

Matrix comparison, Part 1:
Motivation and important issues for measuring the resemblance
between proximity measures or ordination results

This is a preprint of an article accepted for publication in
Journal of the American Society for Information Science and Technology© (forthcoming)

Jesper W. Schneider & Pia Borlund

jws; pb@db.dk

Department of Information Studies,
Royal School of Library and Information Science
Sohngaardsholmsvej 2, 9000 Aalborg, Denmark
Phone: +45 98 15 79 22; Fax: +45 98 15 10 42

Abstract

The present two-part article introduces matrix comparison as a formal means for evaluation purposes in informetric studies such as co-citation analysis. In this, the first part, the motivation behind introducing matrix comparison to informetric studies, as well as two important issues influencing such comparisons, are introduced and discussed. The motivation is spurred by the recent debate on choice of proximity measures and their potential influence upon clustering and ordination results. The two important issues discussed in the present first part are matrix generation and the composition of proximity measures. The present part of the article demonstrates that the approach to matrix generation for the same data set, that is how data is represented and transformed in a matrix, evidently determines the ‘behaviour’ of proximity measures. Two different matrix generation approaches, will therefore in all probability, lead to different proximity rankings of objects, which further lead to different ordination and clustering results for the same set of objects. Further, this part of the article also demonstrates that a resemblance in the composition of formulas indicates whether two proximity measures may produce similar ordination and clustering results. However, as shown in the case of the angular correlation and cosine measures, a small deviation in otherwise similar formulas, can lead to different rankings depending on the contour of the data matrix transformed. Eventually, the ‘behaviour’ of proximity measures, that is whether they

produce similar rankings of objects, is more or less data-specific. Consequently, we recommend the use of empirical matrix comparison techniques for individual studies in order to investigate the degree of resemblance between proximity measures or their ordination results. Part two of the article introduces and demonstrates two related statistical matrix comparison techniques the Mantel test and Procrustes analysis, respectively. These techniques can compare and evaluate the degree of monotonicity between different proximity measures or their ordination results. As such, the Mantel test and Procrustes analysis can be used as statistical validation tools in informetric studies and thus help choosing suitable proximity measures.

1. Introduction

The purpose of this two part article is to introduce matrix comparison as a formal means for evaluation purposes in informetric studies. The present part one of the article outlines the background and motivation for introducing matrix comparison for such purposes. The succeeding second part of the article proceeds by introducing and demonstrating two related statistics of matrix comparison for use in informetric studies, the Mantel test and the Procrustes analysis, respectively.

Recently, a study by Ahlgren, Jarneving, and Rousseau (2003) has questioned the mathematical validity of the product moment correlation coefficient, r , when used as a similarity measure in author co-citation analyses (ACA). The main criticism of r is its sensitivity to zero vector element values, as the coefficient is based upon moments around the mean (e.g., Anderberg, 1973). The authors argue that r is not the optimal choice of similarity measure if the main objective is to generate maps of authors with similar co-citation patterns, as well as to explain changes in such maps over time (Ahlgren, Jarneving, & Rousseau, 2003). As alternatives, the chi-square distance and the cosine similarity measures are suggested. The study has provoked a number of reactions. Bensman (2004) favours the use of r for a number of reasons, among them that it is a measure embedded in inferential statistics and therefore a tool for additional and more powerful analysis. As a basis for prediction, r unquestionably requires a normal distribution. Yet, as Carroll states "...prediction is something you do after you have discovered relationships between variables" (Carroll, 1961, p. 349). In its function as a descriptive proximity measure of relationship between variables, no assumptions are necessary for the computation of r (Carroll, 1961). However, the interpretation of its meaning certainly depends upon the extent to which the data conform to an appropriate statistical model. Even so, Leydesdorff (2005) and Leydesdorff and Bensman (2006)

investigate the application of logarithmic transformations of values prior to the use of r . Their findings indicate no benefit in the mapping results when data undergo logarithmic transformations. White (2003a, p. 1250), however, criticises Ahlgren, Jarneving, and Rousseau (2003) for constructing an idiosyncratic example remote from both theory and practice of ACA. In this respect theory and practice of ACA means the approach developed at Drexel University (White & Griffith, 1981; McCain, 1990). White (2003a) defends the use of r in ACA with a pragmatic argument that actual mapping results are more important than rigid mathematical premises. From this pragmatic perspective, and by use of Ahlgren, Jarneving and Rousseau's (2003) own data, White (2003a) argues that r , cosine, and the chi-square distance measures lead to *almost* identical ordination and clustering results¹. Minor differences between the results do emerge, but according to White (2003a), they have no or little practical impact on the interpretability of the macro structures in relation to ACA studies. One could argue that in more exploratory cases, where macro structures are not perceived *a priori*, minor differences may influence the interpretability of ordination and clustering results. Finally, Leydesdorff and Vaughan (2006) shift the attention of the current debate by emphasizing the importance of matrix generation. The authors demonstrate the fundamental differences between asymmetric occurrence matrices and symmetric co-occurrence matrices (Leydesdorff & Vaughan, 2006). The authors argue that a symmetric matrix of co-occurrence counts is per se a *proximity matrix*, where proximity is a number that "...indicates how similar or how different two objects are, or are perceived to be" (Kruskal & Wish, 1978, p. 7)². Accordingly, a derived proximity measure, as r , should not be applied to such a matrix as the counts already express proximities between objects. Instead, proximity measures should be applied to the asymmetric occurrence matrix in order to transform it into a symmetric co-occurrence matrix of derived proximities (Leydesdorff & Vaughan, 2006). Accordingly, co-occurrence counts are generated and normalized in the transformation process according to the composition of the involved proximity measure. This is the traditional matrix transformation technique used in the vector space model in information retrieval (IR) to compute a measure of vector profile similarity between columns (or rows) of a term-document matrix (occurrence matrix) (Salton, 1968). Equally, in document co-citation analyses (DCA) the basis is an asymmetric occurrence matrix of references and documents, which is transformed, according to a proximity measure, into a symmetric matrix of

¹ *Ordination* is the collective term for multivariate statistical techniques that arrange a set of data with respect to one or more axes in a multidimensional space. Clustering or classification is the act of arranging objects into groups (for definitions see e.g. <http://ordination.okstate.edu/glossary.htm>.)

