

# Going Beyond Counting First Authors in Author Co-citation Analysis

**Dangzhi Zhao**

School of Library and Information Studies, University of Alberta, Edmonton, AB, Canada T6G 2J4. [dzhao@ualberta.ca](mailto:dzhao@ualberta.ca)

The present study examines one of the fundamental aspects of author co-citation analysis (ACA) - the way co-citation counts are defined. Co-citation counting provides the data on which all subsequent statistical analyses and mappings are based, and we compare ACA results based on two different types of co-citation counting - the traditional type that only counts the first one among a cited work's authors on the one hand and a non-traditional type that takes into account the first 5 authors of a cited work on the other hand. Results indicate that the picture produced through this non-traditional author co-citation counting contains more coherent author groups and is therefore considerably clearer. However, this picture represents fewer specialties in the research field being studied than that produced through the traditional first-author co-citation counting when the same number of top-ranked authors is selected and analyzed. Reasons for these effects are discussed.

## Introduction

Since its introduction by White & Griffith (1981), author co-citation analysis (ACA) has gained great popularity in the study of intellectual structures of scholarly fields and of the implied social structures of the corresponding communities. While most studies have applied the general steps and techniques of classic ACA to different research fields with little or no modification, some studies have proposed new techniques for mapping author clusters (White, 2003) or for statistically processing co-citation counts (Ahlgren, Jarneving & Rousseau, 2003). However, few studies (Persson, 2001) have examined some of the fundamental aspects of ACA such as the way that the co-citation counts are defined which provide the raw data on which all subsequent statistical analyses are based.

The present study seeks to contribute to filling this gap, and aims to shed some light on future directions of ACA studies.

## Research questions

ACA is one particular type of co-citation analysis. It is generally accepted that the co-citation concept was discovered independently by Small (1973) and Marshakova (1973), and that document co-citation analysis was introduced by Small (1973) and author co-citation analysis by White & Griffith (1981). Many co-citation analysis studies have been conducted since.

In co-citation analysis, a set of items (authors, documents, journals, etc.) is selected to represent a research area, and relationships between these items are then analyzed using co-citation counts - the number of articles that have cited two items together in the same articles - as similarity measures and multivariate analysis techniques as analysis tools, in order to study the intellectual structure of this research field and to infer some of the characteristics of the corresponding scientific community.

Depending on the units of analysis (documents, authors, journals, etc.) and on the citation / co-citation thresholds, both the macro-structure - the overall map of an entire science with each dot on the map representing a discipline - and the micro-structure - the structure of a single specialty with each dot on the map representing a single document - of a science can be mapped, providing either overviews of research areas or a look at the underlying fine structures (Small, 1999).

ACA takes authors as its units of analysis, which is, when compared with document co-citation analysis, more complex in at least two aspects. First, the interpretation of results is complicated by the fact that an author represents a larger and less homogeneous unit than an article does in terms of what the unit of analysis connotes. This problem was well addressed when ACA was first introduced, as well as in subsequent studies (White, 2003). Second, defining co-citation is tricky because of the existence of multiple authorships. Unlike the first problem, this one has rarely been discussed, partly due to the constraints imposed by the main data source that has so far been used for ACA studies - the set of databases developed by the Institute for Scientific Information (ISI). These databases only index the first authors of cited documents. Although all authors of a cited document can be found if it has also been indexed as one of their source papers (citing papers), the number of such documents, although increasing as time goes on, has been small. As a result, classic ACA only takes into account first authors in the definition of co-citation. Specifically, two authors are considered as

being co-cited when at least one document from each author's oeuvre occurs in the same reference list, an author's oeuvre being defined as all the works with that author as the first author(McCain, 1990). For the purpose of convenience, this is called "first-author co-citation" in the present study.

Taking advantage of citation indexes for scholarly publications on the Web that index all cited authors, the present study attempts to go beyond first-author co-citation and to define an author's oeuvre as all works with this author as one of the authors, which is called "all-author co-citation" in the present study for the purpose of convenience. We hope to see what kind of a picture we can get this way from ACA regarding the intellectual structure of a scholarly field, and how and why it might be different from that resulting from the first-author co-citation-based ACA.

