Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 80, No. 1 (2009) 103–130
DOI: 10.1007/s11192-007-2019-y

A comparative study of first and all-author co-citation counting, and two different matrix generation approaches applied for author co-citation analyses

JESPER W. SCHNEIDER,[a] BIRGER LARSEN,[b] PETER INGWERSEN[b]

[a] *Royal School of Library and Information Science, Aalborg, Denmark*
[b] *Royal School of Library and Information Science, Copenhagen, Denmark*

Aim: The present article contributes to the current methodological debate concerning author co-citation analyses. (ACA) The study compares two different units of analyses, i.e. first- versus inclusive all-author co-citation counting, as well as two different matrix generation approaches, i.e. a conventional multivariate and the so-called Drexel approach, in order to investigate their influence upon mapping results. The aim of the present study is therefore to provide more methodological awareness and empirical evidence concerning author co-citation studies.

Method: The study is based on structured XML documents extracted from the IEEE collection. These data allow the construction of ad-hoc citation indexes, which enables us to carry out the hitherto largest all-author co-citation study. Four ACA are made, combining the different units of analyses with the different matrix generation approaches. The results are evaluated quantitatively by means of multidimensional scaling, factor analysis, Procrustes and Mantel statistics.

Results: The results show that the inclusion of all cited authors can provide a better fit of data in two-dimensional mappings based on MDS, and that inclusive all-author co-citation counting may lead to stronger groupings in the maps. Further, the two matrix generation approaches produce maps that have some resemblances, but also many differences at the more detailed levels. The Drexel approach produces results that have noticeably lower stress values and are more concentrated into groupings. Finally, the study also demonstrates the importance of sparse matrices and their potential problems in connection with factor analysis.

Conclusion: We can confirm that inclusive all-ACA produce more coherent groupings of authors, whereas the present study cannot clearly confirm previous findings that first-ACA identifies more specialties, though some vague indication is given. Most crucially, strong evidence is given to the determining effect that matrix generation approaches have on the mapping of author co-citation data and thus the interpretation of such maps. Evidence is provided for the seemingly advantages of the Drexel approach.

Introduction

Author co-citation analysis (ACA), introduced by WHITE & GRIFFITH [1981], is a technique for mapping the ‘intellectual structure’ of a research field, where the research field is defined as a coherent literature set. The intellectual structure is mapped from the oeuvres of the most cited and co-cited first authors in the particular literature set. Usually, the mapping of a field’s intellectual structure uncovers groupings of authors, which are typically perceived and interpreted as constructs like ‘research specialties’,

Received February 4, 2008; Published online March 18, 2009

Address for correspondence:
JESPER W. SCHNEIDER
E-mail: jws@db.dk

'intellectual bases', or 'paradigmatic positions' within the field (e.g., [WHITE & MCCAIN, 1998; BOYACK, 2004]). Since its introduction, ACA has become a popular and much used technique. However, recently a debate concerning various methodological procedures in ACA has emerged. Especially, the widely adopted approach to ACA developed at Drexel University (aka the Drexel approach) (e.g., [WHITE & GRIFFITH, 1981; MCCAIN, 1990; WHITE & MCCAIN, 1998]) has been the focus of the current debate. Essentially, four methodological issues have been debated in relation to this approach: 1) scalability, i.e., the number of objects (authors) included in author co-citation studies (e.g., [CHEN, 1999; WHITE, 2003A]); 2) units of analyses and their definition, i.e., which cited authors to include in author co-citation studies (e.g., [EOM, 2003; PERSSON, 2001; ROUSSEAU & ZUCCALA, 2004; ZHAO, 2006; ZHAO & STROTMANN, 2007]); 3) choice of proximity measure, i.e. its influence upon grouping and mapping of co-cited authors (e.g., [AHLGREN & AL., 2003; WHITE, 2003B; KLAVANS & BOYACK, 2006; LEYDESDORFF & BENSMAN, 2006; SCHNEIDER & BORLUND, 2007A; 2007B]); and most recently 4) generation and transformation of data and proximity matrices, and how this influence the grouping and mapping of co-cited authors (e.g., [LEYDESDORFF & VAUGHAN, 2006; SCHNEIDER & BORLUND, 2007A]). The evolution and characteristics of the Drexel approach are closely related to the functionalities of the SPSS[1] statistical software and data structure of ISI citation indexes (i.e., [MCCAIN, 1990]). Originally, the approach was characterized by a limited scalability due to restrictions in multidimensional scaling algorithms applied. The unit of analysis is the first cited author in a cited document, a decision dictated by the indexing practices of ISI's citation databases, and the definition of author co-citation is therefore in reality first author co-citation. In the Drexel approach to ACA, the preferred proximity measure is Pearson's product-moment correlation coefficient (r), a choice that seems to rest on the possibility to do factor analysis according to functionalities in SPSS. Finally, and in strong relation to the latter issue of proximity measures, the matrix generation and transformation approach applied has evolved from online searching possibilities in ISI citation indexes through the Dialog database host, and again more available functionalities in SPSS.

The current debate is important, as more methodological awareness and empirical evidence concerning ACA are required. This article contributes empirically to the current debate, as it addresses the *second* and *fourth* methodological issues described above in a comparative study of the mapping effects when applying different units of analysis, first and all-author co-citation counting, based on two different matrix generation approaches. Such a study is feasible since it is based on structured XML documents from the IEEE collection, which allow the construction of ad-hoc citation indexes enabling all-author co-citation analysis.

[1] http://www.spss.com/

The paper is structured as follows. The following section discusses previous research on all-author co-citation analyses and matrix generation approaches. The proceeding section describes the research method of the study, i.e., inclusion criteria, data collection and data analysis. The next section presents and discusses the results, and finally the article ends with a conclusion.

Previous research on units of analysis and matrix generation and transformation

The following section discusses previous theoretical and empirical research findings on all-author co-citation analyses and the applied matrix generation and transformation approaches in author co-citation studies. As indicated above, in several respects the methodological approach to ACA developed at Drexel University has been shaped by specific technical features, which seemingly have brought some constraints or uncertainties to the author co-citation methodology now debated. Most important for the present study is the dependence upon the standardized cited reference strings in Thompson ISI's citation indexes, and the use of the SPSS statistical package as the tool for multivariate analyses (e.g., MCCAIN, 1990).

Units of analysis and their definition in ACA: First- versus all-ACA

The most obvious example of how a technical feature has constrained author co-citation methodology is that the cited reference strings from Thompson ISI only allows for first cited authors as units of analysis in ACA. As a result, the author co-citation methodology only takes into account first cited authors in the definition of author co-citation counts. Two cited authors are considered to be co-cited when at least one cited document from each cited author's oeuvre occurs in the same reference list, an author's oeuvre being defined as all the works with the author as the first cited author [MCCAIN, 1990]. This definition has rarely been challenged. PERSSON [2001] is the first empirical study that compares the potential differences in mapping a field's intellectual structure on the basis of first-author and all-author co-citation analyses. The study is based on 7001 source documents from library and information science journals in the CD-ROM version of Social Science Citation Index 1986–1996. The study investigates how these source documents have been co-cited with each other within the dataset by use of multidimensional scaling (MDS). The co-citations for source documents amount to some 7% of the total number of references in the dataset; the remaining 93% go to non-source documents not indexed by the Thompson ISI citation indexes. The study concludes that first-ACA leaves out several influential researchers compared to all-author ACA, although the subfield structure tends to be just about the same for both units of analysis. The study is somewhat limited due to the dependence on a limited set of source documents, the sparse details provided concerning the definition and calculation of co-citations, and finally the informal evaluation procedures. Nevertheless, the results have indicative value as they are somewhat confirmed in a smaller study done recently by ZHAO [2006].

ZHAO [2006] is the hitherto must detailed theoretical and empirical investigation of first versus all-ACA, including a definition of co-citation counts reminiscent of the definitions given earlier by ROUSSEAU & ZUCCALA [2004]. The study defines three different counting methods: First-ACA; *inclusive* all-ACA; and *exclusive* all-ACA. Likewise, as a consequence of all-ACA, the study redefines "…an author's oeuvre as all works with this author as one of the authors of each of the works." [ZHAO, 2006, P. 1580]. The distinction between inclusive and exclusive all-ACA, refers to the immediate implication of the above definition of all-author co-citation counting of author's oeuvres, as two authors may also be considered as being co-cited when a paper that the two authors co-authored is cited. Thus, co-authorships when cited can also be counted as co-citations. This means that inclusive all-ACA counts cited co-authorships, whereas exclusive all-ACA does not. Typically author co-citations and co-authorships are treated as different units of analysis, where the former is used to map intellectual structures and the latter to investigate research collaboration [ROUSSEAU & ZUCCALA, 2004]. In their definition, ROUSSEAU & ZUCCALA [2004] suggest that such an approach supports the view that authors, regardless of their overall authorship ranking, can contribute substantially to the development of a research area, and that it presents a more accurate portrayal of an individual author's contribution to a research area, where high rates of co-authorship are prevalent (e.g., natural sciences, physics, chemistry). Consequently, inclusive all-author co-citation allows cited co-authorships to be counted, and as a result takes into account connections between authors perceived both by these authors themselves and by the authors of subsequent studies.

