

Comparative Study between First and All-Author Co-Citation Analysis Based on Citation Indexes Generated from XML Data

Jesper Wiborg Schneider, Birger Larsen and Peter Ingwersen

jws@db.dk, blar@db.dk, pi@db.dk

Department of Information Studies, Royal School of Library and information Science
Birketinget 6, DK-2300 Copenhagen S (Denmark)

Abstract

The study presents a comparative analysis between first and all-author co-citation analyses, as well as comparison between two matrix generation approaches. We thus continue the latest research in author co-citation analysis (ACA), where the results of the traditional first-author analyses based on ISI citation indexes are challenged by incorporating all-authors from the cited references. Identifying all cited authors from references in source papers is an extremely cumbersome process if the Thomson ISI citation indexes are used as a basis. Due to the difficulty in obtaining all-author co-citation data few such studies exist. In order to study all-authors co-citation we use a citation index generated from documents in XML code. This allows us to carry out a comparative study between first and all-author co-citation analyses based on the hitherto largest set of references and the broadest domain of research.

Introduction

Author co-citation analysis (ACA), introduced by White and Griffith (1981), is a technique for mapping the 'intellectual structure' of a research field, where the latter 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 a particular literature set. Since its introduction, ACA has become a popular and much used technique. However, recently a debate concerning methodical procedures in ACA has emerged. Especially, the approach to ACA developed at Drexel University (e.g., White & Griffith, 1981; McCain, 1990) has been the focus of the current debate. Essentially, four methodical issues have been debated: 1) scalability (e.g., Chen, 1999), 2) units of analysis and their definition (e.g., Persson, 2001; Zhao, 2006; Rousseau & Zuccala, 2004), 3) the choice of proximity measures (e.g., Ahlgren, Jarneving, & Rousseau, 2003; Schneider & Borlund, 2007a; 2007b), and most recently 4) generation and transformation of matrices (Leydesdorff & Vaughan, 2006; Schneider & Borlund, 2007a). The present paper addresses the second and fourth issues in a comparative study of first and all-author co-citation analysis based on different matrix generation approaches in structured XML documents that allow for the construction of ad-hoc citation indexes.

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

Previous Work on All-author Co-citations and matrix generation

In several respects, the methodical approach to ACA developed at Drexel University has been shaped by specific technical features that have seemingly brought some constraints to the ACA methodology. Most important 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. The most obvious example is that the cited reference strings only allows for first authors as units of analysis in ACA. As a result, ACA methodology only takes into account first authors in the definition of author co-citation counts. Two authors are considered to be co-cited when at least one document from each author's oeuvre occurs in the same reference list of a citing document, where an author's oeuvre is defined as all the works with the author as the first author (McCain, 1990). This definition has rarely been challenged. Persson (2001) is the first empirical study that compares the potential difference in intellectual structure between mappings done by 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 demonstrates that first-author ACA leaves out several influential researchers compared to all-author ACA, although the subfield structure tends to be just about the same for both methods. 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 are indicative as they are somewhat confirmed in a smaller study done by Zhao (2006).

All-author vs. First-author Co-citation Analyses

Zhao (2006) is the hitherto most detailed theoretical and empirical investigation of all-author co-citation analysis, including a definition of co-citation counts reminiscent of the definitions given earlier by Rousseau and Zuccala (2004). The study defines three different counting methods: first-author co-citation; *inclusive* all-author co-citation; and *exclusive* all-author co-citation. Likewise, as a consequence of all-author co-citation analysis, 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-author co-citations 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 into co-citations. This means that inclusive all-author co-citation analysis counts cited co-authorships, whereas exclusive all-author co-citation analysis 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 and Zuccala (2004), in their definition, 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.