We took a simplified approach to all-author co-citation in that we only took into account the first five authors rather than all authors. We hoped that this approach would approximate sufficiently strict all-author co-citation counts, as publications with more than five authors were not expected to occur too frequently based on the statistics from the present study (Table 1), and even if its approximation was insufficient, it would still help us to see beyond the classic first-author co-citation analysis.

One immediate implication of all-author co-citation is that two authors will also be considered as being co-cited when a single paper which they co-authored is cited. In other words, co-authorship when cited is also counted into co-citation, which may be conceptually confusing at the first thought. However, all-author co-citation appears to be an authentic measure of the connectedness between authors because, just like co-citations, co-authorship indicates that authors are related to each other in some sense, and it is actually a closer relationship between authors than that formed by co-citations. It appears to be an even better measure since this way of counting co-citations takes into account more links between related authors, which may make it easier to identify interrelationships among authors. This was partly confirmed by one of our earlier studies (Zhao & Logan, 2002) and is illustrated in the present study.

It would be interesting to do another analysis which excludes cited co-authorship from all-author co-citation counts to see the differences this might make in the ACA results in order to explore the conceptual implications.

Table 1: Distribution of papers by number of authors

# of authors	Papers retrieved from <i>ResearchIndex</i>	
	#	%
0	4	1
1	83	27
2	77	25
3	78	25
4	36	12
5 or more	34	11

## Methodology

### *Data collection*

The research area we analyzed in the present study was XML - eXtensible Markup Language. Although XML has applications in a wide range of areas, the core of XML research belongs to computer science. Thus, we used the NEC Corporation Research Institute's ResearchIndex (now a joint effort of NEC and the School of Information Science and Technology at Pennsylvania State University) to collect citing papers on XML along with their reference lists. ResearchIndex (aka CiteSeer) automatically indexes research papers that both fall within a broadly defined computer science field and are publicly available on the Web. It is a SCI-like tool freely available on the Web, but provides more information on cited papers than SCI, including their full titles and the names of all authors. Studies have shown that author rankings using data from ResearchIndex are highly correlated with those using data from the ISI's SCI when using identical citation counting methods (Zhao, 2003; 2005), implying that using this tool for citation analysis is just as valid as using SCI. More information about ResearchIndex can be found in Lawrence et al (1999), Zhao & Logan (2002), and Zhao & Strotmann (2004).

We developed a Java program to search for all documents indexed by ResearchIndex in "Header" fields under the term "XML" or "eXtensible Markup Language," and to download all of the records that met the search criteria into a local machine. No citation windows were specified in the present study, indicating that publications from all years were used. The algorithm of this program can be found in Zhao (2003). The actual search was conducted on December 18, 2001.

Another program was developed in Java to parse these records, and to store the resulting citation information such as titles, authors, publishing sources and years of both citing and cited documents in a data structure that was convenient for later data analysis such as counting citations and co-citations using multiple methods. Since an earlier study (Zhao & Logan, 2002) had found the existence of duplicates to be one of the major differences between the ISI databases and ResearchIndex, the citing documents were examined first by another Java program and then manually to remove possible duplicates. Citations made by these duplicates were removed as well. The data structure and the algorithms of these programs can also be found in Zhao (2003).

This way, we collected 312 publications which made 4,578 citations altogether.

### ***Data analysis***

Based on the perception of all the authors of these 312 publications, we conducted an ACA using both first-author and all-author co-citation counts.

We followed the commonly accepted steps and techniques of ACA (McCain, 1990; White & McCain, 1998; White, 2003; Zhao, 2003) except for the way we defined co-citation, as discussed earlier. Core sets of authors were selected based on "citedness"-- the number of citations they received. Two sets of highly visible authors were thus selected using two different citation counting methods -- straight counts and complete counts. Simply put, when a paper with N authors is cited, with straight counts, only the number of citations of the first author of this paper increases by 1, and with complete counts, full credit is given to all authors of the paper, i.e. the number of citations of each of the N authors increases by 1. However, similar to our simplified approach to all-author co-citation counts discussed earlier, we also took a simplified approach to complete counts in that it only took into account the first five authors rather than all authors.