² The general term proximity refers to *dissimilarity*, *distance* or *similarity*. Two objects are 'close' when their dissimilarity or distance is small or their similarity is large.

derived proximities (Morris, 2005). This matrix transformation technique is very different from the ‘Drexel-style’³ traditionally used in ACA studies. The ‘Drexel-style’ ACA uses online retrieval procedures, where co-citation counts are obtained by searching the cited author field in ISI’s citation indices (e.g., White & Griffith, 1981; McCain, 1990). Subsequently, a symmetric matrix of directly obtained co-citation counts is constructed and r is computed from this matrix. According to Leydesdorff and Vaughan (2006), the latter approach is questionable, as derived measures should be computed from an asymmetric occurrence matrix and not from a corresponding symmetric proximity matrix. However, if the aim is the construction of a proximity matrix based on direct co-citation counts, then the ‘Drexel-style’ sampling procedure is very suitable. As stressed by Leydesdorff and Vaughan (2006), direct co-citation counts are indeed proximities, and White (2003b) has demonstrated that they are very suitable as input for spatial ordination techniques or network scaling.

The present dispute reopens the discussion about different proximity measures’ affect upon ordination and clustering results in informetric studies. From this follows, which measures to choose in such studies, how to validate them, and how to compare ordination and clustering results based on different measures for the same data set. As noted by Oberski (1988) and more recently by Leydesdorff (2005), it remains troublesome that one has such a wealth of both proximity measures and clustering and ordination algorithms available that one is able to generate almost any representation from a set of data. It is indeed a general problem acknowledged in many fields where ordination and clustering techniques are applied for multivariate data analysis (e.g., Williams, Clifford, & Lance, 1971; Everitt, 1979; Pielou, 1984; Jackson, Somers, & Harvey, 1989; Legendre & Legendre, 1998). According to Jackson, Somers, and Harvey (1989), the objective character of the analyses is compromised by the subjective choices of proximity measure as well as ordination and clustering techniques.

General studies that compare proximity measures and their influence upon ordination and clustering are fairly common (e.g., Austin, 1976; Baroni-Urbani & Buser, 1976; Fasham, 1977; Hubálek, 1982; Gower & Legendre, 1986; Kenkel & Orloci, 1986). In addition, several specific studies have investigated the composition and use of proximity measures in IR and informetrics (e.g., Jones & Curtice, 1967; McGill, Koll, & Noreault, 1979; Jones & Furnas, 1987; Leydesdorff, 1987; Oberski, 1988; Wang, Wong, & Yao, 1992; Ellis, Furner-Hines, & Willett, 1993; Gmür, 2003; Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2006). Characteristically, these

³ The term ‘Drexel-style’ is taken from White (2003a, p. 1251).

studies have generated different results, prompting a general acceptance that the ‘behaviour’ of proximity measures is data-specific. As a result, the choice of a proximity measure is largely subjective and often based on tradition or on posterior criteria such as the ‘interpretability’ of the results, rather than theory, mathematical validity, or comparative empirical investigations. As Gordon (1987, p. 127) suggests, “...human ingenuity is quite capable of providing a *post hoc* justification of dubious classifications”.

Accordingly, numerous measures are available and some produce comparable results irrespective of the analytical technique applied, but studies usually do not justify their preference for any one measure in particular. The qualities and characteristics of proximity measures need to be better understood, so that the most appropriately ones can be employed. In fact, White (2003a) encourages more in-depth comparative studies of proximity measures in order to better understand their affect upon ordination and clustering results. Motivated by the revived debate, the present article takes up this encouragement.

The purpose of this two-part article is to introduce two related statistical techniques for matrix comparison. Matrix comparison is a practical tool that investigates the degree of empirical resemblance between two or more proximity matrices. In part two, we introduce two techniques widely used in numerical ecology, the Mantel test (Mantel, 1967; Mantel & Valand, 1970; Dietz, 1983) and Procrustes analysis (Schönemann, 1966; Schönemann & Carroll, 1970; Gower, 1971). We demonstrate the application of the Mantel tests for direct comparison of different proximity measures, and the Procrustes analysis for comparison of ordination results based on different proximity measures. Empirical comparison of two proximity measures or two ordination results is a useful statistical tool for validation in informetric studies. In the case of the Mantel test, the degree of resemblance between two measures forecast their potentially different affect upon the eventual ordination and clustering results. In principle, two measures with a very strong resemblance most likely produce identical ordination or clustering results, thus, the choice of measure between the two becomes less important. Alternatively, or as a supplement, Procrustes analysis compares the actual ordination results without investigating the underlying proximity measures, but by matching two configurations of the same n objects in a multidimensional space. We demonstrate the application of the Mantel test and Procrustes analysis on a small hypothetical data set in order to show their simplicity, flexibility, as well as different emphasis.

The present first part of the article addresses two important issues that influence the empirical comparison of two or more proximity matrices. The first issue concerns the generation of matrices.

This issue is especially important in relation to co-citation studies, as several different approaches are employed. Different matrix generation approaches, for the same sample of objects using the same proximity measure, will undoubtedly produce different ordination and clustering results. This important issue may seem obvious, but it is hardly ever discussed in co-citation studies. As a result, we find it important to emphasize the role of matrix generation in co-citation studies.

The second issue concerns the composition of proximity measures. Basic knowledge of the composition of proximity measures is an advantage, not only when choosing a measure for analysis, but also when choosing alternative measures for validation purposes. A resemblance in the composition of formulas indicates whether two measures may produce similar ordination and clustering results. Hence, we find it important to emphasize some generic compositional characteristics of proximity measures.

The article is organized into the following sections. Section 2 introduces the process of matrix generation and discusses its influence upon matrix comparison. Section 3 introduces the composition of proximity measures and discusses its influence upon matrix comparison. Finally, section 4 is a brief summary of the issues discussed in the present first part of the article, which serves as a transition to the second part of the article that introduces and demonstrates the two related matrix comparison techniques, the Mantel test and Procrustes analysis, respectively.