Besides the novel definition of all-author co-citation counting, Zhao (2006) adheres to a traditional Drexel-approach to ACA (see below). The dataset was rather small: it consisted of 312 publications in PDF on the subject of XML identified using CiteSeer¹. The 312 publications contained 4578 citations, 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 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 count 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-author co-citation analysis. However, for the time being, the results of the empirical study must be treated carefully until we have more substantial evidence that may or may not support its findings. The motivation for the present paper is therefore to continue the work of Zhao (2006) by further investigating inclusive all-author co-citation analysis in order to bring about deeper empirical understanding and evidence concerning this novel counting approach. The present study is the first in a series that addresses the research possibilities inherent in a citation index based on source documents formatted in XML. One such possibility is all-author co-citation analysis, and the present study is based on the hitherto largest set of citing documents applied in an all-author co-citation analysis.

Co-citation Matrix Generation

Most recently the role played by matrices in co-citation analyses has received attention. Leydesdorff and Vaughan (2006) demonstrate the fundamental difference between asymmetric data matrices ($n \times m$) and symmetric proximity matrices ($n \times n$), arguing that symmetric matrices of co-occurrence counts are *per se* proximity matrices and should be treated as such.

¹ <http://citeseer.ist.psu.edu/>

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 symmetric proximity matrix. However, the desire to apply factor analysis to ACA as a more detailed exploratory tool in order to identify latent structures and thus help interpret the mapping results, necessitates a symmetric proximity matrix of covariance or correlation coefficients. In traditional multivariate analyses such proximity matrices are derived from an asymmetric data matrix of variables by cases. However, such a matrix is not available in the Drexel-approach due to the paired online counting. As a result an unorthodox procedure is devised, where the proximity matrix of co-citation counts are transformed into an additional proximity matrix of derived correlation coefficients of first author co-citation profiles. Note that a linear transformation of a symmetric proximity matrix is not straightforward. A theoretical problem arises, as all relations in a symmetric matrix occur twice which evidently leads to a magnification. Further, the transformation also causes a fundamental problem in relation to the interpretation and treatment of diagonal values. In SPSS several possibilities for treating diagonal values are available, and the most commonly used in ACA is to treat the diagonals as missing data (e.g., McCain, 1990). Hence, rows are treated as cases and columns as variables, yet this procedure is only allowable when computing correlation matrices. 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, 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. The question is whether the different approaches to matrix generation make a significant difference in the interpretation and mapping of ACA?

Consequently, the present study explores two of the recently debated methodical issues of ACA, first-author versus all-author co-citation analysis, and the influence of different matrix generation approaches upon mapping results. The major 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 does different matrix generation techniques influence the interpretation and mapping of author co-citation data?

In order to answer these questions, we perform two first-author and two inclusive all-author co-citation analyses, one pair commencing from a data matrix and one pair commencing from a proximity matrix of co-citation counts.

Method

Data set

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)². 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
- **Cited year**

² See Malik et al. (2005) and <http://inex.is.informatik.uni-duisburg.de/2006/> for more information on INEX.

- **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, Bollacker & Lawrence, 1998) based on extraction of reference data from PDF files like those of Zhao (2006), e.g., CiteSeer.IST, Google Scholar³ and REXA⁴, have to deal with many problems of segmentation and disambiguation of data from the raw PDF files. However, using the XML data 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, Chan & Martinson (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 this 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, the next step was to select one of the references 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) reference 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 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.

Data analysis

We perform two first-author and two inclusive all-author co-citation analyses based on the two dataset and two different approaches to matrix generation. 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. The one pair of first-author and inclusive all-author co-citation analyses commences from an $n \times n$ proximity matrix, which corresponds to the Drexel-approach. The other pair of first-author and inclusive all-author co-citation analyses commences from an $n \times m$ data matrix, 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 $n \times m$ data matrix (e.g., Lattin, Carroll, & Green, 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 employ a proximity matrix as their input. Most commonly, such an input proximity matrix is *derived* from the traditional data matrix, by some suitable proximity measure (e.g., Lattin, Carroll, & Green, 2003). The

³ <http://scholar.google.com/>

⁴ <http://rexa.info/>

transformation of the data matrix by use of a proximity measure results in an $n \times n$ 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 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 MDS.