Besides the novel definition of all-author co-citation counting, ZHAO [2006] adheres to a traditional Drexel-approach to ACA. The dataset is rather small. It consists of 312 publications in PDF-format on the subject of XML identified using Citeseer.[2] The 312 publications contained 4578 references, which was used as a basis for the co-citation analysis. The results of the study indicate that all-author co-citation counting creates more coherent groups of authors, which therefore supposedly should be considerably clearer to identify and interpret. Nevertheless, due to the straightforward application of citation thresholds for including cited authors in the study, the results also show that all-author co-citation counts can lead to identification of fewer specialties in a research field compared to first-author co-citation counting, that is, when the same number of top-ranked authors is selected and analyzed [ZHAO, 2006].

ZHAO [2006] undoubtedly contributes considerably to our understanding of all-ACA. However, for the time being, the results of the empirical study must be treated with care until we have more substantial evidence that may or may not support its

[2] http://citeseer.ist.psu.edu/

findings. The motivation for the present article is therefore to continue the work of ZHAO [2006] by further investigating inclusive all-ACA in order to bring about deeper empirical understanding and evidence concerning this novel counting approach. In this study we work with a considerably larger data set – the hitherto largest set of citing documents applied in an all-author co-citation analysis.

Generation and transformation of matrices in author co-citation studies: A conventional multivariate approach versus the Drexel approach

Most recently the important role played by matrices in co-citation analyses has received attention. LEYDESDORFF & VAUGHAN [2006] demonstrate the fundamental difference between asymmetric data (e.g. occurrence) matrices ($n \times m$) and symmetric proximity (e.g. co-occurrence) matrices ($n \times n$), arguing that symmetric matrices of co-occurrence counts are *per se* proximity matrices and should be treated as such. This claim is supported by BORG & GROENEN [2005, P. 126], who define co-occurrence data as directly obtained proximities.

In the Drexel-approach to ACA, first author co-citation counts are obtained by online retrieval. Subsequently the co-citation counts are entered into a pre-constructed square symmetric matrix, which is to be considered a proximity matrix [BORG & GROENEN, 2005; LEYDESDORFF & VAUGHN; 2006; SCHNEIDER & BORLUND, 2007A]. However, the desire to apply factor analysis in ACA, as a more detailed exploratory tool in order to identify latent structures and intellectual groupings, and thus help interpret the mapping results, necessitates a symmetric proximity matrix of covariance or correlation coefficients (e.g., [MULAIK, 1972]). In traditional multivariate analyses such proximity matrices are derived from an asymmetric data matrix of variables by cases (objects by observations) (see numerous multivariate statistical textbooks, e.g. [LATTIN & AL., 2003]). However, such a matrix is not available in the Drexel approach. Here a square symmetric proximity matrix is generated directly on the basis of paired author co-citation counts retrieved online (e.g., [MCCAIN, 1990]. As a result an unorthodox procedure is devised, where the proximity matrix of directly obtained co-citation counts is transformed into an additional ‘proximity matrix’ of derived correlation coefficients of first-author co-citation profiles among the included authors (see [SCHNEIDER & BORLUND, 2007A]). Note that a linear transformation of a symmetric proximity matrix is not straightforward (there are also problems in relation to some transaction matrices, see [PRICE, 1981]). A theoretical problem arises, as all relations in a symmetric matrix occur twice, a fact that evidently leads to a magnification when computing correlations in author co-citation profiles [SCHNEIDER & BORLUND, 2007A]. Further, the transformation also causes a fundamental problem in relation to the interpretation and treatment of diagonal values in the original proximity matrix of raw co-citation counts. In SPSS several possibilities for treating diagonal values are available, and the most commonly used in ACA is to treat the diagonal values as missing data (e.g., [MCCAIN, 1990]). Hence, the proximity matrix is in effect treated as a data matrix, since rows are treated as cases and columns as variables; yet this procedure is only allowable when computing correlation matrices in SPSS. The same practice is not applicable in SPSS if one wishes to transform the proximity matrix of co-citation counts into a similarity or distance measure. Missing data beyond doubt causes some loss of information in the matrix and therefore likely influence the ensuing ordination or clustering results. WHITE [2003], 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 approach, where the directly obtained proximity matrix is treated as a data matrix for transformation purposes. The latter problems could be avoided if a conventional multivariate approach to matrix generation and transformation was applied [SCHNEIDER & BORLUND, 2007A]. The pending question however, is whether the different approaches to matrix generation and transformation make a significant difference in the actual mapping and interpretation of intellectual structures in author co-citation studies – both first- and all-author ACA? A further motivation for the present article is therefore also to investigate the influence of matrix generation and transformation approach in ACA, again in order to bring about deeper empirical understanding and evidence concerning these different data representation and transformation approaches.

Consequently, in a combined study the present article explores empirically two of the recently debated methodological issues of ACA, first-author versus all-author co-citation analysis, as well as the influence of different matrix generation and transformation approaches upon mapping results, a conventional multivariate approach versus the Drexel approach. The overall research questions explored in the study are:

- To what extent does a different data set support the previous findings of first-author versus all-author co-citation analysis?
- To what extent do two different matrix generation and transformation techniques influence the interpretation and mapping of author co-citation data?

In order to answer these questions, we perform, and subsequently compare, four author co-citation analyses in the present article: Two first-ACA based on the same set of objects and two *inclusive* all-ACA likewise based on identical objects. One pair of first- and all-ACA is based on a conventional matrix approach commencing from a data matrix, while the other pair of first-and all-ACA is based on the Drexel approach commencing from a proximity matrix of co-citation counts. The four author co-citation analyses are illustrated in Table 1.

Table 1. Delimitation of the four ACA performed

		Matrix generation approaches in ACA:	
		Conventional	Drexel
Units of analyses in ACA:	First-author	×	×
	Inclusive all-authors	×	×

In order to evaluate the effects when applying different units of analysis, as well as different matrix generation approaches, in author co-citation studies, we apply MDS, factor analysis, Procrustes and Mantel statistics; the latter two statistics are introduced and demonstrated in SCHNEIDER & BORLUND [2007B]. Consequently, the evaluations in the present study are quantitative, no author names or qualitative assessments are applied. We have chosen this approach since our aim is a more general methodological one, where the focus is on the general interpretability of latent structure visualizations.

Method

Data collection: Creation of a citation index

The following section outlines the data collection process, the inclusion criteria, and the data analyses applied in the present study. The citation index used in the current study was extracted from a corpus of full text XML documents. The corpus consists of 16,819 articles from the journals of the IEEE Computer Society from the years 1995 to 2004, and is part of the Information Retrieval test collection created by the Initiative for the Evaluation of XML Retrieval (INEX).[3] While the Document Type Definitions (DTDs) used by publishers mainly aid in controlling the printing and publication process many of the so-called *elements* identified by XML tags have many other potential uses. In this paper we extract the following data from the XML version of the IEEE CS journals to form a citation index:

- ***All* cited authors** including their order in the cited document
- **Cited titles** of the cited articles or books
- **Cited journal name** of cited articles

[3] See MALIK & AL. [2006] and http://inex.is.informatik.uni-duisburg.de/2006/ for more information on INEX.