There are no strict rules regarding thresholds for citation-based author selection in author co-citation analysis studies (McCain, 1990). Assuming that the more authors the better a research field is represented, the present study used low thresholds to allow 100 authors to be included in the final multivariate analysis, the maximum number of variables possible when using ALSCAL, the

multidimensional scaling routine in SPSS (version 10.0).

A Java program was developed to count both all-author co-citation frequencies and first-author co-citation frequencies, and to record them in two separate matrixes. These co-citation matrixes were then cleaned by deleting authors who were co-cited with very few other authors based on the assumption that authors who have little connection with the rest of the field are not good representatives of the field. Specifically, an author was deleted if the corresponding row/column contained more than 95% zero value cells. The resulting matrixes were then converted to Pearson's r correlation matrixes that were in turn used as input to the two multivariate analysis procedures employed: Factor Analysis (FA) and Multi Dimensional Scaling (MDS).

Factors were extracted by Principal Component Analysis (PCA) with an oblique rotation (SPSS Direct OBLIMIN). An oblique rotation was chosen because it is often more appropriate than orthogonal rotations when it can be expected theoretically that the resulting factors (in this case, specialties) would in reality be correlated (Hair et al., 1998). The number of factors extracted was determined based on Kaiser's rule of eigenvalue greater than 1 because the resulting model fit was adequate as represented by total variance explained, communalities, and correlation residuals as discussed below (Hair et al., 1998).

The multidimensional scaling procedure used in this study was SPSS ALSCAL as many studies have done (White & McCain, 1998; Kreuzman, 2001) which produced powerful two-dimensional solutions with the squared correlation (RSQ) value and the corresponding Stress 1 measure being 0.99 & 0.05 (Figure 1) and 0.98 & 0.07 (Figure 2) respectively. The two-dimensional maps (MDS maps as shown in Figure 1 & 2) were visualized using LaTeX from the coordinates resulting from the ALSCAL procedure.

With the aid of both factor analysis and multidimensional scaling techniques, the grouping of the scholars within each of the two sets of authors was analyzed, and results compared.

## **Results and discussion**

We will base our discussion mainly on the factor analysis results presented in Table 2 and Table 3, complemented by MDS maps as presented in Figure 1 and Figure 2, because factor analysis applied in ACA has been shown to provide clear and revealing results as

to the nature of a discipline (White and McCain, 1998).

Kaiser's rule of eigenvalue greater than one resulted in a five-factor model from all-author co-citation analysis (Table 2) and an eleven-factor model from first-author co-citation analysis (Table 3). They respectively account for 97.2% and 96% of the total variance, and the differences between observed and implied correlations are for the most part (almost 100%) smaller than 0.05 in both cases. The factor names shown on top of Tables 2 & 3 correspond to column headings indicated by numbers and were given based on an examination of the cited articles written by authors in the corresponding factors. Following White and McCain's example (White & McCain, 1998), authors are ranked in the factor on which they load most highly and their loadings on other factors that are above 0.4, if any, are also presented, indicating their contributions to more than one specialty. If an author does not load 0.4 or higher on any of the factors, the author's highest loading, whatever it may be, is presented.

Since the factor analysis result based on first-author co-citation counts is large in both dimensions (number of authors and number of factors), unlike Table 2 that shows all factors in both left and right half of the table, the left part of Table 3 only shows factors 1 to 3 and all other factors on which the authors who load mostly on factors 1 to 3 have secondary loadings while the right part shows factors 4 to 11 and all other factors on which the authors who load mostly on factors 4 to 11 have secondary loadings.