The set of cited authors chosen for analysis is 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.

The following matrices are generated: one first-author data matrix (75×2002), one all-author data matrix (75×3161), one proximity matrix of directly obtained co-citation counts between first-authors (75×75), and finally one proximity matrix of directly obtained co-citation counts between all-authors (75×75). 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 possible implications described in the introduction.

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*) that does not have the constraint of orthogonality as the factor rotation usually used in ACA. For want of a better solution, we apply the Kaiser-Guttman rule for determination of the number of factors extracted. 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 tools when investigating the grouping of authors and whether there is a difference between first and all-author co-citation analysis. The investigation into the likely difference of results when employing different matrix generation techniques is statistically validated by use of Procrustes analysis.

Procrustes analysis is a statistical technique for comparing two sets of data configurations for the same set of objects (Schönemann & Carroll; 1970; Gower, 1971). The technique is thoroughly introduced and demonstrated in Schneider and 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. A typical Procrustes analysis simply provides a descriptive summary and graphical comparison of two configurations of points. 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, 2007a for details).

Results and discussion

As MDS configurations are the basis for our analyses we commence by presenting the four MDS solutions. Note that we in the present analysis do not map author names; the present evaluation is solely based on multivariate techniques, Procrustes analysis and Mantel test in combination with manual inspection.

Figures 1 to 4 below illustrates the four MDS configurations. Figures 1 and 3 are based on the Drexel-approach, and Figures 2 and 4 on the conventional approach.

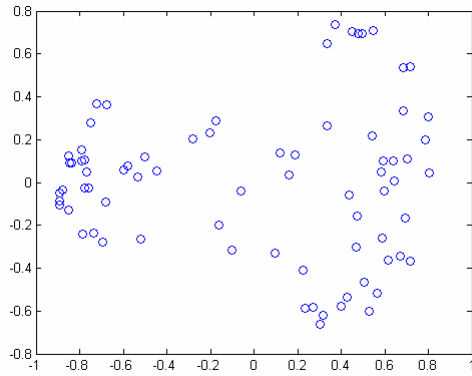


Figure 1. MDS configuration of first-author co-citation analysis – Drexel-approach.

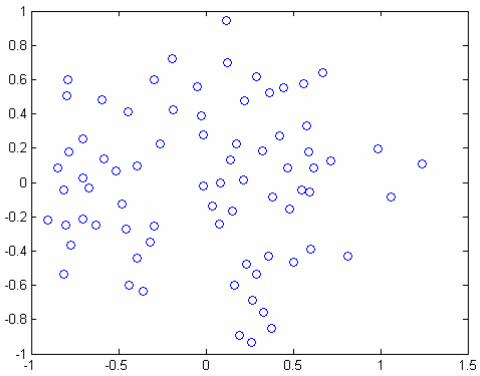


Figure 2. MDS configuration of first-author co-citation analysis – conventional approach.

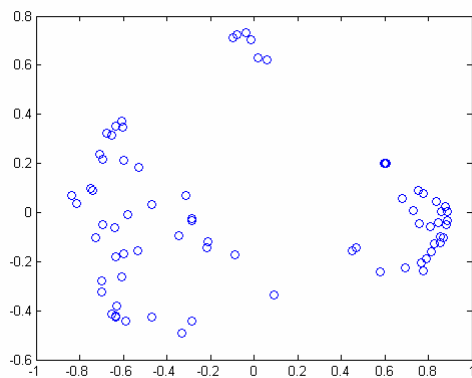


Figure 3. MDS configuration of all-author co-citation analysis – Drexel-approach.