- **Cited year**
- **Cited volume** and **Cited issue** of cited articles
- **Cited page numbers** (begin and end numbers)

For our purpose this data set is ideal: Compared to the Thomson ISI citation indexes we have direct access to all cited authors (100% coverage), and by working directly on the source files used in the production of the original citing articles we also have a range of high quality input data to generate the citation index. By relying on XML data we avoid some of the errors that other approaches have to deal with: So-called autonomous citation indexing [GILES & AL., 1998] based on extraction of reference data from PDF files like those of ZHAO [2006], e.g., CiteSeerIST, Google Scholar[4] and Rexa,[5] have to deal with many problems of segmentation and disambiguation of data from the raw PDF files. The stringent and detailed XML tagging available in our data set makes it possible to unambiguously identify and extract most of the data needed for building a citation index. However, we still have errors arising from the citing authors, that is, errors originating in sloppy reference practise. Some studies have reported quite large shares of references (some more than 50%) with errors originating from citing authors; see e.g. [LOK & AL., 2001] for examples from the medical domain. Our XML data makes it possible to investigate novel ways of automatically detecting such errors. This is, however, a subject for future research and not treated further here. For the present study we have chosen an approach proposed by GLÄNZEL [1996] where each reference is represented by a cluster-key consisting of the two last digits of the cited year, the first four letters of the last name of the author, the volume number, the start page number and the first letter of the name of the cited journal. A key for GLÄNZEL [1996] would then be: 96-GLAN-35-2-167-S. By reducing the references to this short form the key catches many of the variants arising from sloppy referencing. Errors in any of the key's elements are of course not corrected for, but the effect of these is hard to study on large datasets such as ours.

The 16,819 articles from the IEEE CS journals contained a total of 212,657 references (12.6 on average). After application of the cluster-key this was reduced to 132,311 unique references. For each of the cluster-keys that occurred more than once, one of the references was selected to represent the whole cluster in order to be able to extract the cited authors needed in our analysis. For this purpose the cited titles were analysed: if one of the cited titles in a cluster had more occurrences than the rest this was chosen; if there two or more shared the top position the longest (and most specific) cited title was chosen. Finally the data of the representative reference was used to represent all the references in the cluster. For the present study we have extracted two datasets, one with the ID of each citing article plus the cited first authors, and a parallel

[4] http://scholar.google.com/

[5] http://rexa.info/

dataset with all cited authors extracted. The first-author dataset consists of 198,865 pairs of document IDs and cited first authors (13,792 references had no cited authors). The all-author dataset consists of 414,729 pairs.

Inclusion criteria

As indicated above, we perform two first-author and two inclusive all-author co-citation analyses, based on two slightly different dataset, as well as two different matrix generation approaches. Obviously, the data set containing only first cited authors is the basis for the two first-ACA, and likewise, the data set containing all cited authors is the basis for the two all-ACA. The commonly accepted Drexel-approach is applied, as well as an approach based on the conventional procedures for multivariate statistical analysis outlined in numerous textbooks, and emphasised by LEYDESDORFF & VAUGHAN [2006] as well as SCHNEIDER & BORLUND [2007A]. The one pair of first-author and inclusive all-author co-citation analyses commences from an $n \times n$ symmetric proximity matrix (a cited author-by-cited author matrix of co-occurrences), which corresponds to the Drexel-approach. The other pair of first-author and inclusive all-author co-citation analyses commences from an $n \times m$ asymmetric data matrix (a cited author-by-citing document matrix of occurrences), which corresponds to conventional multivariate data analysis.

The basic components in the Drexel-approach are given above and are outlined in MCCAIN [1990]. An approach to ACA based on conventional procedures for multivariate data analysis could well include the same elements as the Drexel-approach; however, there is one tremendously important difference: Multivariate data analyses most often commence from an $n \times m$ multivariate data matrix (e.g., [LATTIN & AL., 2003]). Factor analysis seeks a solution that focuses on the decomposition of the covariance or correlation matrix. This implies that the data matrix *must* be transformed into such a specific type of proximity matrix. Some multivariate techniques, such as MDS or 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, by some suitable proximity measure (e.g., [LATTIN & AL., 2003]). The transformation of the data matrix by use of a proximity measure results in an $n \times n$ symmetric matrix of inter-object proximities. Alternatively, as mentioned above, a proximity matrix can also be generated *directly*, for example from co-citation counts. Such data are then perceived to be proximities, which can be added directly to a proximity matrix. Accordingly, we employ a conventional approach where we commence from a multivariate data matrix. This data matrix is the basis for factor analysis and a proximity matrix of correlations is derived from it. Subsequently, the derived correlation matrix is used as input for non-metric MDS.

Contrary to the Drexel approach, the set of cited authors chosen for analysis is exclusively determined by citation frequency. Two sets of cited authors are needed. A set comprising first-authors of cited references only, and a second set comprising all-authors of the cited references. The latter is inclusive co-authorships, as described in above. Straight counting is applied in both instances. Note that we remove duplicate authors from individual references lists. For example, if cited author *X* appears 5 times in a specific reference list, he or she is only counted once, and the multivariate data matrix is thus binary. The main motivation for invoking duplicate removal is to reduce the likely effect of self-citations, especially in the case of all-author co-citation counting where multiple authorships can lead to an excessive number of self-citations. Contrary to ZHAO [2006], who limited the cited authors to the first five, we include *all* authors of a cited reference in the counting.

An arbitrary citation threshold of 75 cited authors are chosen for both cases of author co-citation. The overlap in authors between the two sets is 41, approximately 55%.

The following matrices are generated:

- one first-author data matrix (75×2002 – cited authors by citing documents),
- one all-author data matrix (75×3161 – cited authors by citing documents),
- one proximity matrix of directly obtained co-citation counts between first-authors (75×75 – cited authors by cited authors), and finally
- one proximity matrix of directly obtained co-citation counts between all-authors (75×75 – cited authors by cited authors).

The latter two proximity matrices are simply a multiplication of the respective data matrices with their transpose. Following the Drexel-approach, the two proximity matrices are transformed into matrices of correlation coefficients. Diagonal values are treated as missing data with the implications described in the introduction.

Data analyses and evaluation procedures

As our aim with the study is a more general methodological one, we do not map author names. The investigation is focused upon interpretability of latent structures when applying different units of analyses and different approaches to matrix generation. As such, the domain under study is to certain degree irrelevant. Consequently, the present evaluations are solely based on multivariate analyses, such as MDS and factor analysis, as well as Procustes and Mantel statistics

Non-metric MDS is applied to all four correlation matrices, i.e. the two derived from the data matrices and the two derived from the initial proximity matrices. We employ the PROXSCAL scaling routine in SPSS. Factors are extracted by principal components analysis with an oblique rotation (Oblimin in *XLSTAT*[6]). The traditional approach to factor rotation in ACA is orthogonal (e.g., [WHITE & MCCAIN, 1998]). When extracting factors, we are forced to impose the assumption that the common factors are mutually uncorrelated (orthogonality) in order to identify the solutions. Theoretically, however, subsequent orthogonal rotation is not appropriate if we expect the resulting factors to be correlated in reality; in the present case such factors would correspond to for example specialties within a research field. By not constraining the rotation to be orthogonal, it may be possible to better approximate a simple structure in the transformed loadings matrix and thereby improve the interpretability of the solution. Oblique rotation aligns the factor axes as closely as possible with groups of variables from the original data set, whether or not the resulting factors are uncorrelated. Thus, the objective function for oblique rotation is similar to that of orthogonal rotation, i.e., minimize the sum of the cross-products of squared factor loadings, but without the constraint of orthogonality. Finally, for want of a better solution, we apply the Kaiser-Guttman rule for determination of the number of factors to extract. The Kaiser-Guttman rule suggests that only those factors with associated eigenvalues that are strictly greater than 1 should be kept.

MDS and factor analysis are thus applied as exploratory data analysis tools when investigating the grouping of authors and whether there is a significant difference between first and all-author co-citation analysis. The investigation into the likely influence upon such analyses when employing different matrix generation techniques is examined statistically by use of the Mantel and Procrustes statistics. The Mantel test is a technique used to estimate the resemblance between two proximity matrices computed about the same objects [MANTEL, 1967].

Procrustes analysis is a statistical technique for comparing two sets of data configurations for the same set of objects [SCHÖNEMANN & CARROL; 1970; GOWER, 1971]. The technique is thoroughly introduced and demonstrated in SCHNEIDER & BORLUND [2007B]. Several approaches to Procrustes analysis exist; we employ the least squares optimization criterion and an orthogonal transformation matrix. The objective is to minimize the sum of the squared deviations, m^2, between points through translating, rotating and dilating one configuration to match the other target configuration. The deviations between points are called vector residuals. A small vector residual indicates a close agreement between the corresponding points. A typical Procrustes analysis simply provides a descriptive summary and graphical comparison of two configurations of points. Although there is a measure of fit provided (m^2), there is no formal means of assessing whether the fit is better than expected by chance. However by employing a permutation approach to one of the data sets, we can determine whether the original m^2 is smaller than expected due to chance (see [SCHNEIDER & BORLUND, 2007B] for details).