Table 2: Factor Analysis of 100 authors in the XML research field (all-author co-citation)

1 - XML or semi-structured databases; 2 - Foundations of XML or semi-structured data mgt. & proc.;  
 3 - Programming for / processing of XML data; 4 - The Semantic Web; 5 - XML and Relational Database

Authors	1	2	3	4	5	Authors	1	2	3	4	5
E. Bertino	0.97					L. Libkin		0.97			
L. Tanca	0.94					F. Neven		0.96			
S. Comai	0.94					W. Tan		0.92			
E. Damiani	0.94					Richard Hull		0.89			
S. Paraboschi	0.92					Scott Weinstein		0.89			
P. Fraternali	0.91					C. S. Hara		0.88			
S. Ceri	0.88					Wenfei Fan		0.87			
J. Lapp	0.86					Philip Wadler		0.82			
D. Schach	0.86					Victor Vianu	0.43	0.77			

E. Damiani	0.94			Richard Han	0.89		
S. Paraboschi	0.92			Scott Weinstein	0.89		
P. Fraternali	0.91			C. S. Hara	0.88		
S. Ceri	0.88			Wenfei Fan	0.87		
J. Lapp	0.86			Philip Wadler	0.82		
D. Schach	0.86			Victor Vianu	0.43	0.77	
A. Bonifati	0.85			H. Thompson		0.75	
B. Ludascher	0.74			J. Clark		0.74	0.43
J. L. Wiener	0.74			V. Apparao		0.71	
J. Robie	0.73			M. Champion		0.71	
J. McHugh	0.72			S. B. Davidson	0.45	0.71	
D. Quass	0.72			S. DeRose		0.70	
A. Levy	0.72			H. Hosoya		0.69	0.43
R. Goldman	0.70			B. Pierce		0.68	0.45
G. Hillebrand	0.70			C. Beeri		0.67	
A. Deutsch	0.70			J. Simeon	0.46	0.59	
Y. Sagiv	0.70			Matthew Fuchs		0.58	0.58
M. F. Fernandez	0.69			Tova Milo		0.58	
J. Widom	0.69			K. Smaga	0.44	0.54	
D. Florescu	0.68			M. Murata		0.53	0.46
A. Malhotra	0.68	0.50		W. van der Aalst	-0.46		0.71
D. Suci	0.68			D. Lee	-0.48		0.71
D. Maier	0.67			I. Jacobs			0.69
A. O. Mendelzon	0.66			D. Megginson	0.63		0.65
G. Weikum	0.65			A. Le Hors			0.65
H. Garcia-Molina	0.65			M. Felleisen	-0.51		0.61
S. Abiteboul	0.65			T. Bray		0.53	0.60
J. D. Ullman	0.64	0.61		J. Paoli		0.52	0.60
S. S. Chawathe	0.64			C. Sperberg-MacQueen		0.52	0.59
D. D. Chamberlin	0.63			D. McKelvie	-0.51		0.45
D. Beech	0.63		0.41	R. Guha			0.95
S. Cluet	0.62			S. Decker			0.93
L. Mignet	0.61			D. Brickley			0.92
G. Ghelli	0.59			R. Swick			0.91
M. J. Franklin	0.59			O. Lassila			0.91
Y. Papakonstantino	0.58			Michael Erdmann			0.87



D. Beech	0.63			0.41	R. Guha											0.95
S. Cluet	0.62				S. Decker											0.93
L. Mignet	0.61				D. Brickley											0.92
G. Ghelli	0.59				R. Swick											0.91
M. J. Franklin	0.59				O. Lassila											0.91
Y. Papakonstantino	0.58				Michael Erdmann											0.87
Pavel Velikhov	0.57				Dieter Fensel											0.83
R. Cattell	0.53				I. Horrocks											0.76
C. Delobel	0.51				T. Berners-Lee											0.73
V. Christophides	0.51				Michael Hanus											0.51
P. Buneman	0.50	0.44			M. Carey											0.55
G. Moerkotte	0.47				Gang He										-0.41	0.54
J. Chen	0.45				J. Shanmugasundara											0.54
D. Kossmann	0.44			0.40	Kristin Tufte											0.54
M. Kersten	0.43			-0.40	Chun Zhang											0.51
E. Maler	0.32				David J. DeWitt											0.45
A. Gupta		0.99			J. F. Naughton			0.42								0.44

Table 3: Factor Analysis of 100 authors in the XML research field (first-author co-citation)

1- XML or semi-structured DBs; 2 - Foundations XML data mgt.; 3 - The Semantic Web; 4 - Prog. / proc. XML data; 5- NLP; 6 - Version mgt.; 7 - Functional and Logic Prog.; 8 - DB and IR foundations; 9 – KM; 10 - Access ctl.; 11 - Data integration

Authors	1	2	3	4	8	Authors	1	2	4	5	6	7	8	9	10	11
L. Fegaras	0.84					D. Megginson			-0.84							
S. Adler	0.79					D. Lee			-0.74							
J. F. Naughton	0.74					A. Aho			-0.70							
P. Atzeni	0.61					N. Klarlund			-0.67							
D. Maier	0.56					M. Murata		0.42	-0.64							
D.D. Chamberlin	0.56					R. Bourret			-0.62							
R. Cattell	0.55					E. Maler			-0.61							
M. J. Carey	0.53					D. Fallside			-0.51							
D. Beech	0.48					C.-C. Kanne			-0.44							
F. Bancilhon	0.45	0.43				L. Wood			-0.43	0.43						
J. Shanmugasundar	0.43					A. Schmidt			-0.41							
V. Christophides	0.42					J. Bosak	0.35		-0.37							
J. Widom	0.41					H. Jagadish				0.82				0.4		
J. Miller	0.40					M. Kay				0.74						
S. Cluet	0.39	0.38				N. Walsh				0.67						

D. Beech	0.48			C.-C. Kanne	-0.44		
F. Bancilhon	0.45	0.43		L. Wood	-0.43	0.43	
J. Shanmugasundar	0.43			A. Schmidt	-0.41		
V. Christophides	0.42			J. Bosak	0.35	-0.37	
J. Widom	0.41			H. Jagadish		0.82	0.4
J. Miller	0.40			M. Kay		0.74	
S. Cluet	0.39	0.38		N. Walsh		0.67	
A. Y. Levy	0.38			J. K. Ousterhout		0.66	
D. Florescu	0.36			C. Barras		0.63	
J. McHugh	0.36			D. McKelvie		0.36	
S. Abiteboul	0.35	0.31		A. Albano		-0.33	
A. Deutsch	0.34		0.33	G. Ghelli		-0.33	
M. Fernandez	0.34			L. Liu			0.95
R. Goldman	0.33			A. Marian			0.94
C. Baru		1.03		S.-Y. Chien			0.88
S. Cosmadakis		0.99		J. Chen			0.60
F. Neven		0.97		S. S. Chawathe			0.36
W. Fan		0.92		E. Harold			0.91
R. Ramakrishnan		0.90		H. Boley			0.72
D. Calvanese		0.89		M. Hanus			0.69
V. Apparao		0.88		C. Goldfarb			0.84
J. Ullman		0.84		H. Meuss			0.76
P. Wadler		0.84		M. P. Marcus			0.75
H. Thompson		0.74		G. Navarro			0.74
T. Bray		0.72		R. Baeza-Yates			0.59
C. Beeri		0.67		E. Baralis			0.57
J. E. Hopcroft		0.67	-0.43	J. Paredaens			0.53
J. Clark		0.64		A. Bonifati			0.51
H. Hosoya		0.61		Stefano Ceri			0.45
P. Fankhauser		0.57		A. Aiken			0.45
P. Buneman		0.56		H. Liefke			0.44
A. Davidson		0.53		S. DeRose			0.44
A. Sahuguet		0.49	-0.44	J. Robie	0.33		0.38
L. Cardelli		0.42		C. Freitag			0.81
Y. Papakonstantino		0.40		P. McBrien			0.51
T. Milo		0.38		E. Bertino			1.03
D. Fensel			0.92	E. Damiani			0.90
D. Brickley			0.92	A. