7.2, it does follow from formula (31), showing that formula (31) is also correct if one of the two (or even both!) networks consists of one point.

3°) Taking $\beta_1 = \beta_2 = 1$ and $C_1 = C_2 = 1$ gives the disjoint union of two complete networks. Its compactness is:

$$\begin{aligned}
C &= \frac{N_1(N_1-1)^2 + N_2(N_2-1)^2 + N_1N_2(N_1+N_2-2)}{(N_1+N_2)(N_1+N_2-1)^2} \\
&= \frac{N_1(N_1-1) + N_2(N_2-1)}{(N_1+N_2)(N_1+N_2-1)} \quad (34)
\end{aligned}$$

as obtained by Botafogo, Rivlin and Shneiderman (1992). If, moreover, $N_1 = N_2 = n$

then the BRS-compactness is equal to $\frac{n-1}{2n-1}$, which tends to 1/2 if n tends to infinity.

7.4 Theorem

If $N_1 = N_2 = N$, $C_1 = C_2 = c$ and $\beta_1 = \beta_2 = b$ then:

$$C = \frac{c(N-1)^2 + bN(N-1)}{(2N-1)^2} < c \quad (35)$$

Proof.

We know by (20) that $b \leq \frac{3(n-1)}{n+1}c$, hence:

$$\begin{aligned}
C &= \frac{n-1}{(2n-1)^2} (c(n-1) + bn) \\
&\leq \frac{n-1}{(2n-1)^2} \left(c(n-1) + n \frac{3(n-1)}{n+1} c \right) \\
&= \left(\frac{n-1}{2n-1} \right)^2 c \left(1 + \frac{3n}{n+1} \right) < \left(\frac{1}{2} \right)^2 c 4 = c
\end{aligned}$$

We observe that this result as well as that of theorem 5.1 are derived using formula (20), which in itself is a result of the general formula (19) giving an upper bound for the sum of all distances between connected nodes in a network.

We end this section by showing that the conditions $N_1 = N_2 = N$, $C_1 = C_2 = c$ and

$\beta_1 = \beta_2 = b$ do not imply that the two graphs are isomorphic. This implies also that the result of theorem 7.4 is not only valid for identical graphs for also for some non-isomorphic ones.

Consider the graphs G_1 and G_2 (see Fig. 5). All links are assumed to be bi-directional. They have both 7 nodes ($n = 7$) and clearly have a connection coefficient of 1 ($b = 1$). Finally, their Σ -values are 64, so that they have the same compactness

value: $c = \frac{7}{6} \frac{64}{7 \cdot 6^2} = \frac{115}{126}$. Moreover, the two graphs G_1 and G_2 are non-isomorphic

as G_1 has a point with out-degree 1 (namely point 2), while G_2 does not have such a point. Using the same construction, both now with unidirectional links leads to another example.

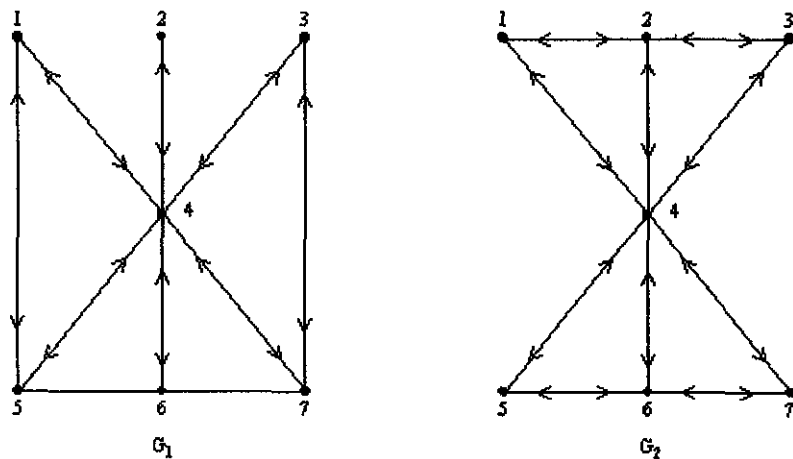


Fig. 5 Two non-isomorphic graphs with the same compactness value

8. Unidirectional networks

For unidirectional, acyclic networks, such as citation networks, De Bra (2000) introduced another convention, leading to the following definition.