2. Matrix generation

The present section addresses the issue of matrix generation and its importance in relation to co-citation studies. As stated above, different matrix generation approaches, for the same sample of objects using the same proximity measure, will undoubtedly produce different ordination and clustering results. While this may seem obvious, it is hardly ever discussed in co-citation studies.

A co-occurrence analysis appears to be a traditional bivariate analysis; however, co-occurrence analyses are most often applied to multivariate data sets and are therefore treated as multivariate analyses (e.g., Lattin, Carroll, & Green, 2003). Multivariate data analyses are essentially algebraic operations on vectors and matrices. For that reason, objects and attributes are represented as a multivariate data matrix. We let m denote the number of attributes and n denote the number of objects in the sample. Each column vector of the matrix is an attribute (i.e. a measured property of the objects in the sample), and each row vector corresponds to an object (i.e. a list of all measured properties for one object in the sample). This is an asymmetric $n \times m$ multivariate data matrix, \mathbf{X} ; the elements x_{ij} , of which give the values of m attributes for each of n objects. The data matrix is

termed a two-way two-mode matrix indicating that the rows and columns correspond to different things. According to most multivariate statistical textbooks, a multivariate data matrix is the traditional basis for a multivariate analysis (e.g., Krzanowski & Marriott, 1994; Everitt & Rabe-Hesketh, 1997; Lattin, Carroll, & Green, 2003; Cox, 2005). Indeed, as demonstrated by Krauze and McGinnis (1979) and elaborated on by Egghe and Rousseau (1990), informetric data such as publication, reference, and citation data are most conveniently described by matrix algebra and initially represented as an $n \times m$ data matrix.

Some multivariate techniques, such as multidimensional scaling (MDS) and cluster analysis employ a proximity matrix as their input. Most commonly, such an input proximity matrix is *derived* from the traditional $n \times m$ data matrix, \mathbf{X} , by some suitable proximity measure. The transformation of \mathbf{X} by use of a proximity measure results in an $n \times n$ symmetric matrix of inter-object proximities. Such a proximity matrix is termed a two-way one-mode matrix indicating that both rows and columns refer to the same objects (Everitt & Rabe-Hesketh, 1997, p. 5). Alternatively, a proximity matrix can also be generated *directly*; for example from physical distances such as mileage between cities, by collecting subjective proximity judgements from people, or indeed co-occurrence data such as actual co-citation counts (Feger & De Boeck, 1993; Everitt & Rabe-Hesketh, 1997). Such data are perceived to be proximities, which can be added directly to a proximity matrix (e.g., Leydesdorff & Vaughan, 2006).

Other multivariate techniques, such as principal components analysis and exploratory factor analysis seek to reduce dimensions or uncover latent factors in the $n \times m$ multivariate data matrix \mathbf{X} (e.g., Lattin, Carroll, & Green, 2003). The underlying models for the two techniques are different, though they share a common solution that focuses on the decomposition of the covariance or correlation matrix of \mathbf{X} . This implies that the data matrix *must* be transformed into a specific type of proximity matrix, i.e. covariance matrix or its standardized complement the correlation matrix. This is a conventional conception of multivariate data analysis found in numerous textbooks (e.g., Lattin, Carroll, & Green, 2003; Cox, 2005). Nevertheless, the methodical consensus is not so apparent when it comes to co-occurrence analyses in informetric studies. In the following subsection, we exemplify the methodical discrepancy between the paradigmatic approaches to matrix generation within DCA and ACA.

2.1 The case of document and author co-citation analysis

The paradigmatic approach to DCA, pioneered for example by Small and colleagues, in principle follows the conventional procedure to matrix generation and subsequent multivariate analyses outlined above (Small, 1973; Small & Griffith, 1974; Small & Greenlee, 1980; Small & Sweeny, 1985; Small, Sweeney, & Greenlee, 1985). Cited documents are the objects under study, thus citing documents become the attributes of the study. Notice that the measured properties of the objects are occurrences and the data are always binary. The $n \times m$ data matrix is therefore a binary occurrence matrix. A simple multiplication of the occurrence matrix with its transpose produces a matrix of actual co-occurrence counts. However, much more commonly, the relative co-citation strength between objects is *derived* by transforming the occurrence matrix by some proximity measure to an $n \times n$ symmetric co-occurrence matrix of proximities. This is a so-called ‘global approach’, where the relative co-citation strength between objects is based on resemblances in vector profiles for the cited documents (Cronbach & Gleser, 1953; Ahlgren, Jarneving, & Rousseau, 2003). Traditionally, the resulting proximity matrix is the basis for subsequent cluster analysis (Gmür, 2003).

As indicated in the introduction, the paradigmatic approach to ACA, developed at Drexel University, is very different from that of DCA (e.g., McCain, 1990; White, 2003a). In ACA, the conventional preceding step where the $n \times m$ data matrix is generated is omitted. Instead, actual co-citation counts between objects are obtained through online retrieval, and a proximity matrix is generated directly from these counts. As stated above, it is entirely permissible to generate a matrix based on directly obtained proximities (e.g., Kruskal & Wish, 1978; Cox & Cox, 2000). It is therefore important to emphasise that such directly obtained values are proximities *per se*, and according to Leydesdorff and Vaughan (2006), should be applied as such. The latter has been the exception and not the rule in ACA (for exceptions, see for example Persson (1994) and White (2003b)). In fact, the ‘Drexel-style’ extends the conventional *modus operandi* by transforming the directly obtained proximity matrix of co-citation counts into an equivalent proximity matrix – the correlation matrix r . According to McCain (1990, p. 436) the advantage of this transformation is the possibility of comparing ‘co-citation profiles’ and not just actual co-citation counts between a pair of authors. Further, McCain (1990) emphasizes that the correlation matrix also makes it possible to use factor analysis in ACA, where indeed a correlation matrix is mandatory, as noted above.