[6] http://www.xlstat.com

To sum up, the inclusion criteria and data analyses performed are:

- 75 most cited authors for every ACA;
- Straight counting with normalization for duplicate authors in a reference list;
- The basis for the conventional approach is the generation of two $n \times m$ data matrices of cited authors by citing documents: A data matrix of first cited authors by citing documents, $n \times m = 75 \times 2002$; and a data matrix of all cited authors by citing documents, $n \times m = 75 \times 3161$. These two data matrices are transformed into two $n_r \times n_r$ proximity matrices of correlations between co-cited authors;
- The original data matrices are the basis for the factor analyses, while the proximity matrices are the basis for non-metric MDS;
- The basis for the Drexel approach is the generation of two $n \times n$ proximity matrices of directly obtained raw co-citation counts between cited authors: A proximity matrix of first-author co-citation counts, $n \times n = 75 \times 75$; and a proximity matrix of all-author co-citation counts, $n \times n = 75 \times 75$. These two proximity matrices are treated as data matrices and transformed into two subsequent $n_r \times n_r$ proximity matrices of correlations between author co-citation profiles;
- The original proximity matrices are the basis for the factor analyses, while the correlation matrices are the basis for non-metric MDS.

The next section will present the results of the different analyses and provide a discussion of these results.

Results and discussion

According to the research questions posed in the introduction, two overall investigations are made: First we compare the two different matrix generation approaches by subjecting their MDS configurations to two Procrustes analyses; and subsequently we investigate the potential differences between first and all-author co-citation analyses by use of the MDS configurations, factor analysis, and the Mantel statistic. Note that the present evaluation is solely based on the above mentioned multivariate analyses, Procrustes and Mantel statistics, in combination with manual inspection.

Comparison of matrix generation approaches: the conventional versus the Drexel approach to ACA

As the non-metric MDS configurations are the basis for both analyses we commence by presenting the four non-metric MDS solutions. Figures 1 to 4 illustrates the four

MDS configurations. Figures 1 and 3 are based on the Drexel-approach, and Figures 2 and 4 on the conventional approach. The configurations are mapped using the Matlab[7] software.

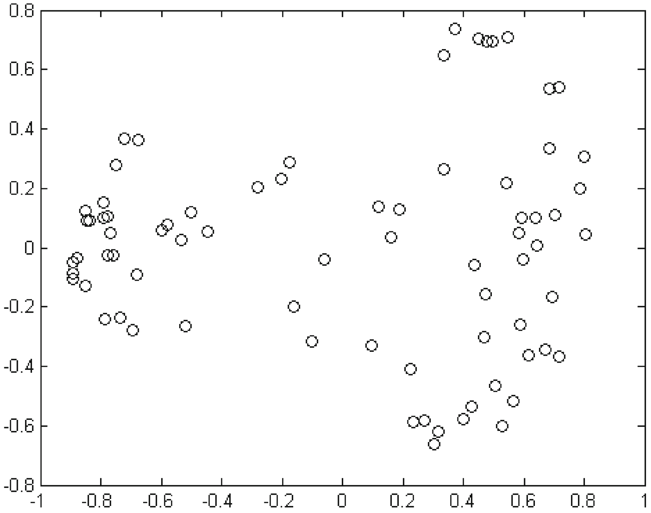


Figure 1. MDS configuration of first-author co-citation analysis – Drexel-approach

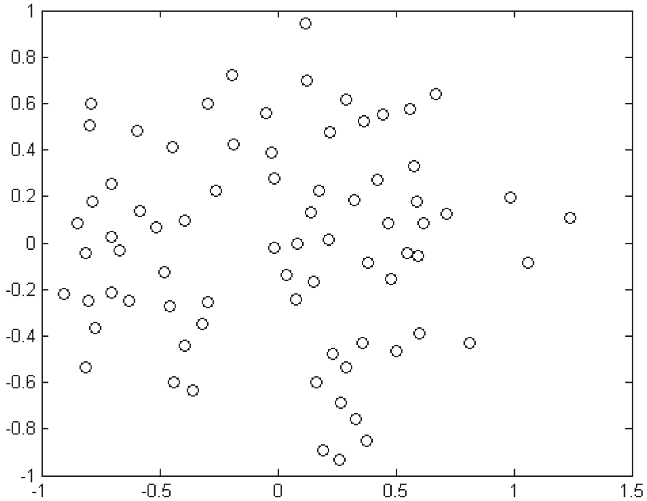


Figure 2. MDS configuration of first-author co-citation analysis – conventional approach

[7] http://www.mathworks.com/

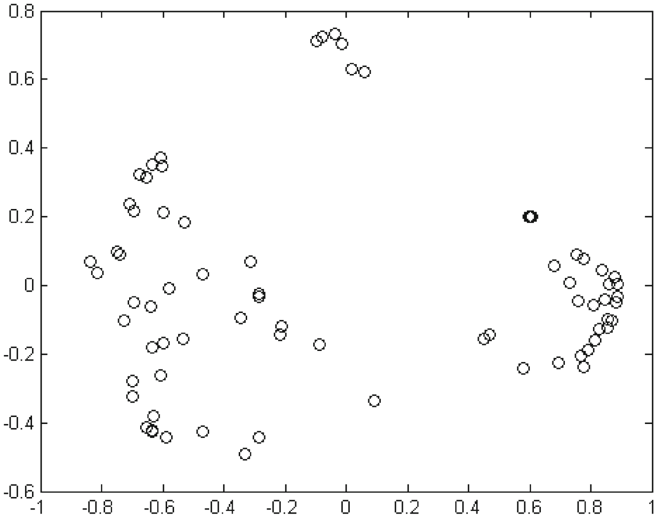


Figure 3. MDS configuration of all-author co-citation analysis – Drexel-approach

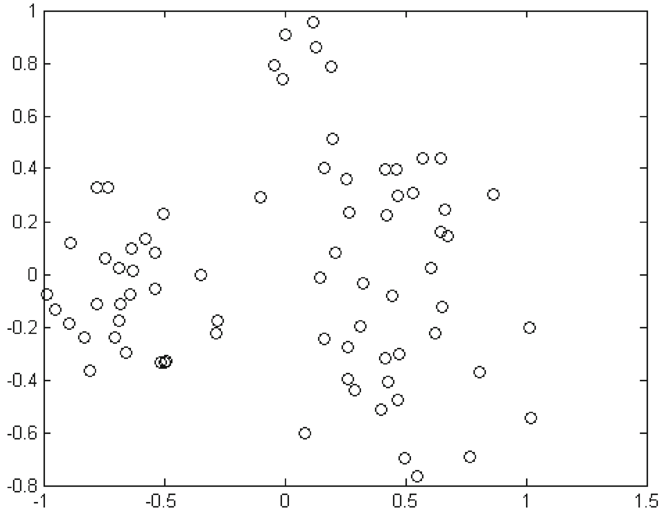


Figure 4. MDS configuration of all-author co-citation analysis – conventional approach

A manual inspection of these configurations indicates that Figures 1 and 3 contain some noticeable structures. Likewise some structure is visible in Figure 4, while it is very difficult to identify any structure in Figure 2. Table 2 below gives the Stress-1 values for the 4 configurations.

Table 2. Stress-1 values for the four MDS-configurations illustrated in Figures 1–4

	Stress-1 values			
Configuration	First-author/ conventional approach	First-author/ Drexel-approach	All-author/ conventional approach	All-author/ Drexel-approach
2 dimensions	0.251	0.174	0.219	0.132
3 dimensions	0.186	0.102	0.181	0.096

None of these values are powerful. It can be inferred that the visibility of structure or rather lack of it in the non-metric MDS configurations above clearly is related to the meagre Stress-1 values. The most acceptable results for 2-dimensions are the two configurations based on the Drexel-approach. The consensus cut of level of 0.2 for Stress-1 values in non-metric MDS is only achieved for all configurations if we include a third dimension. Hence we have four configurations based on two different counting methods and two different matrix generation approaches. None of these configurations have remarkably low Stress-1 values; however, it is noticeable that the best and indeed acceptable configurations are based on the same matrix generation approach, i.e., the Drexel-approach. It is also noteworthy that the all-author configurations have lower Stress-1 values compared to first-author configurations in all cases. In particular, there is a noticeable reduction in Stress-1 in the 2-dimensional solution – indicating that the inclusion of all cited authors better fit the underlying data. This is of some importance as most MDS-based maps are presented in two dimensions. From the manual inspection, it also seems that the all-author ACA maps result in stronger concentrations of the co-cited authors into clusters, regardless of the approach to matrix generation.

One way of investigating whether the two matrix generation approaches provide different ordinations is to compare their solutions, based on the same proximity measure, and for the same set of objects. Procrustes analysis compares and evaluates the resemblance between ordination solutions. Two Procrustes analyses are done, one for the two first-author configurations, and one for the two all-author configurations. Note that the comparable configurations contain the same set of objects, a requirement in Procrustes analysis. This implies that a Procrustes analysis is not possible between first and all-author configurations unless they contain identical objects. The target configuration in both cases is the MDS solution with the lowest Stress-1 value, i.e., the Drexel-approach in both cases. Consequently, the configurations based on the conventional approach are subjected to translation, rotation and dilation in order to match them with the target configurations. The remaining residuals for corresponding points in the two configurations after the least squares fitting provide the m^2 statistic. The m^2 statistic is subject to a permutation procedure in order to test the statistical significance of the comparison.

Figures 5 and 6 gives the Procrustean superimposition plots for the two matching pairs: First-ACA for the same set of objects based on a Drexel and a conventional matrix generation approach, respectively; and all-ACA for the same set of objects based on a Drexel and a conventional approach, respectively.

At 1000 permutations, the Procrustes statistic, m^2, for the first-author comparison is 0.47 (p=0.0001) and m^2 for the all-author comparison is 0.50 (p=0.0001). The Procrustes statistic is essentially a dissimilarity measure of fit, thus the two results clearly indicate a statistical significant difference in the ordination produced by the two matrix generation approaches.

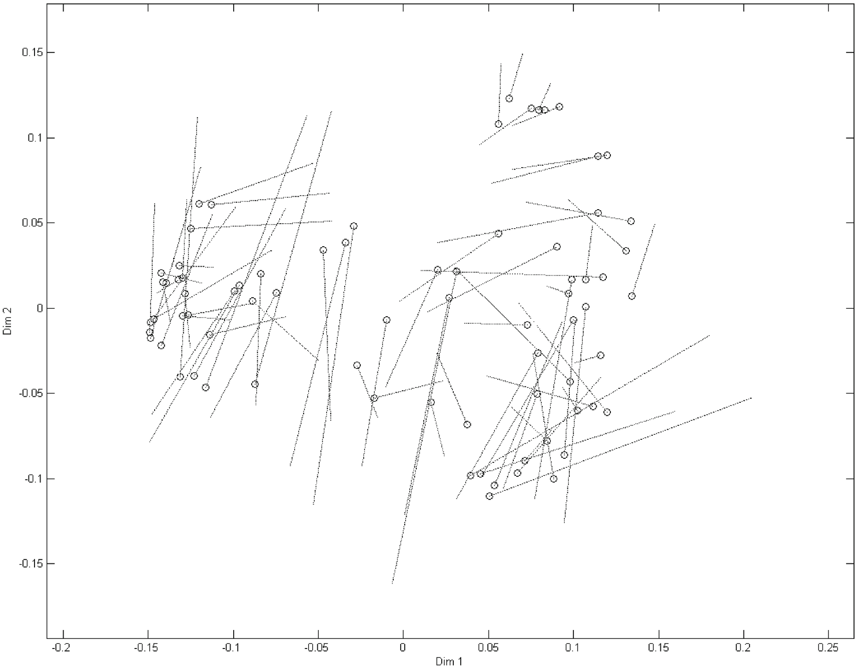


Figure 5. Procrustean superimposition plot of the Drexel (X) and conventional (Y) first-author co-citation configurations; Procrustean m^2statistic = 0.47

The circles in the Procrustean superimposition plots are the authors plotted from the target configurations, in this case the configurations based on the Drexel approach, as they provided the best fit according to the Stress-1 values in Table 2. The lines in the plots indicate residuals after matching the configurations based on the conventional approach.

A small vector residual indicates a close resemblance between the corresponding points and vice versa. It is evident from Figures 5 and 6 that some objects are represented very differently depending on the applied matrix generation approach. However, from the circles in the two plots, we can identify visible structures, and we can also observe that numerous residuals, of different length, move within these structures. Few, if any, shift between major groupings. It is therefore questionable whether the different configurations also mean different overall groupings. It seems that the two matrix generation approaches produces ordinations of some resemblance, however, in the present cases the Drexel approach undoubtedly provides the best and most comprehensible configurations. Further, it is also apparent that the all-ACA seem to have the most coherent structures among the mapped authors.

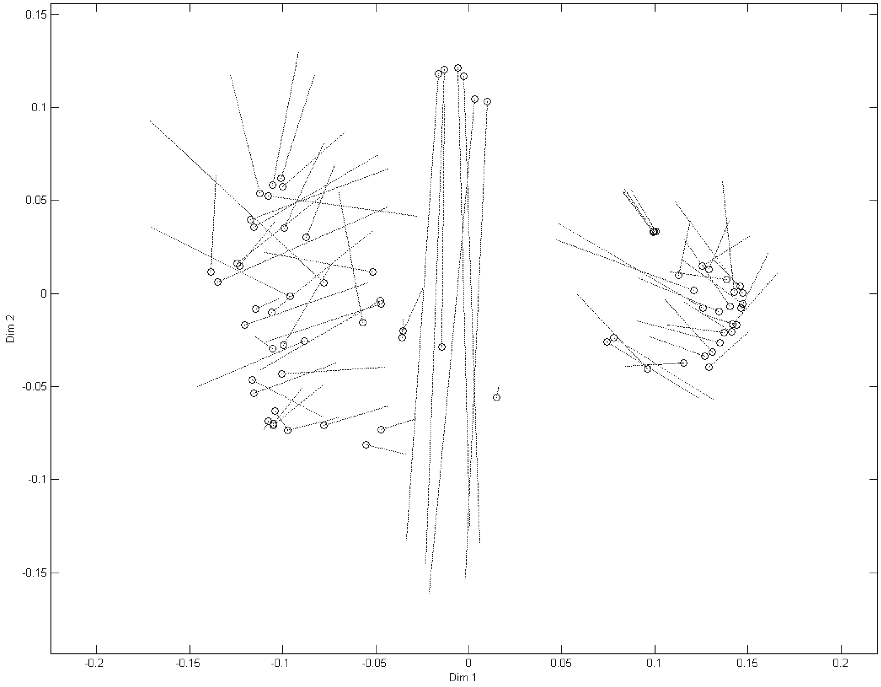


Figure 6. Procrustean superimposition plot of the Drexel (X) and conventional (Y) all-author co-citation configurations; Procrustean m^2statistic = 0.50

Comparison of different units of analysis: first- versus all-author co-citation analysis

Factor analysis is traditionally applied in ACA to elaborate on the mapping results, i.e., to assist in finding latent structures and groupings among objects. Traditionally major factors in ACA are interpreted as 'research specialties' or 'paradigmatic positions' (cf. [WHITE & MCCAIN, 1998]). Hence in order to investigate whether the present data sets support the previous findings of first- versus all-author co-analysis we compare the four ACA on the basis of exploratory factor analyses.

The most evident indictor for the prevailing latent structure in the data set is the major factors. Nevertheless, the determination and optimal extraction of major factors is a long standing dispute within the statistical community. A common rule of thumb often used for dropping the least important factors from the analysis is the Kaiser-Guttman rule. Utilizing the Kaiser-Guttman rule for extraction of factors in our present analyses result in a 29-factor model for the conventional approach to first-ACA, explaining 59% of the variance; a 25-factor model for the conventional approach to all-ACA, explaining 63% of the variance; a 16-factor model for the Drexel approach to first-ACA, explaining 80% of the variance; and finally a 15-factor model for the Drexel-approach to all-ACA, explaining 86% of the variance. The results of the four factor analyses are summarized in Table 3 below.

Table 3. Summary of factors extracted and variance explained in the four factor analyses

	Conventional approach	Drexel approach
First-author co-citation analysis	29 factors/ 59% variance explained Cronbach's α: 0.45	16 factors/ 80% variance explained Cronbach's α: 0.78
All-author co-citation analysis	25 factors/ 63% variance explained Cronbach's α: 0.64	15 factors/ 86% variance explained Cronbach's α: 0.93

From Table 3 it is apparent that there is a huge difference between the factor solutions for the two matrix generation procedures, i.e., the vertical axes. In addition, it is also noticeable that only a very small difference in factor solutions exits between first- and all-ACA, i.e., the horizontal axis.

Extraction of a smaller set of factors explaining a majority of the variance in the data set, no doubt implies that the latent structures in the data set should be more explicable and visible. In the present study, the factor solutions from the Drexel approach comes closets to this objective by providing a 15 and 16 factor solution for first- and all-ACA, respectively. In ZHAO [2006] the inclusive all-ACA provided a 5-factor model, explaining 97% of the variance, and the first-ACA provided an 11-factor model, explaining 96% of the variance – indeed impressive results, which are interpreted so that first-ACA provides more specialties compared to all-ACA, the latter however provides more coherent groups. When comparing the present factor solutions to those of ZHAO [2006], some consideration concerning the differences in character of the two data sets are needed. ZHAO [2006] is based on a somewhat restricted subject domain, whereas the present data set contains authors from a substantial number of sub-disciplines including the domain treated in ZHAO [2006]. On the basis of these characteristics, one would expect a larger scattering and less coherence in the present data set compared to the relative small set provided by ZHAO [2006]. Nevertheless, as Table 3 indicates a considerable amount of variance can be explained in the present data set.

Let us consider the two factor solutions based on the conventional approach first. As indicated by the Stress-1 values, and demonstrated by the Procrustes analyses, the configurations based on the conventional approach on the whole have a poorer fit between its observations and derived proximities resulting in inferior mappings, compared to the Drexel-approach. This is complemented by the internal consistency reliability measure of Cronbach's alpha often used in factor analyses [CRONBACH, 1951]. Factors are sets of original variables which are thought to measure a latent construct. Variables in a factor will normally be more inter-correlated with each other than with other variables representing other latent constructs. Cronbach's alpha is the common test of whether items are sufficiently interrelated to justify their combination in a factor. The standardized coefficient can be calculated as follows:

$$\alpha = \frac{k \cdot \bar{r}}{[1 + (k-1) \cdot \bar{r}]} \tag{1}$$

where k is the number of variables used in the factor and $\bar{r}$ is the average inter-item correlation among k items. Cronbach's alpha ranges from 0 to 1; the higher the average inter-item correlation, the greater the value of α, reflecting a higher internal consistency for the factor. Controversy exits in relation to what constitutes a good α, however, NUNNALLY [1978] recommends that α should be above 0.7.

In the present study, we average the values of α form the individual factors in an ACA, in order to get an impression of the overall ability of the particular author co-citation study's ability to create internal consistent factors, and thus its ability to create comprehensive factor solutions. This means that higher average values indicate a more comprehensive structure. Table 3 above provides the averaged coefficients of Cronbach's alpha after oblique rotation for the four author co-citation studies. Again we see that the highest values, and according to NUNNALLY [1978] the only reliable values, are the ones obtained when the Drexel approach to ACA is applied. However, the conventional approach based on all-author co-citations is close to the threshold of 0.7. The latter indicates that all-author co-citation counting produces more coherent groupings, compared to first-author counting, as suggested by ZHAO [2006]. We can therefore conclude that the averaged values of Cronbach's alpha complement the previous indications suggested by the Stress-1 values and the Procrustes statistics, and that the configurations based on the conventional matrix approach on the whole have a poorer fit and provide a less comprehensible structure, compared to the Drexel approach. In addition, all-author co-citation counting seems to produce more coherent groupings, compared to first-author counting. Note however, that first-author counting, combined with the Drexel approach to matrix generation, still produces a factor solution with fewer factors, where more variance is explained, as well as a higher α value. This suggests that the matrix generation and transformation approach is a determining cause when considering the outcome of an ACA.

Obviously, the two data matrices (n×m), upon which the conventional approach is based, are extremely sparse. Whereas the two proximity matrices (n×n), upon which the Drexel-approach is based, are relatively dense; the latter no doubt is profoundly influenced by the fact that the proximity matrices are treated as data matrices, doing this, means that the proximity values appear twice on each side of the diagonal. Treating a matrix this way is unorthodox to be sure; nevertheless, it seems to provide very clear results in ACA. Hence, the conventional transformation of sparse n×m into n_r×n_r and the Drexel transformation of the dense n×n into n_r×n_r yield different configurations for the same set objects.

Dense matrices contain few zero connections and consequently indicate a denser network of objects more suitable for measuring correlation coefficients (cf. [SCHNEIDER & BORLUND, 2007A]). On the other hand, sparse matrices contain a considerable number of zero values, which makes computation of correlation coefficients more vulnerable (cf. [SCHNEIDER & BORLUND, 2007A]). Remember that the basis for the Drexel-approach is a matrix of co-citations obtained by multiplying the data matrices by their transpose. Both types of matrices are transformed into matrices of correlation coefficients. As demonstrated above, the two approaches yield different configurations. If the data contour of the matrices is irrelevant to a proximity transformation, then the same set of objects represented by two different matrices should obtain monotone rankings. If however, zero values influence the computation of the correlation coefficient, then rankings will deviate. Two Mantel tests [MANTEL, 1967; SCHNEIDER & BORLUND, 2007B] between the two pairs of correlation matrices for the conventional and Drexel-approach respectively (i.e. the latter computed from the data matrices (75×2002 first-author, and 75×3161 all-author), and the former from the co-citation matrices (75×75 both) prove this. The Mantel test is a statistical test of the correlation between two matrices, where the matrices must be of the same rank. To overcome the problem of non-independent proximities within the matrices, a random permutation strategy and a non-parametric correlation model are applied in order to assess significance (i.e., see [SCHNEIDER & BORLUND, 2007B] for details).

The Mantel statistic for the first- and all-author correlation matrices are outlined in Table 4.

Table 4. Non-parametric Mantel tests for correlation matrices produced by the conventional and Drexel approaches to the first- and all-author co-citation analyses

Conventional versus Drexel approach for the first-author co-citation matrices	*rho*=0.683 (*p*=0.0001)
Conventional versus Drexel approach for the all-author co-citation matrices	*rho*=0.793 (*p*=0.0001)

This clearly indicates a decreasing monotonicity between the rankings of objects. Note that the Mantel statistic is considerably higher for the all-author correlation matrices, a fact also indicated by the Stress-1 values for the respective pair of configurations. At first glance, this seems to contradict our claim as the dimensionality of the data matrix for the all-author co-citation analysis is considerably larger than its counterpart for the first-author analysis, 75×3161 and 75×2002, respectively. However, when computing the density of these matrices (i.e., a function of the number of non-zero values in the matrix compared to its size), it becomes clear that the larger all-author data matrix has a larger density (0.05) compared to the first-author data matrix (0.04).

Likewise, if we interpret the entropy statistic [SHANNON & WEAVER, 1949] as a measure for the ‘amount of mix’ found in two-class variables within a matrix (binary values of zero and one), we get similar indications. The special case of entropy for a random variable with two outcomes is the binary entropy function:

$$H(X) = H_b(p) = -p \log p - (1-p) \log(1-p) \qquad (2)$$

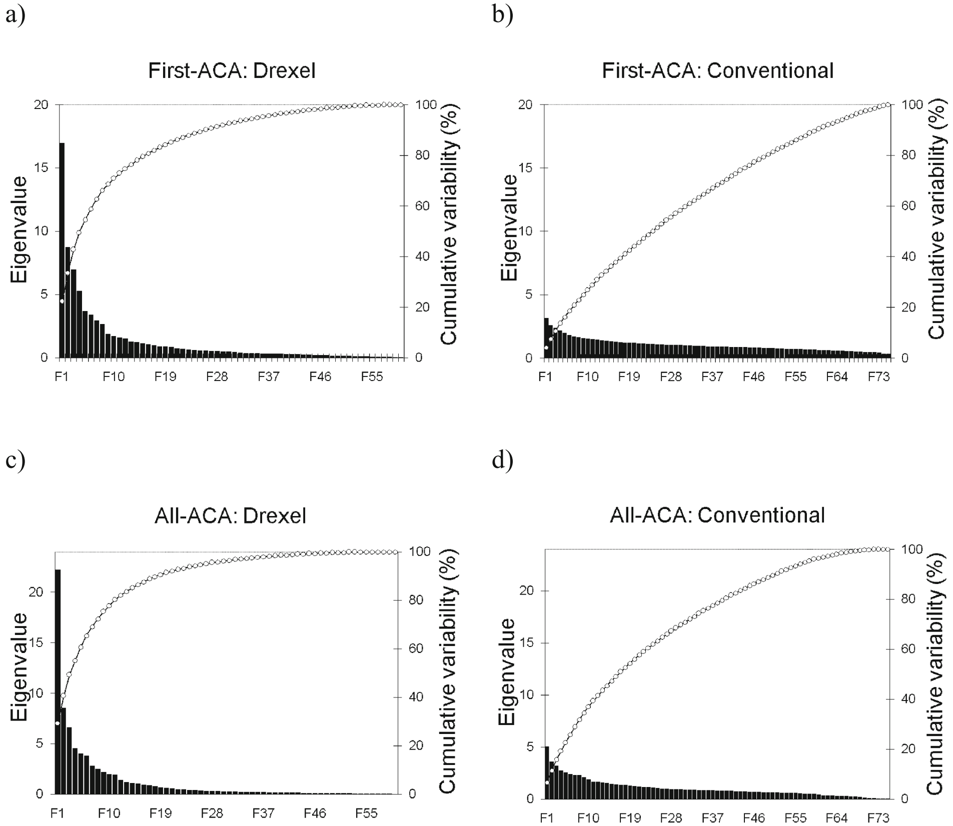
where H is entropy, and p is the probability of one class (0.5), which class is indifferent, either ones or zeros in the present case. We use the base 2 logarithm, which means that the calculated entropy is in units of bits. Generally, the more distinct values in the matrix, and the more evenly they are distributed, the greater the entropy. For example, matrices with completely even distributions of values possess 1 bit of entropy when there are two-class values. The less evenly those values are distributed, the lower the entropy. The entropy statistics for the two matrices are given in Table 5.

Table 5. Entropy statistics for the two data matrices

First-author data matrix	$H(X) = 0.22$
All-author data matrix	$H(X) = 0.27$

The entropy statistics indicate that the values in the larger all-author data matrix are more distinct and evenly distributed compared to the first-author matrix. This supports the density measures for the two matrices, and the latter confirms that the distinct values expressed in the entropy statistics are the non-zero values (ones). Consequently, even though the all-author data matrix is larger it also contains more non-zero values.

The present results of the Mantel and entropy statistics are tremendously important for a factor analysis, as the latter is based on a decomposition of a covariance or a correlation matrix. Applying factor analysis to a correlation matrix with only low inter-correlations, a likely consequence of transforming sparse matrices, will require factor solutions with nearly as many factors (principal components) as there are original variables, thereby defeating the data reduction purposes of factor analysis. This is clearly the case in the present analysis for the matrices based on the conventional approach as indicated in Figures 7b and 7d below.



Figures 7a–d. Factorial scree plots of eigenvalues and cumulative variability explained

What is desirable when choosing major factors is clearly a shape of the histogram like the ones in Figures 7a and 7c, preferably even steeper, and likewise, cumulative variability curves like those on Figures 7a and 7c, where the total explained variance is spread among relatively few major factors.

Accordingly, for the present analysis only the factor analyses based on the Drexel-approach are reliable for identification of major factors or rather ‘specialties’ within the IEEE data set; a solution supported by the averaged Cronbach’s alpha values outlined in Table 3. Consequently, we have a 16-factor model for the Drexel approach to first-author co-citation analysis, explaining 80% of the variance; and a 15-factor model for the Drexel-approach to all-author co-citation analysis, explaining 86% of the variance.

A first glance at the eigenvalues for the extracted factors in Table 6 below reveals a remarkable similarity. A close manual inspection of the factor patterns (not included in the paper), before and after oblique rotation, reveals an even more significant similarity between the extracted factors. Remember that the two sets had an overlap of 41 authors. These authors represent in total 13 common factors in the first-author case, and 14 factors in the all-author case. The large majority of the remaining authors in both cases are spread among these common factors. The few authors that are left over in both cases represent the remaining few factors. Not unusually only one author are affiliated with such a factor, and it is indeed questionable whether such a one-object factor contributes to the uncovering of ‘specialties’.

Figures 8 and 9 demonstrate the similarity in structure, for the un-rotated factors 1 and 2, between the two author co-citation counting methods. Crosses indicate exclusive authors for the particular analysis, whereas circles indicate overlapping authors between the two analyses.

Table 6. Eigenvalues for first- and all-author co-citation analysis

	First-author co-citation analysis				All-author co-citation analysis		
	Eigenvalue	Variability (%)	Cumulative %		Eigenvalue	Variability (%)	Cumulative %
F1	16.944	22.233	22.233	F1	22.156	29.289	29.289
F2	8.689	11.402	33.634	F2	8.508	11.247	40.536
F3	6.926	9.088	42.722	F3	6.593	8.716	49.251
F4	5.248	6.886	49.609	F4	4.513	5,965	5.217
F5	3.650	4.790	54.398	F5	4.009	5.299	60,516
F6	3.376	4.429	58.828	F6	3.756	4.965	65.481
F7	2.934	3.850	62.678	F7	2.764	3.654	69.135
F8	2.645	3.470	66.148	F8	2.474	3.270	72.405
F9	1.841	2.415	68.564	F9	2.178	2.879	75,283
F10	1.706	2.238	70.802	F10	1.949	2.576	77.860
F11	1.564	2.052	72.854	F11	1.891	2.499	80.359
F12	1.484	1.947	74.801	F12	1.361	1.799	82.158
F13	1.241	1,628	76.429	F13	1.167	1,543	83.701
F14	1.180	1.548	77.977	F14	1.051	1.390	85.090
F15	1.103	1.447	79.424	F15	1.023	1.353	86.443
F16	1.013	1.330	80.754				

It seems that the inclusive all-author co-citation counting produces a slightly more coherent grouping between the authors. At least its structure does support and elaborate the MDS solution presented in Figure 3. The explanation for this finding is likely to be found in the special counting of co-authorships as devised by inclusive all-author co-citation counting. Theoretically, such cited co-authorships, other things being equal, will produce stronger connections between groups of cited authors in a network.

It is difficult to determine whether the small differences in extracted factors between the two analyses indicate whether first-author co-citation analysis is able to identify more specialties. Some evidence in the data set suggests that several of these authors perhaps should be treated as outliers. However, a comparison of the mapping of exclusive and overlapping authors in Figures 8 and 9, suggest a more coherent distribution of overlapping authors in the all-ACA, whereas a few exclusive authors in the first-ACA are plotted at some distance from overlapping authors indicating a specific grouping. The latter may be interpreted as the ability of first-ACA to produce more groupings compared to all-ACA, as suggested by ZHAO [2006]. Even so, the differences in the present analyses are so small, that caution must be imposed. Consequently, from the present study we cannot confirm that all-author co-citation counts can lead to identification of fewer specialties, however, the study do confirm that all-author co-citation counting produce more coherent groupings amongst authors.

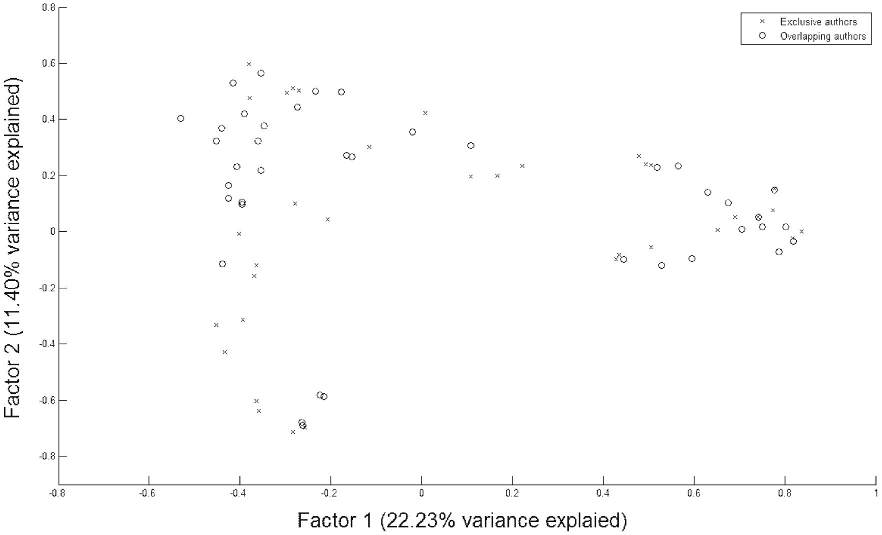


Figure 8. Mapping of un-rotated factors 1 and 2 for the first-author co-citation analysis; variability for F1: 22.23% and F2: 11.40%; cumulative variability: 33.63%