8.1 Definition: De Bra's symmetric citation distance matrix

Given a set of N documents, then De Bra (2000) defines the citation distance matrix D_{DB} as follows: $d(i, j)$ is equal to the length of the shortest path (in number of links) in the citation network from document i to document j, if such a path exists. Further: $d(i, j) = d(j, i)$, and $d(i, i) = 0$. Finally, all other entries of the D_{DB} -matrix are equal to N. Note that the matrix D_{DB} is not the D_B -matrix of the corresponding undirected network. The relation between the D_{DB} -matrix, the D_B -matrix and the D_B -matrix of the corresponding undirected network is illustrated in the following example (Fig.6).

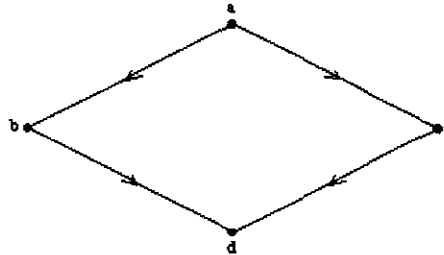


Fig.6 A network used to illustrate the difference between the general and the De Bra approach

The D_B , D_{DB} and D_B -matrix of the corresponding undirected network are:

	a	b	c	d		a	b	c	d		a	b	c	d
a	0	1	1	2		0	1	1	2		0	1	1	2
b	4	0	4	1		1	0	4	1		1	0	2	1
c	4	4	0	1		1	4	0	1		1	2	0	1
d	4	4	4	0		2	1	1	0		2	1	1	0

De Bra made the citation matrix artificially symmetric. We show now that, for acyclic unidirectional networks, this is not really necessary.

8.2 Different representations of an acyclic unidirectional network

An acyclic unidirectional network such as a citation network, can be described in the

following three ways: 1) as a general network using the BRS distance matrix (where the network just happens to be unidirectional and acyclic); 2) using the BRS distance matrix, but it is given that the network is unidirectional and acyclic (this condition influences the min-value in the compactness formula); 3) using De Bra's conventions.

In all three cases $MAX = N(N^2 - N)$. In the first case $MIN = N^2 - N$; in the second one it is $N(N^2 - 1)/2$; and in the last one it is again $N^2 - N$. This leads to the following compactness formulae (denoted respectively as C , C_{BIU} and C_{DB}):

$$C = \frac{N^3 - N^2 - \left(\sum_{(i,j) \in A_B} d(i,j) + \frac{(N^2 - N)}{2} N \right)}{(N^3 - N^2) - (N^2 - N)} \quad (36)$$

$$C_{BIU} = \frac{N^3 - N^2 - \left(\sum_{(i,j) \in A_B} d(i,j) + \frac{(N^2 - N)}{2} N \right)}{(N^3 - N^2) - \frac{N(N^2 - 1)}{2}} \quad (37)$$

and finally:

$$C_{DB} = \frac{N^3 - N^2 - \left(2 \sum_{(i,j) \in A_B} d(i,j) \right)}{(N^3 - N^2) - (N^2 - N)} \quad (38)$$

These three formulae are all special cases of the general form introduced in Definition 2.

The next theorem shows the relation between the three compactness formulae.

8.3 Theorem

For unidirectional, acyclic networks such as citation networks, we have:

1) $C_{B|U} = C_{DB}$

2) $2C = C_{B|U} = C_{DB}$

Proof. Denoting $\sum_{(i,j) \in A_B} d(i,j)$ simply by Σ' and $N^2 - N$ by m , we have:

$$C_{B|U} = \frac{Nm - \Sigma' - \frac{Nm}{2}}{Nm - \frac{m(N+1)}{2}} = \frac{Nm - 2\Sigma'}{Nm - m} = C_{DB}$$

Further:

$$C = \frac{Nm - \Sigma' - \frac{Nm}{2}}{Nm - m} = \frac{1}{2} \frac{Nm - 2\Sigma'}{Nm - m} = \frac{1}{2} C_{DB} = \frac{1}{2} C_{B|U}$$

This proves the theorem.

8.4 Corollaries and comments

It is easy to check that the previous result is also true in general, i.e. with $\varphi(N)$ instead of N .

As the BRS-compactness value C of an acyclic, unidirectional network is at most 0.5, this also implies that De Bra's measure can be considered as a renormalization (resulting again in values between 0 and 1) of the BRS-value for the case of acyclic, unidirectional networks.

In (Fang & Rousseau, 2001) the compactness of some small lattice citation networks has been calculated using De Bra's formula (De Bra, 2000).

8.5 The non-uniqueness of the De Bra matrix description

If one reverses all arrows in a digraph, then the new network will be called the *reversed network*. The operation of reversing all arrows in a network is called *reversion*. It is clear that the De Bra matrix of a citation network and that of its reversion are the same. Observe that, generally, citation networks that are each other's reversion, are non-isomorphic.

Property. De Bra's description is non-unique. By this we mean that there exist citation networks that are non-isomorphic and are not each other's reversion, and yet yield the same De Bra matrix representation.

It suffices to give an example. The citation networks represented by Figs. 7 a and b are clearly non-isomorphic and not each other's reversion. Yet, they both have the following De Bra matrix representation.

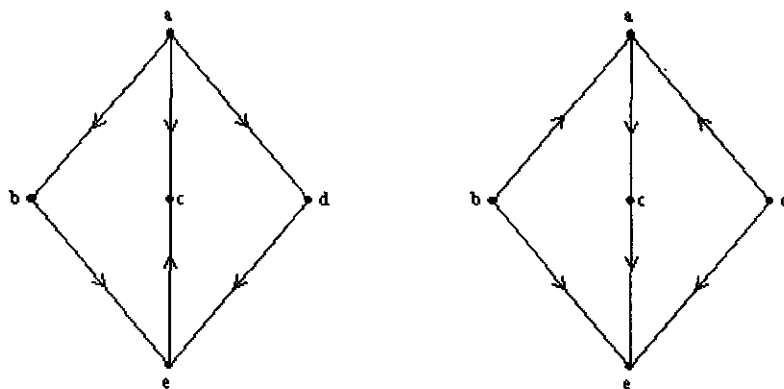


Fig.7 Two non-isomorphic graphs with the same De bra matrix representation

Matrix representation

$$\begin{pmatrix} & a & b & c & d & e \\ a & 0 & 1 & 1 & 1 & 2 \\ b & 1 & 0 & 2 & 5 & 1 \\ c & 1 & 2 & 0 & 2 & 1 \\ d & 1 & 5 & 2 & 0 & 1 \\ e & 2 & 1 & 1 & 1 & 0 \end{pmatrix}$$

It is clear that this non-uniqueness shows the non-optimality of De Bra's representation. Consequently, it seems better to stick to the original BRS matrix representation.

9. Calculation of the compactness of a balanced tree

In section 3 we calculated the connection coefficient of a balanced tree of depth d . We will now continue the calculations in order to obtain this tree's compacted value.

First we need its Σ -value, denoted as Σ_d . This is obtained as follows:

$$\begin{aligned} \Sigma_d &= \sum_{i=1}^d b^i + 2 \sum_{i=2}^d b^i + \dots + d \sum_{i=d}^d b^i \\ &= \frac{(b^{d+1} - b) + 2(b^{d+1} - b^2) + \dots + d(b^{d+1} - b^d)}{b-1} \\ &= \frac{\frac{d(d+1)}{2} b^{d+1} - \frac{db^{d+2} - (d+1)b^{d+1} + b}{(b-1)^2}}{b-1} \\ &= \frac{db^{d+3}(d+1) - 2db^{d+2}(d+2) + b^{d+1}(d+1)(d+2) - 2b}{2(b-1)^3} \end{aligned} \quad (39)$$

This leads to the following compactness value:

$$\begin{aligned}
C_d &= \frac{(db^{d+1} - (d+1)b^d + 1)(b^{d+1} - 1)}{(b^{d+1} - 1)(b^d - 1)b(b^d - 1)} \\
&\quad - \frac{(db^{d+3}(d+1) - 2db^{d+2}(d+2) + b^{d+1}(d+1)(d+2) - 2b)(b-1)^3}{2(b-1)^3(b^{d+1} - 1)b^2(b^d - 1)^2} \\
&= \frac{b^{d+1}(2b^{d+1}(bd - d - 1) - b^2d(d+1) + 2b(d^2 + d + 1) - d(d+1))}{2(b^d - 1)^2(b^{d+1} - 1)} \quad (40)
\end{aligned}$$

Taking $d = 1$ in equation (40) gives $C_1 = \frac{1}{b+1}$. If, moreover, we take the limit for b

tending to 1 in (40) we find the compactness value of a unidirectional chain of length

d , namely $\frac{2d+1}{6d}$ (checked by computer).

10. Calculation of the compactness of an ensemble

In this section we present another construction of a unidirectional network based on simple building blocks. This construction generalizes the unidirectional chain. We will compute its BRS-compactness and study its limiting behavior. We are convinced that examples such as this one, are important in order to gain experience with this measure of cohesion. Moreover, the more complicated an example is, the more it resembles real-world networks, and, hence, can be used for modeling purposes.

Construction of an ensemble

Consider L 'levels'. Each level j consists of n_j nodes. Nodes at a fixed level are disconnected between each other, but are connected to each node at level $j+1$ (except of course nodes at level L). Connections are unidirectional and no other

connections exist. This graph will be called an ensemble. An example, with 4 levels is given in Fig. 8.

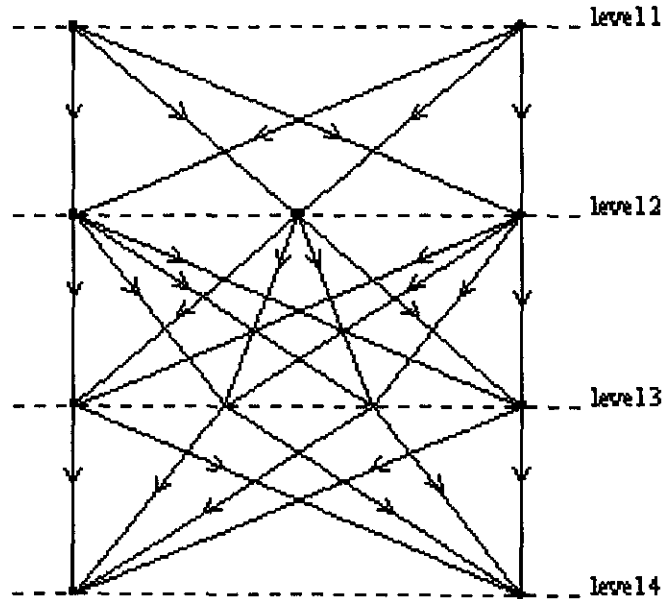


Fig.8 Ensemble with 4 levels

The total number of nodes in the ensemble is $N = \sum_{j=1}^L n_j$.

We now determine the Σ -value of the BRS-representation. The nodes at level 1 contribute:

$$n_1 \text{ times } ([n_2 + 2n_3 + \dots + (L-1)n_L] + (n_1-1) \text{ times } N) ;$$

nodes at level 2 contribute:

$$n_2 \text{ times } ([n_3 + 2n_4 + \dots + (L-2)n_L] + [(n_2-1) + n_1] \text{ times } N) ;$$

in general, nodes at level j contribute ($j = 1, \dots, L-1$):

$$n_j \text{ times } ([n_{j+1} + 2n_{j+2} + \dots + (L-j)n_L] + [(n_j-1) + n_1 + \dots + n_{j-1}] \text{ times } N) ;$$

finally at level L we have $n_L((n_L - 1) + n_1 + \dots + n_{L-1}) \text{ times } N$.

This leads to the following total:

$$N \sum_{j=1}^L n_j \left(\sum_{k=1}^j n_k - 1 \right) + \sum_{j=1}^{L-1} n_j \left(\sum_{k=1}^{L-j} k n_{j+k} \right)$$

We next consider the special case that all n_j are equal, hence $N = nL$. Then the total is:

$$\begin{aligned} nL \sum_{j=1}^L n(nj-1) + \sum_{j=1}^{L-1} n \left(\sum_{k=1}^{L-j} k n \right) \\ = n^3 L \frac{L(L+1)}{2} - n^2 L^2 + n^2 \sum_{j=1}^{L-1} \frac{(L-j)(L-j+1)}{2} \end{aligned}$$

Putting $L-j = k$ leads to:

$$\begin{aligned} n^2 L^2 \left(n \frac{L+1}{2} - 1 \right) + n^2 \sum_{k=1}^{L-1} \frac{k(k+1)}{2} \\ = n^2 L^2 \left(n \frac{L+1}{2} - 1 \right) + n^2 \frac{(L-1)L(L+1)}{6} \\ = n^2 L \left(\frac{nL(L+1)}{2} - L + \frac{L^2-1}{6} \right). \end{aligned}$$

Consequently, the BRS-compactness value is:

$$\begin{aligned} C &= \frac{n^3 L^3 - n^2 L^2 - n^2 L \left(\frac{nL(L+1)}{2} - L + \frac{L^2-1}{6} \right)}{n^3 L^3 - 2n^2 L^2 + nL} \\ &= \frac{n(3nL^2 - 3nL - L^2 + 1)}{6(n^2 L^2 - 2nL + 1)} \end{aligned}$$

In particular, if $n = 1$ (a unidirectional chain consisting of L nodes), C is equal to:

$$\frac{2L-1}{6(L-1)}$$

This result is in agreement with Corollary 3 of Theorem 4.1, with $\varphi(N) = N = L$. Hence we see that a chain consisting of two nodes has a BRS- compactness value of 0.5 (the maximum value for a unidirectional network), a chain consisting of three nodes of 5/12, for four nodes it is 7/18, and so on, with a limiting value of 1/3.

We fix n and consider the limit for $L \rightarrow \infty$. Then the limiting value is:

$$\frac{3n^2 - n}{6n^2} = \frac{1}{2} - \frac{1}{6n}$$

For $n = 1$, this is $2/6$, for $n = 2$ it is $5/12$, for $n = 3$ it is $8/18$ and so on. If also n tends to infinity, we find the value 0.5 , as expected for a unidirectional network.

We next fix L , and consider the limit for $n \rightarrow \infty$. This limit value is equal to

$$\frac{3L^2 - 3L}{6L^2} = \frac{1}{2} - \frac{1}{2L}$$

If now, L tends to infinity, we find (again) 0.5 . For $L = 1$, we find 0 , also as it is expected to be.

10. Conclusion

The Net, citation networks as well as scientific collaboration networks are nowadays in the center of attention. We hope that the structural measure of cohesion, namely BRS-compactness, studied in this article will prove to be a useful element for their description. The fact that this measure has the well-known Wiener index as main component leads to the suggestion to find and apply more topological indices. These indices play an important role in the description of molecular graphs in computational and mathematical chemistry (Gutman & Polansky, 1986; Rouvray, 1986; Trinajstić, 1992). It has, moreover, been shown that the Wiener index is correlated to a large number of physiochemical properties such as boiling point, melting point, refractive index, surface tension and viscosity of chemical molecules. There seems to be no

reason why they could not play an equally important role to characterize networks in the context of the information sciences.

One clear restriction of the measure studied here and by Botafogo, Rivlin and Shneiderman is the fact that it relates to unweighted networks. Yet, there are usually many links between the nodes in a graph, be it authors that receive many citations from the same colleagues, or sites on the Internet that are connected through many links. This leads to a weighted graph structure that will be studied in a following paper (Egghe & Rousseau, 2001).

We conclude with an open problem. Given a rational number between 0 and 1, does there exist a graph with that particular BRS-compactness value?

References

- Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks
Science, 286, 509-512.
- Berge, C. (1967). *Théorie des graphes et ses applications*. Paris: Dunod.
- 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., & 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).
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R.,
Tomkins, A., & Wiener, J. (2000). Graph structure in the web. *Proceedings of*

- the 9th International World Wide Web Conference (Amsterdam).
- Calvi, L. & De Bra, P. (1997). Using dynamic hypertext to create multi-purpose textbooks. Proceedings of ED-MEDIA 97 (Calgary).
- Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S.R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Hypersearching the Web. *Scientific American* 280(6), 54-60.
- Chen, W.K. (1971). *Applied graph theory*. Amsterdam: North-Holland.
- 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.
- Ding, Y., Foo, S., & Chowdhury, G. (1998). A bibliometric analysis of collaboration in the field of information retrieval. *International Information & Library Review*, 30, 367-376.
- Doyle, J.K., & Graver, J.E. (1977). Mean distance in a graph. *Discrete Mathematics*, 17, 147-154.
- Egghe, L. & Rousseau, R. (2001). BRS-compactness for weighted graphs (in preparation).
- Entringer, R.C., Jackson, D.E. & Snyder, D.A. (1976). Distance in graphs. *Czechoslovak Mathematical Journal*, 26, 283-296.
- Fang, Y., & Rousseau, R. (2001). Lattices in citation networks: an investigation into the structure of citation graphs. *Scientometrics* (to appear).
- Geller, N. (1978). On the citation influence methodology of Pinski and Narin. *Information Processing and Management* 14, 93-95.
- Gibbons, A. (1985). *Algorithmic graph theory*. Cambridge (UK): Cambridge University

Press.

Gutman, I., & Polansky, O. (1986). *Mathematical concepts in organic chemistry*.

Berlin: Springer.

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., & 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.

Kim, Y.W. and Kim, J.H. (1990). A model of knowledge based information retrieval with hierarchical concept graph. *Journal of Documentation*, 46, 113-136.

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

Kleinberg, J., Kumar, S.R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). The Web as a graph: measurements, models, and methods. *Proceedings of the Fifth Annual International Computing and Combinatorics Conference*.

Knuth, D. (1969). *The art of computer programming. Vol.1. Fundamental algorithms*. Reading (MA): Addison-Wesley.

Kretschmer, H. (1999). Types of two-dimensional and three-dimensional collaboration patterns. In: Macias-Chapula, C. (Ed.): *Proceedings of the seventh conference of the international society for scientometrics and informetrics*. Universidad de Colima (Mexico), 244-266.

- Leazer, G.H. & 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. & Harrison, R. (1998). Applying metrics to the evaluation of educational hypermedia applications. *Journal of Universal Computer Science*, 4, 4. To be found at:
http://www.iicm.edu/jucs_4_4/applying_metrics_to_the/paper.html
- Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science*, 98(2), 404-409.
- Ng, C.P., & Teh, H.H. (1966). On finite graphs of diameter 2. *Nanta Mathematica*, 1, 72-75.
- Pinski, G. & Narin, F. (1976). Citation influences for journal aggregates of scientific publications: theory, with applications to the literature of physics. *Information Processing and Management*, 12, 297-312.
- Plesník, J. (1984). On the sum of all distances in a graph or digraph. *Journal of Graph Theory*, 8, 1-21.
- Price, D. J. de Solla (1965). Networks of scientific papers. *Science*, 149, 510-515.
- Pritchard, A. (1984). *On the structure of information transfer networks*. M. Phil. thesis, School of Librarianship, Polytechnic of North London (UK).
- Randić, M. (1975). On characterization of molecular branching. *Journal of the American Chemical Society*, 97, 6609-6615.
- Rivlin, E., Botafogo, R., & Shneiderman, B. (1994). Navigating in hyperspace: designing a structure-based toolbox. *Communications of the ACM*, 37(2), 87-96.

- Rouvray, D.H. (1986). Predicting chemistry from topology. *Scientific American*, 255(3), 36-43.
- Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2), 97-108.
- Shepherd, M.A., Watters, C.R. , & Cai, Y. (1990). Transient hypergraphs for citation networks. *Information Processing and Management*, 26, 395-412.
- Trinajstić, N. (1992). *Chemical graph theory*. Boca Raton: CRC Press.
- Wiener, H. (1947). Structural determination of paraffin boiling points. *Journal of the American Chemical Society*, 69, 17-20.
- Wilson, R. J. (1972). *Introduction to Graph Theory*. London: Longman.