Obviously, it is possible to transform a symmetric proximity matrix into an equivalent one. This is nevertheless an unorthodox procedure, which to our knowledge is not commonly described in textbooks on multivariate analyses. As argued by Leydesdorff and Vaughn (2006), derived proximities are normally obtained from a transformation of an $n \times m$ data matrix, and not from a corresponding symmetric proximity matrix. This is also the common practice in factor analysis, where the correlation matrix is derived from a data matrix and not an equivalent proximity matrix (e.g., Mulaik, 1972). Actually, a linear transformation of a symmetric proximity matrix is not straightforward. The transformation causes a fundamental problem in relation to the interpretation and treatment of the diagonal values. A symmetric matrix is one where $x_{ij} = x_{ji}$; if $x_{ij} \neq x_{ji}$ then the matrix is asymmetric. It is customary to transform asymmetric matrices into symmetric ones; this also holds true in the special case where the asymmetric matrix is square, such as transaction matrices used in informetric studies (Price, 1981). Here the diagonal values are meaningful and cause no immediate problems in relation to the subsequent transformation (Price, 1981; Noma, 1982). Likewise, the diagonal values in a traditionally derived proximity matrix are also meaningful⁴. Diagonal values in a proximity matrix, although meaningful, are indifferent to MDS or cluster analysis. Only one set of the symmetric off-diagonal values are needed as input. As ‘Drexel-style’ ACA prescribes a linear transformation of a symmetric proximity matrix, attention must be given to the diagonal values in the row or column vectors. Several solutions have been investigated, and it seems that the preferred option is to treat the diagonal values as missing data (McCain, 1990; White, 2003a). Missing data beyond doubt causes some loss of information in the matrix and therefore likely influence the ensuing ordination or clustering results. White (2003a, p. 1250), nevertheless, asserts that the treatment of diagonal values is a minor problem. This may be true, depending on the data set at hand, but the generic problem arises due to the unorthodox matrix generation (transformation) approach and could be avoided if a more traditional procedure was applied.

We do not argue against the ‘Drexel-style’ ACA, but we stress that it is unorthodox seen in relation to traditional approaches to multivariate analyses. Consequently, we emphasize that comparison of two proximity matrices, or two ordination results, for the same set of objects but based on different approaches to matrix generation most certainly will produce markedly different results.

⁴ In a similarity matrix, the diagonal values are 0, i.e. the object’s similarity with itself; and in a dissimilarity or distance matrix, the diagonal values are 0, i.e. the object’s dissimilarity or distance to itself.

Consider the following hypothetical $n \times m$ occurrence matrix, \mathbf{X} , and a corresponding hypothetical $n \times n$ proximity matrix of co-occurrence counts, \mathbf{P} , for the same set of objects.

TABLE 1. Occurrence matrix \mathbf{X} .

$$\mathbf{X} = \begin{bmatrix} A & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ B & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ C & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ D & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ E & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

TABLE 2. Co-occurrence matrix \mathbf{P} .

$$\mathbf{P} = \begin{bmatrix} & A & B & C & D & E \\ A & & 2 & 3 & 4 & 2 \\ B & & & 2 & 2 & 3 \\ C & & & & 4 & 2 \\ D & & & & & 4 \\ E & & & & & \end{bmatrix}$$

Given a self consistent data source, the actual co-occurrence counts between objects in the two matrices are identical. The co-occurrence matrix is simply \mathbf{X} multiplied with its transpose, $\mathbf{X}\mathbf{X}^T$. The diagonal in \mathbf{P} is left out as the elements are treated as missing values.

The *derived* correlation matrices, \mathbf{C}_X and \mathbf{C}_P , as well as their corresponding non-metric MDS solutions are shown below in Tables 3 and 4, and Figure 1 and 2, respectively. The objects and proximity measure chosen are the same, but the matrix generation procedures, and thus the representation of the objects, differ.

TABLE 3. Correlation matrix \mathbf{C}_X .

$$\mathbf{C}_X = \begin{bmatrix} & A & B & C & D & E \\ A & 1.00 & -0.20 & 0.20 & 0.41 & -0.20 \\ B & & 1.00 & -0.20 & -0.41 & 0.20 \\ C & & & 1.00 & 0.41 & -0.20 \\ D & & & & 1.00 & 0.41 \\ E & & & & & 1.00 \end{bmatrix}$$

TABLE 4. Correlation matrix \mathbf{C}_P .

$$\mathbf{C}_P = \begin{bmatrix} & A & B & C & D & E \\ A & 1.00 & -0.87 & 1.00 & 0.41 & 0.58 \\ B & & 1.00 & -0.87 & 0.11 & 0.06 \\ C & & & 1.00 & 0.41 & 0.58 \\ D & & & & 1.00 & -0.52 \\ E & & & & & 1.00 \end{bmatrix}$$

Clearly, in the present diminutive example, object correlations derived from different matrix generation procedures, i.e., $n \times m$ ‘occurrence profiles’, \mathbf{C}_X , and $n \times n$ ‘co-occurrence profiles’, \mathbf{C}_P , diverge considerably.

Notice that objects A and C have a perfect linear correlation in \mathbf{C}_P (Table 4). Accordingly, the non-metric MDS solution, shown below in Figure 2, assigns identical coordinates for the two objects, as the stress value is zero. Hence, A and C are superimposed on top of each other.

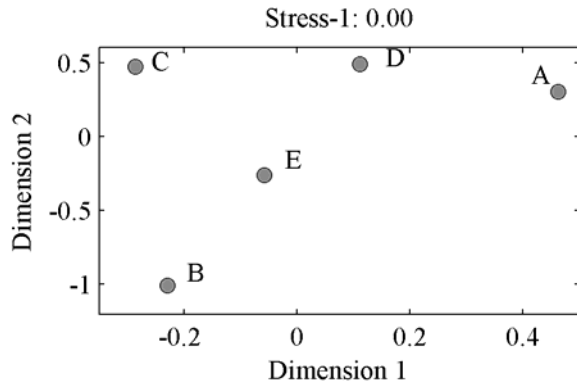


FIG. 1. Non-metric MDS solution of C_X .