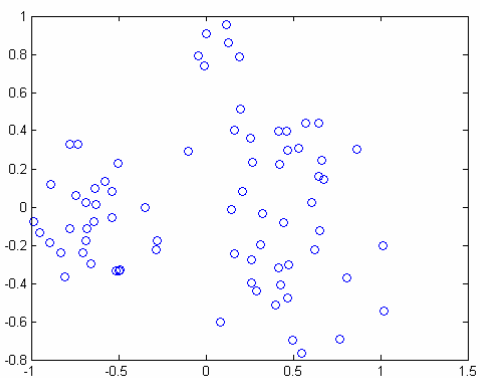


Figure 4. MDS configuration of all-author co-citation analysis – conventional approach.

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 1 below gives the Stress-1 values for the 4 configurations.

Table 1. 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 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 two 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 2D 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 2D. 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.

Comparison of matrix generation techniques

One way of investigating whether the two matrix generation approaches provide different ordinations is to compare their solutions, based on the same proximity measure, for the same set of objects. Procrustes analysis compares and evaluates the resemblance between ordination solutions. Two Procrustes analyses are carried out, one for the two first-author configurations, and one for the two all-author configurations. Note 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 permutation procedure in order to test the significance of the comparison. Figures 5 and 6 below give the Procrustean superimposition plots for the two matching pairs.

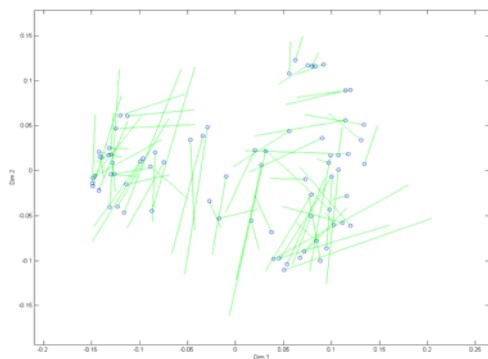


Figure 5. Procrustean superimposition plot of first-author co-citation analysis for the same set of objects based on a Drexel and a conventional matrix generation approach.

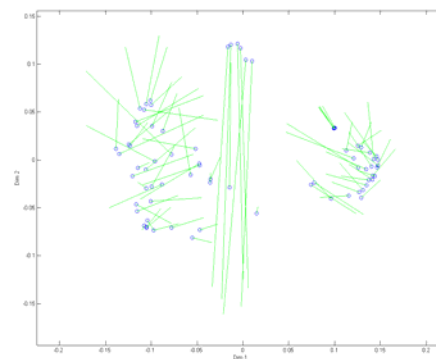


Figure 6. Procrustean superimposition plot of all-author co-citation analysis for the same set of objects based on a Drexel and a conventional approach.

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 significant difference in the ordination produced by the two matrix generation approaches. The circles in the Procrustean superimposition plots, Figures 5 and 6 are the authors plotted from the Drexel-approach configurations. 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. So 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.

First-author 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 among objects. Traditionally major factors in ACA are interpreted as 'research specialties' (cf. White & McCain, 1998). Hence in order to investigate whether the present data set support the previous findings of first-author versus all-author co-analysis we compare the four ACA on the basis of exploratory factor analyses. The most evident indicator 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 practice is to use the Kaiser-Guttman rule.

The Kaiser-Guttman rule for extraction of factors in the present analysis results in a 29-factor model for the conventional approach to first-author co-citation analysis, explaining 59% of the variance; a 25-factor model for the conventional approach to all-author co-citation analysis, explaining 63% of the variance; a 16-factor model for the Drexel approach to first-author co-citation analysis, explaining 80% of the variance; and finally a 15-factor model for the Drexel-approach to all-author co-citation analysis, explaining 86% of the variance. Extraction of a smaller set of factors explaining a majority of the variance in the data set, no doubt implies that the latent structure in the data set should be more explicable and visible. In Zhao (2006) the inclusive all-author co-citation analysis provided a 5-factor model, explaining 97% of the variance, and the first-author co-citation analysis provided an 11-factor model, explaining 96% of the variance – indeed impressive results, which are interpreted so that first-author co-citation analysis provides more specialties compared to all-author co-citation analysis, the latter however provides more coherent groups.

Consider the two factor solutions based on the conventional approach first. As indicated by the Stress-1 values, 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. Obviously, the two data matrices, upon which the conventional approach is based, are extremely sparse. Whereas the two proximity matrices, 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. Dense matrices contain few zero connections and consequently indicate a denser network of objects more suitable for measuring correlation coefficients. 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, and the former from the co-citation matrices) prove this. The Mantel statistic for the first-author correlation matrices gives $\rho = 0.683$ ($p=0.0001$) and the Mantel statistic for the all-author correlation matrices gives $\rho = 0.793$ ($p=0.0001$). This clearly indicates a decreasing monotonicity between the rankings of objects. This is illustrated in Figures 7 and 8.

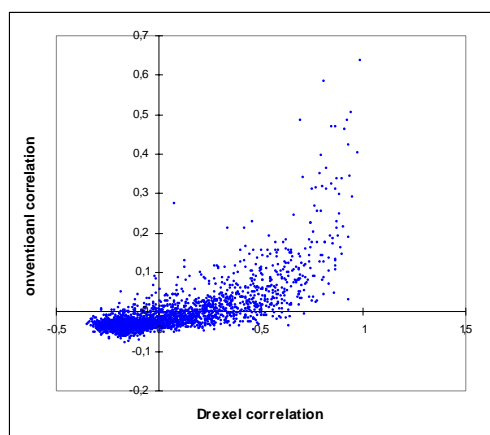


Figure 7. Mantel test for correlation matrices produced by the conventional and Drexel approach for the first-author co-citation analysis.

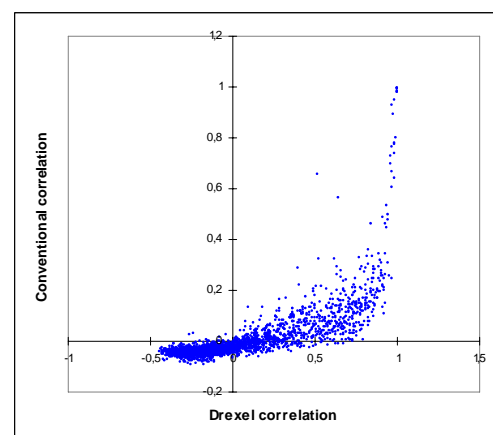


Figure 8. Mantel test for correlation matrices produced by the conventional and Drexel approach for the all-author co-citation analysis.

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, 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). Consequently, even though the all-author data matrix is larger it also contains more non-zero values. The latter results are tremendously important for a factor analysis as it decomposes a covariance or a correlation matrix. Applying factor analysis to a correlation matrix with only low inter-correlations, a likely consequence of 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 10 and 12 below.

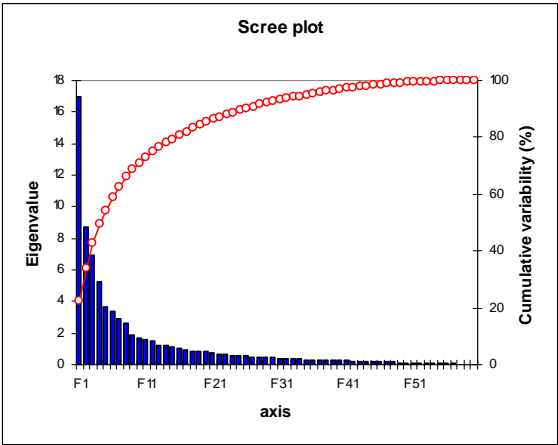


Figure 9. Factor analysis scree plot of first-author co-citation analysis – Drexel approach.

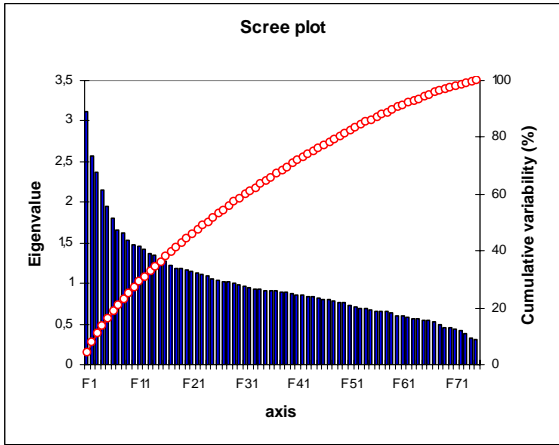


Figure 10. Factor analysis scree plot of first-author co-citation analysis – Conventional approach.

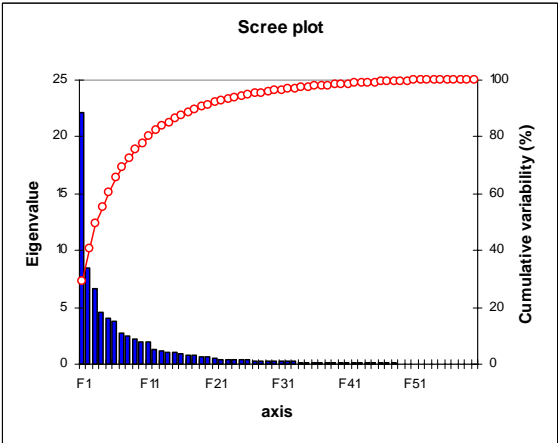


Figure 11. Factor analysis scree plot of all-author co-citation analysis – Drexel approach.

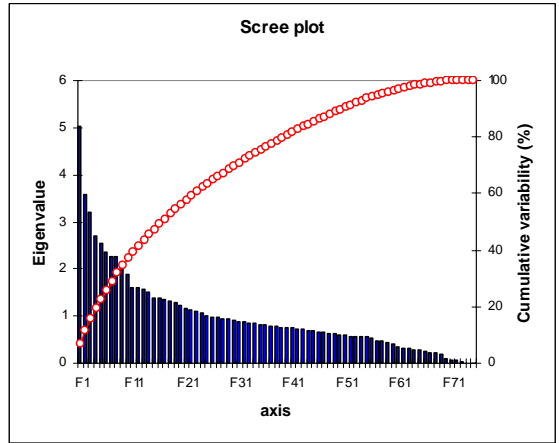


Figure 12. Factor analysis scree plot of all-author co-citation analysis – Conventional approach.

What is desirable when choosing major factors is clearly a shape of the histogram like the ones in Figures 9 and 11, preferably even steeper, and the cumulative variability curves like the one in Figure 11, where the total explained variance is spread among relatively few major factors. Consequently, 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. 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 2 reveals a remarkable similarity.

Table 2. 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				

A close inspection of the factor patterns (not included in the paper), before and after oblique rotation, reveals an even more significant similarity between of 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 left over authors in both cases represents the remaining few factors, not unusually only one author are affiliated with such a factor. Figures 13 and 14 below demonstrate the similarity in structure, for factors 1 and 2, between the two author co-citation counting methods.

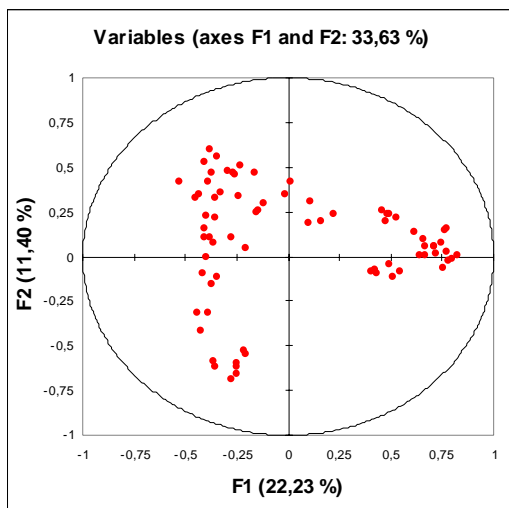


Figure 13. Mapping of two factors for the first-author co-citation analysis.

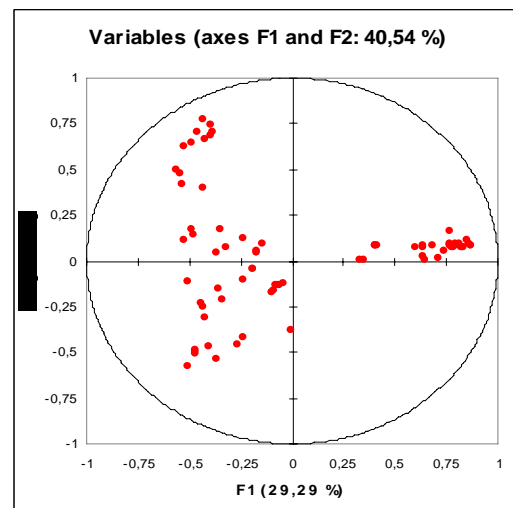


Figure 14. Mapping of two factors for the all-author co-citation analysis.

Perhaps the all-author solution gives a slightly more coherent grouping between the authors? At least its structure does support and elaborate the MDS solution presented in Figure 3. We cannot 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. 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.

Conclusion

In this study we have presented a comparative analysis of first-author versus all-author co-citation analysis on the largest dataset so far used in an all-author ACA study. The results show that the inclusion of all cited authors can aid in producing 2D MDS visualisations that better fit the underlying data (i.e., have lower stress values), and that all-author ACA may lead to stronger concentration in the maps. In addition, we also study the differences between using two approaches for matrix generation: the popular Drexel-approach, as well as a conventional approach based on multivariate statistical analysis. Overall the two methods produce maps that are have some resemblances, but also many differences at the detail level. The Drexel-approach produces results that a have noticeable lower stress values and are more concentrated into groups.

The study also demonstrates the importance of sparse matrices and their potential problems in factor analysis. Sparse matrices are more suitable for principal components analysis. In the present study the sparse matrices deflated the extraction of factors. While, we can confirm that all-author co-citation analysis produce more coherent groupings, the present study cannot confirm previous findings that first-author co-citation analysis identifies more specialties. We need to investigate this further, as the removal of duplicates may have affected this result.

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. In future work we will addresses the research possibilities inherent in a citation index based on source documents formatted in XML.

References

- Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *JASIST*, 54(6), 550-560.
- Chen, C. (1999): Visualising semantic spaces and author co-citation networks in digital libraries. *IP&M*, 35(3), 401-420.
- 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, 89-98.
- Glänzel, W. (1996): The Need for Standards in Bibliometric Research and Technology, *Scientometrics*, 35 (2), 1996, 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* (pp. 138-149). Edinburgh: Edinburgh University Press.
- Lattin, J., Carroll, J.D., & Green, P.E. (2003). *Analyzing Multivariate Data*. Pacific Grove, CA: Brooks/Cole - Thompson Learning.
- Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence Matrices and their Application in Information Science: Extending ACA to the Web Environment. *JASIST*, 57(12), 1616-1628.
- Lok, C.K.W, Chan, M.T.V & 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 et al.: *Proceedings of INEX 2005*. Berlin: Springer 2006, 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. *JASIS*, 41(6), 433-443.
- Persson, O. (2001): All author citations versus first author citations. *Scientometrics*, 50(2), 339-344.
- Rousseau, R. & Zuccala, A. (2004): A classification of author co-citations: definitions and search strategies. *JASIST*, 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. *JASIST* (accepted for publication).
- 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. *JASIST* (accepted for publication).
- 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.
- White, H. D. (2003). Author Cocitation Analysis and Pearson's r . *JASIST*, 54(31), 250–259.
- White, H.D., & Griffith, B.C. (1981). Author co-citation: A literature measure of intellectual structure. *JASIS*, 32(3), 163-171.
- Zhao, D. (2006): *Towards all-author co-citation analysis*. *IP&M*, 42(6), 1578-1591.