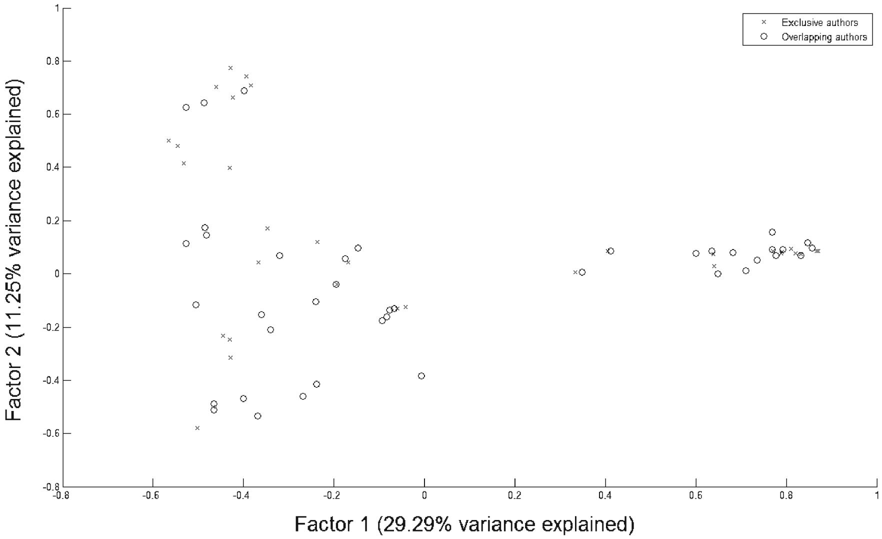


Figure 9. Mapping of un-rotated factors 1and 2 for the all-author co-citation analysis; variability for F1: 29.29% and F2: 11.25%; cumulative variability: 40.54%

Conclusion

The present article has presented a comparative study of mapping effects when different units of analyses as well as different matrix generation approaches are applied in author co-citation analyses. Specifically, the study investigated first-author versus inclusive all-author co-citation counting, and the Drexel versus conventional approach to matrix generation and transformation. Further, the study is based on the largest dataset so far used in an all-author co-citation study. The dataset was drawn from full-text scholarly articles formatted in XML that allows precise extraction of a large range of features; most notably in relation to this study all cited authors. As such, the combined study explores empirically two of the recently debated methodological issues of ACA.

The results show that the inclusion of all cited authors can aid in producing two-dimensional mappings based on MDS that better fit the underlying data (i.e., have lower stress values), and that inclusive all-author co-citation counting may lead to stronger groupings in the maps. The latter is likely due to co-authorships being counted as devised by the inclusive all-author counting strategy.

In addition, we also study the potential divergence in mapping results when different matrix generation and transformation approaches are applied: the popular Drexel approach, as well as a conventional approach based on multivariate statistical analysis. Overall the two approaches produce maps that have some resemblances, but also many differences at the more detailed levels. The Drexel approach produces results that have noticeably lower stress values and are more concentrated into groupings, which makes these maps more comprehensible. This finding is most likely due to the unorthodox matrix generation and transformation approach, where the basis is a dense proximity matrix that is treated as a data matrix. In relation to the previous finding, the study also demonstrates the importance of sparse matrices and their potential problems in connection with factor analysis. In the present study the sparse matrices deflated the extraction of factors, whereas the dense matrices, resulting from the unorthodox Drexel approach, produced much better factor solutions explaining a satisfactory amount of the variability in these matrices. Indeed sparse matrices are probably more suitable for principal components analysis. Factor analysis and principal components analysis are different in their goals and in their underlying models. Roughly speaking, one should use principal components analysis to summarize or approximate data using fewer dimensions and factor analysis for an explanatory model of the correlations among the data (see for example, [LATTIN & AL., 2003]).

Thus, to answer the two research questions, while we can confirm that inclusive all-author co-citation analysis produce more coherent groupings of authors, the present study cannot clearly confirm previous findings that first-author co-citation analysis identifies more specialties, though some vague indication is given. We need to investigate this further, as the removal of duplicates may have affected this result. And most crucially, strong evidence is given in the present study to the determining effect that matrix generation and transformation approaches have on the mapping of author co-citation data and thus the interpretation of such maps. Indeed this finding is general as its affects all studies where a Drexel-like approach is applied to co-occurrence data.

*

This paper is extended and revised from a paper presented at the ISSI 2007 conference [SCHNEIDER & AL., 2007].

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.

BORG, I., GROENEN, P. J. F. (2005), *Modern Multidimensional Scaling: Theory and applications.*2nd. Springer Science & Business Media: New York.

BOYACK, K. (2004), Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences*, 101 : 5192–5199.

CHEN, C. (1999), Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35 (3) : 401–420.

CRONBACH, L. J. (1951), Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3) : 297–334.

EOM, S. B. (2003), *Author Cocitation Analysis using Custom Bibliographical Databases: An Introduction to the SAS systems*. The Edwin Mellen Press, Lewiston, New York.

GILES, C. L., BOLLACKER, K., LAWRENCE, S. (1998), CiteSeer: An Automatic Citation Indexing System. In: *Third ACM Conference on Digital Libraries*. ACM Press, New York, pp. 89-98.

GLÄNZEL, W. (1996), The need for standards in bibliometric research and technology, *Scientometrics*, 35 (2) : 167–176.

GOWER, J. C. (1971), Statistical methods for comparing different multivariate analyses of the same data. In: HODSON, KENDALL, TAUTU (Eds), *Mathematics in the Archaeological and Historical Sciences*. Edinburgh: Edinburgh University Press, pp. 138–149.

KLAVANS, D., BOYACK, K. (2006), Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57 (2) : 251–263.

LATTIN, J., CARROLL, J. D., GREEN, P. E. (2003), *Analyzing Multivariate Data*. Pacific Grove, CA: Brooks/Cole - Thompson Learning.

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.

LOK, C. K. W, CHAN, M. T. W., MARTINSON, I. M. (2001), Risk factors for citation errors in peer-reviewed nursing journals. *Journal of Advanced Nursing*, 32 (2) : 223–229.

MALIK, S., KAZAI, G., LALMAS, M., FUHR, N. (2006), Overview of INEX 2005. In: FUHR & AL.: *Proceedings of INEX 2005*. Berlin: Springer 2006, pp. 1–15. (LNCS 3977)

MANTEL, N. (1967), A technique of disease clustering and a generalized regression approach. *Cancer Research,* 27 : 209–220.

MCCAIN, K. W. (1990), Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science,* 41 (6) : 433–443.

MULAIK, S. A. (1972), *The foundations of factor analysis*. New York: McGraw-Hill.

NUNNALLY, J. C. (1978), *Psychometric theory*. 2nd edition. New York: McGraw-Hill.

PERSSON, O. (2001), All author citations versus first author citations. *Scientometrics*, 50 (2) : 339–344.

PRICE, D. J. DE SOLLA (1981), The analysis of square matrices of scientometric transaction, *Scientometrics*, 3 (1) (1981) 55–63.

ROUSSEAU, R., ZUCCALA, A. (2004), A classification of author co-citations: definitions and search strategies. *Journal of the American Society for Information Science and Technology*, 55 (6) : 513–629.

SCHNEIDER, J. W., BORLUND, P. (2007A): Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, 58 (11) : 1586–1595.

SCHNEIDER, J. W., BORLUND, P. (2007B): Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology*, 58 (11) : 1596–1609.

SCHNEIDER, J. W., LARSEN, B., INGWERSEN, P. (2007), Comparative study between first and all-author co-citation analysis based on citation indexes generated from XML data. In: *Proceedings of ISSI 2007, 11th International Conference of the International Society for Scientometrics and Informetrics,* Eds. Torres-Salinas, D. & Moed, H. CINCDOC, Madrid, pp. 696–707.

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.

SHANNON, C. E., WEAVER, W. (1949), *The Mathematical Theory of Communication*. University of Illinois Press; Urbana IL.

WHITE, H. D. (2003a), 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. (2003b), Author Cocitation Analysis and Pearson's r. *Journal of the American Society for Information Science and Technology,* 54 (31) : 250–259.

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.

WHITE, H. D., MCCAIN, K. (1998), Visualizing a discipline: An author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science,* 49 (4) : 327–355.

ZHAO, D. (2006), Towards all-author co-citation analysis. *Information Processing & Management*, 42 (6) : 1578–1591.

ZHAO, D., STROTMANN, A. (2007), Can citation analysis of web publications better detect research fronts? *Journal of the American Society for Information Science and Technology*, 58 (9) : 1285–1302.