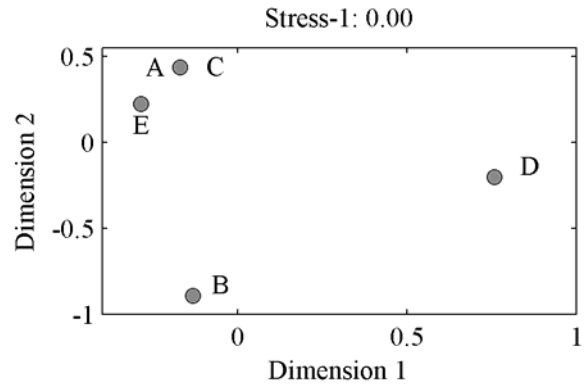


FIG. 2. Non-metric MDS solution of C_P .

Evidently, the ‘behaviour’ of proximity measures is dependent on how data are represented in a matrix. In all probability, two different matrix generation procedures lead to different proximity rankings of objects, which further lead to different ordination and clustering results. This is important to bear in mind when conducting multivariate analyses, and especially in relation to matrix comparison, or comparison of ordination results, the focus of the present article. Accordingly, comparisons of proximity matrices or ordination results are influenced by the way the comparable matrices are generated.

Obviously, data sampling is a fundamental issue in relation to the above presented discrepancy in matrix generation within co-citation analyses. While the issue of sampling theory and its application in co-citation analyses is very important (e.g., Åström, 2002; Nicolaisen, 2004), it is not the focus of the present article. However, we do wish to emphasise that sampling procedures eventually influence matrix generation. Today, a common practice in DCA is to retrieve and download a sample of citing documents, typically delimited by time, subject, and/or journals (Gmür, 2003). Highly cited documents, identified by some arbitrary threshold, define the objects under study, and citing documents become the attributes for the study. This creates the previously mentioned $n \times m$ data matrix. Conversely, the common practice in ACA is to select a sample of authors in a domain based on their citation counts, and most often supplemented subjectively by a number of other ‘influential’ authors in the domain. Comparable to DCA, time, subject, and/or journals most often delimit the sample. The chosen sample of authors is paired, and their co-citation counts are retrieved online. Data are not downloaded, instead proximities are retrieved directly from the citation indices. This creates the aforementioned $n \times n$ proximity matrix. It thus appears that sampling strategies for the paradigmatic approaches to DCA and ACA have evolved indiscriminately, more affected by access to citation and co-citation data than traditional sampling

theory or traditional multivariate procedures. This is especially the case with the ‘Drexel-style’ ACA, whereas DCA to a larger degree at least resembles traditional multivariate procedures.

We think it is important to reconsider and investigate the paradigmatic sampling and matrix generation strategies traditionally applied in co-citation studies. For example, with the relatively easy access to large downloadable data sets via Web of Science®, the basis for ACA may as well be an $n \times m$ data matrix. Such an approach would be more in line with traditional procedures of multivariate analyses. However, we need to verify in general whether diverse matrix generation techniques produce substantially different results in domain visualization studies. As we will demonstrate in part two of this article, the Mantel and the Procrustes statistics are two techniques that can be used for such verification purposes.

3. The composition of proximity measures and its relation to monotonicity

The previous section stresses the influence of matrix generation in relation to comparison of proximity measures. Another important aspect to consider in this respect is the composition of proximity measures. Obviously, the composition of proximity measures, strongly influence their potential degree of mutual *monotonicity*. Different proximity measures will treat a data set differently, since measures emphasize, or are sensitive to, the ‘contour’ of the data in a matrix (Jones & Furnas, 1987). The contour of the data refers to the total variability in a matrix and defines the ‘shape’ of that matrix (Cronbach & Gleser, 1953; Jones & Furnas, 1987). Theoretically, functions that resemble each other in all probability have a high degree of monotonicity between them. Two measures are completely monotonic to each other if the ranking of all measurements of proximity between pairs of objects in a specific set is the same using one measure as it is using the other (Hubálek, 1982; Ellis, Furner-Hines, & Willett, 1993). In principle, universal monotonicity can be demonstrated by simple algebraic comparison of two formulas. Nonetheless, to a certain degree a monotonic relationship is relative depending of the specific contour of the data in a matrix under investigation. Small deviations in what seem comparable formulas may affect the degree of monotonicity. We therefore recommend that the comparison of proximity measures should be done empirically for a given data set by use of matrix comparison, as demonstrated in part two of this article.

It is not the scope of the present article to introduce proximity measures and their algebraic compositions. We refer to excellent general overviews by Anderberg (1973), Sneath and Sokal (1973), and Gower and Legendre (1986), and for a specific library and information science

perspective, Ellis, Furner-Hines, and Willett (1993). The present purpose is mainly to accentuate that knowledge of the composition of proximity measures is essential when choosing a suitable measure for a given analysis, and subsequently for validation purposes, when the degree of monotonicity between the chosen measure and some other measures are investigated by use of matrix comparison⁵.

An almost endless number of proximity measures exist (e.g., Sneath & Sokal, 1973). Most proximity measures have a binary and a quantitative expression, and they can be described and classified in numerous ways. Sneath and Sokal (1973) demonstrate that the numerator of symmetric proximity measures contains one of two basic functions. For binary data, the basic functions are one of *match* or one of *mismatch* between pair wise elements in two object vectors. For quantitative data, the basic functions are one of *co-occurrence* or one of *difference* between pair wise elements in two object vectors. The latter quantitative functions are also known as the *inner product* and the *difference sum* (Ellis, Furner-Hines, & Willett, 1993). It is possible for either of these functions to be used on their own as proximity measures, their coefficients, however, vary in direct proportion with the number of pair wise elements compared. Accordingly, the functions do not have an upper limit, unless they are normalized by some function in the denominator.

In general, all symmetric proximity measures can be classified according to which of these two basic functions they contain in their numerator. Distance measures are based on a binary mismatching or quantitative difference function in the numerator (Sneath & Sokal, 1973). Similarity measures are based on a binary matching or quantitative co-occurrence function in the numerator (Sneath & Sokal, 1973). Similarity measures are sometimes further subdivided into correlation measures, which are based on a modification of the matching or co-occurrence functions in the numerator. Generally, measures based on the difference sum accentuate the difference in size or magnitude between elements in object profiles, whereas measures of similarity generally attend to the similarity in ‘shape’ between object profiles (Cronbach & Gleser, 1953; Jones & Furnas, 1987). Here ‘shape’ refers to the residual variability left in a profile, when size or magnitude differences are discarded (i.e., means and scatter according to Cronbach and Gleser (1953)). For example, in IR, arguments have been made that the direction of a document vector in term space approximates the *topic* of the document, while the length of a document vector is an indication of the *intensity* with which these topics are treated (Jones & Furnas, 1987). In IR the similarity in topic ‘shapes’ between two vectors are sought, whereas the intensity with which these topics are

⁵ The present discussion focuses on proximity measures applied to vector representations. However, as indicated by one reviewer a variety of index-based similarity measures exist as well.

treated is indifferent (Salton, 1968). This leads to the application of cosine as a measure of angle similarity and thus direction between vectors.

Consider two object vectors:

$$A = [3, 0, 0, 0, 0, 0, 0, 0, 0, 0],$$

$$B = [0, 0, 0, 0, 0, 0, 0, 0, 0, 4].$$

If we calculate the Euclidean distance measure between these two vectors, without normalization, we get a coefficient of five. Now consider another pair of object vectors:

$$C = [3, 2, 4, 0, 1, 2, 3, 1, 2, 0],$$

$$D = [0, 2, 4, 0, 1, 2, 3, 1, 2, 4].$$

Comparable to the first pair, A and B , the Euclidean distance coefficient between C and D is likewise five. Notice that A and B do not share any common attributes, while C and D share seven attributes. Intuitively, the pair C and D seems more similar than A and B , which do not share any attributes. This small example demonstrates that mismatch or missing attributes play an important role in distance measures.

The number of attributes measured on objects defines dimensionality. In high dimensions most vectors are sparse as they only contain a handful of non-zero attribute values. Accordingly, the presence of an attribute is a lot more important than the absence of an attribute. As a result, distance measures do not work as well in high dimension as in low dimensions (Jones & Furnas, 1987).

Conversely, similarity measures such as cosine (Ochiai, 1957; Salton, 1963) and Jaccard (Jaccard, 1901; Tanimoto, 1958) focus on the mutual presence of attributes rather than their mutual absence. In the example above, the cosine similarity between A and B is zero, whereas the cosine similarity between C and D is 0.76. These coefficients are much more in line with the intuitive understanding that A and B are not similar and C and D are seemingly more similar. Nonetheless, similarity measures such as cosine and Jaccard may also have problems in high dimensions. Object vectors become increasingly ‘sparse’ as the dimensionality increases, whereas, on average the similarity between objects in a hyperspace is very low. Presence data in high dimensions are most likely scattered in space, with few regions of density (Scott, 1992). Accordingly, numerous similarities between objects, especially outside or between these density regions, cannot be trusted as the similarity between pairs of objects on average is low (Scott, 1992). This is known as the ‘curse of dimensionality’ (Bellman, 1961). As a result, data in higher dimensions are more difficult to interpret for ordination and cluster techniques. This is the main reason for the application of

dimensionality reduction techniques. Consider the following, O is similar to P , and P is similar to Q , but this does not necessarily imply that O is similar to Q . Nonetheless, single link clustering may join these objects despite their low similarities (Anderberg, 1973). An often-mentioned issue in this respect is whether similarity measures behave as Euclidean metrics (e.g., Gower & Legendre, 1986). However, too much should not be made of this. As thoroughly demonstrated by Gower and Legendre (1986), a simple monotonic transformation often restores the required Euclidean properties for most measures.

Consequently, we expect measures based on the difference sum and the inner product, respectively, to show the largest deviation in monotonicity between them on a given data set. While, in general the numerator distinguishes between distance and similarity measures, the composition of the denominator distinguishes the individual distance or similarity measures from each other (Sneath & Sokal, 1973). The denominator is a normalization function. Traditionally, this normalization ensures that the proximity coefficients remain within a specific range, such as that bounded by 0 and 1, or by -1 and +1. Yule (1912) compared a range of similarity measures with the normalization choices of arithmetic mean, geometric mean or median. Yule (1912) found that all forms of average are measures of analogous properties, but they do not provide the same values. Accordingly, Yule (1912) concludes that the various similarity measures differ precisely the same way.

Gower and Legendre (1986) make the important point that the nature of the data should strongly influence the choice of measure. The case of r is imperative in this respect. Consider the quantitative expressions of the cosine and r similarity measures:

$$\text{cosine} = \frac{\sum(x_i \cdot y_i)}{\sqrt{\sum(x_i)^2 \cdot \sum(y_i)^2}} \quad (1)$$

$$r = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} \quad (2)$$

The composition of these two measures is almost identical. The product moment correlation is the cosine of the angle between two vectors, where these vectors are centred relative to a zero mean, and reduced to have unit variance, that is vectors of unit length. In principle, the numerators in cosine and r contain an inner product function and the denominators contain a normalization function based on the Euclidean length of the compared vectors. The numerator in cosine is the inner product between two vectors of *original* values, whereas the numerator in r is the inner

product between two vectors of *centred* values. The latter is the product moment (covariance) of x and y . The denominator in cosine is the product of Euclidean lengths for x and y , respectively. The denominator in r is the product of moment normalized Euclidean lengths for x and y . Accordingly, the essential difference between the two measures is that cosine is based on original values (deviations from the origin) while r is based on centred values (deviations about the mean). Notice that r contains two normalization functions. First, comparable vectors are moment normalized to centre their element values relative to a zero mean, and secondly, like the cosine measure, the moment normalized vectors are then further normalized to unit length. According to Jones and Furnas (1987, p. 434), the use of moment normalization in r is a limitation and may result in loss of potentially useful information. This point is also stressed by Anderberg (1973, p. 74), who argues that r is less discriminating than the cosine measure. In fact, the argument goes back to Cronbach and Gleser (1953), who criticized the use of r in certain contexts due to the abovementioned neglect of important information. Accordingly, centred values are less informative than original values. Further, as indicated by Ahlgren, Jarneving and Rousseau (2003), the computation of centred values is sensitive to zero element values in the vectors. Theoretically, the informativeness of r is therefore to a certain degree bounded to its sensitivity to zero values. In fact, Jones and Furnas (1987) question the validity of r as a similarity measure, while appraising its use for prediction in inferential statistic. This point is also addressed by Lattin, Carroll and Green (2003, p. 277), who note that r is not always an appropriate measure of similarity as correlation is a measure of covariance, which is a type of proximity but not necessarily of similarity. Consider two objects measured across four attributes $O = [1, 2, 1, 2]$ and $P = [1, 1, 2, 2]$. Intuitively, we would judge these objects highly similar, which is also the indication of the cosine coefficient of 0.9. Remember that r measures linear correlation, so across the four attributes O and P exhibits a correlation coefficient of 0.0. It is therefore important to keep in mind that r is especially vulnerable to the nature of a data set.

With this in mind, we still expect r and cosine to have a high degree of monotonicity between them for a given data set, as they both measure the shape similarity of object profiles. Nonetheless, the fact that r reflects centred values makes it sensitive to sparse matrices, as large number of zero values most likely influence the calculation of product moments. In principle, as dimensions increase and vectors become sparser, the monotonicity between r and cosine may to some extent decline. The latter is confirmed in several empirical investigations (McGill, Koll, & Noreault,

1979; Hubálek, 1982), where the degree of monotonicity between cosine and r varies in correlation levels between 0.7 and 1.

The choice of proximity measure should be guided by the type of variability that is deemed relevant to the application at hand. For example, a noticeable feature of $n \times m$ data matrices is their sparseness. One should therefore consider whether r is the right choice of measure for a given $n \times m$ data matrix. An angular similarity measure may still be the most pertinent choice, if ratios or direction come closest to estimating the type of proximity deemed relevant by the analyst (Anderberg, 1973). The latter seems the pertinent choice in current co-occurrence analyses (Salton, 1963; Ahlgren, Jarneving, & Rousseau, 2003; Leydesdorff & Vaughan, 2006). However, as Gower and Legendre (1986) conclude, it is not possible to give a definitive answer to what measure is best to use in all circumstances. Accordingly, we call upon the use of empirically based matrix comparison as a means to evaluate the chosen proximity measure. High degrees of monotonicity with other relevant measures indicate that clustering and ordination results will not differ markedly between these measures. Depending on the purpose of evaluation, one should carefully consider which measures to compare. Measures that resemble each other in their composition most likely have high degrees of monotonicity between them. A comprehensive validation, where one seeks a confirmation of an ordination result, should compare proximity measures meaningful to the study at hand, and preferably ones that deviate to some degree in their composition.

4. Brief summary

The present article outlines the background and motivation for introducing matrix comparison in informetric studies. Section two and three addresses two important issues that influence the empirical comparison of two or more proximity matrices. The first issue concerns the generation of matrices. This issue is hardly ever discussed in co-citation studies even though several different approaches are employed. Different matrix generation approaches, for the same sample of objects using the same proximity measure, will undoubtedly produce different ordination and clustering results. We therefore find it important to emphasize the role of matrix generation in co-citation studies.

The second issue concerns the composition of proximity measures. Basic knowledge of the composition of proximity measures is an advantage, not only when choosing a measure for analysis, but also when choosing alternative measures for validation purposes. A resemblance in the composition of formulas indicates whether two measures may produce similar ordination and

clustering results. We therefore find it important to emphasize some generic compositional characteristics of proximity measures. However, the ‘behaviour’ of proximity measures is eventually data-specific. Consequently, we recommend the use of empirical matrix comparison techniques in order to investigate the degree of resemblance between proximity measures or their ordination results. With this knowledge in mind we proceed to the actual techniques of empirical matrix comparison. Part two of this article introduces and demonstrates two related matrix comparison techniques, the Mantel test and Procrustes analysis.

References

- Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6), 550-560.
- Anderberg, M. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Austin, M. P. (1976). Performances of four ordination techniques assuming three different nonlinear response models. *Vegetatio*, 42, 11-21.
- Baroni-Urbani, C., & Buser, M. W. (1976). Similarity of Binary Data. *Systematic Zoology*, 25(3), 251-259.
- Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.
- Bensman, S. J. (2004). Pearson's r and author cocitation analysis: A commentary on the controversy. *Journal of the American Society for Information Science and Technology*, 55(10), 935-935.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 26(4), 347-372.
- Cox, T. F. (2005). *An Introduction to Multivariate Data Analysis*. London: Hodder Arnold.
- Cox, T. F., & Cox, M. A. A. (2000). *Multidimensional Scaling* (2 ed.). Boca Raton: Chapman & Hall/CRC.
- Cronbach, L. J., & Gleser, G. C. (1953). Assessing Similarity between Profiles. *Psychological Bulletin*, 50(6), 456-473.
- Dietz, E. J. (1983). Permutation Tests for Association between two Distance Matrices. *Systematic Zoology*, 32(1), 21-26.
- Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics - Quantitative Methods in Library, Documentation and Information Science*. Amsterdam: Elsevier.
- Ellis, D., Furner-Hines, J., & Willett, P. (1993). Measuring the degree of similarity between objects in text retrieval systems. *Perspectives in Information Management*, 3(2), 128-149.
- Everitt, B. S. (1979). Unresolved Problems in Cluster-Analysis. *Biometrics*, 35(1), 169-181.
- Everitt, B. S., & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. London: Arnold.

- Fasham, M. J. R. (1977). Comparison of Nonmetric Multidimensional-Scaling, Principal Components and Reciprocal Averaging for Ordination of Simulated Coenoclines, and Coenoplanes. *Ecology*, 58(3), 551-561.
- Feger, H., & De Boeck, P. (1993). Categories and concepts: introduction to data analysis. In J. van Mechelen, J. Hampton, R.S. Michalski & P. Theuns (Eds.), *Categories and concepts: theoretical views and inductive data analysis*. London: Academic Press.
- Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57(1), 27–57.
- Gordon, A. D. (1987). A Review of Hierarchical-Classification. *Journal of the Royal Statistical Society, A 150*, 119-137.
- Gower, J. C. (1971). Statistical methods for comparing different multivariate analyses of the same data. In J. R. Hodson, D. G. Kendall & P. Tautu (Eds.), *Mathematics in the Archaeological and Historical Sciences* (pp. 138-149). Edinburgh: Edinburgh University Press.
- Gower, J. C., & Legendre, P. (1986). Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, 3(1), 5-48.
- Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews of the Cambridge Philosophical Society*, 57, 669-689.
- Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin et la Société Vaudoise des Sciences Naturelles*, 37, 241-272.
- Jackson, D. A., Somers, K. M., & Harvey, H. H. (1989). Similarity Coefficients - Measures of Co-Occurrence and Association or Simply Measures of Occurrence. *American Naturalist*, 133(3), 436-453.
- Jones, P. E., & Curtice, R. M. (1967). A framework for comparing term association measures. *American Documentation*, 18, 153-161.
- Jones, W. P., & Furnas, G. W. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science*, 38(6), 420-442.
- Kenkel, N. C., & Orloci, L. (1986). Applying Metric and Nonmetric Multidimensional-Scaling to Ecological-Studies - Some New Results. *Ecology*, 67(4), 919-928.
- Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2), 251-263.

- Krauze, T. K., & McGinnis, R. (1979). Matrix Analysis of Scientific Specialties and Careers in Science. *Scientometrics*, 1(5-6), 419-444.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills: Sage Publications.
- Krzanowski, W. J., & Marriott, F. H. C. (1994). *Multivariate Analysis, Part 1: Distributions, ordination and inference*. London: Edward Arnold.
- Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing Multivariate Data*. Pacific Grove, CA: Brooks/Cole - Thompson Learning.
- Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2 ed.). Amsterdam: Elsevier.
- Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, 11, 281-320.
- Leydesdorff, L. (2005). Similarity measures, author cocitation analysis, and information theory. *Journal of the American Society for Information Science and Technology*, 56(7), 769-772.
- Leydesdorff, L., & Bensman, S. (2006). Classification and powerlaws: The logarithmic transformation. *Journal of the American Society for Information Science and Technology*, 57(11), 1470-1486.
- Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence Matrices and their Application in Information Science: Extending ACA to the Web Environment. *Journal of the American Society for Information Science and Technology*, 57(12), 1616-1628.
- Mantel, N. (1967). A technique of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209-220.
- Mantel, N., & Valand, R. S. (1970). A technique of nonparametric multivariate analysis. *Biometrics*, 26, 547-558.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6), 433-443.
- McGill, M., Koll, M., & Noreault, T. (1979). *An evaluation of factors affecting document ranking for information retrieval systems* (NTIS report No. PB80-119506). Springfield, VA: US Department of Commerce.
- Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, 56(12), 1250-1273.
- Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. New York: McGraw-Hill.

- Nicolaisen, J. (2004). *Social behavior and scientific practice – missing pieces of the citation puzzle*. PhD dissertation. Copenhagen: Department of Information Studies, Royal School of Library and Information Science, 2004.
- Noma, E. (1982). An Improved Method for Analyzing Square Scientometric Transaction Matrices. *Scientometrics*, 4(4), 297-316.
- Oberski, J. E. J. (1988). Some statistical aspects of co-citation cluster analysis and a judgement by physicists. In A. F. J. van Raan (Ed.), *Handbook of Quantitative Studies of Science and Technology* (pp. 431-462). Amsterdam: Elsevier Science Publishers.
- Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletins of the Japanese Society for Scientific Fisheries*, 22, 526-530.
- Persson, O. (1994). The intellectual base and research front of JASIS 1986-1990. *Journal of the American Society for Information Science*, 45(1), 31-38.
- Pielou, E. C. (1984). *The interpretation of ecological data: A primer on classification and ordination*. New York: Wiley.
- Price, D. J. D. S. (1981). The analysis of square matrices of scientometric transaction. *Scientometrics*, 3(1), 55-63.
- Salton, G. (1963). Associative document retrieval techniques using bibliographic information. *Journal of the Association for Computing Machinery*, 10, 440-457.
- Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.
- Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1), 1-10.
- Schönemann, P. H., & Carroll, R. M. (1970). Fitting One Matrix to Another under Choice of a Central Dilation and a Rigid Motion. *Psychometrika*, 35(2), 245-256.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons Inc.
- Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269.
- Small, H., & Greenlee, E. (1980). Citation context analysis of a co-citation cluster: Recombinant DNA. *Scientometrics*, 2(4), 277-301.
- Small, H., & Griffith, B. C. (1974). The structure of scientific literature I: Identifying and graphing specialities. *Science Studies*, 4(1), 17-40.

- Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations. I. A comparison of methods. *Scientometrics*, 7(3-6), 391-409.
- Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations. II. Mapping science. *Scientometrics*, 8(5-6), 321-340.
- Sneath, P., & Sokal, R. (1973). *Numerical taxonomy : The principles and practice of numerical classification*. San Francisco, CA: W. H. Freeman.
- Tanimoto, T. T. (1958). *An elementary mathematical theory of classification and prediction* (IBM Internal Report).
- Wang, Z. W., Wong, S. K. M., & Yao, Y. Y. (1992). An analysis of vector space models based on computational geometry. In *Proceedings of the ACM/SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, 1992* (pp. 152-160).
- White, H. D. (2003a). Author Cocitation Analysis and Pearson's r. *Journal of the American Society for Information Science and Technology*, 54(31), 250-259.
- White, H. D. (2003b). Pathfinder Networks and Author Cocitation Analysis: A Remapping of Paradigmatic Information Scientists. *Journal of the American Society for Information Science and Technology*, 54(5), 423-434.
- White, H. D., & Griffith, B. C. (1981). Author co-citation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163-171.
- Williams, W. T., Clifford, H. T., & Lance, G. N. (1971). Group-Size Dependence - Rationale for Choice between Numerical Classifications. *Computer Journal*, 14(2), 157-162.
- Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6), 579-652.
- Åström, F. (2002). Visualizing Library and Information Science concept spaces through keyword and citation based maps and clusters. In H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari (Eds.), *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science, Seattle, WA, USA July 21-25, 2002* (pp. 185-197). Greenwood Village, CO: Libraries Unlimited.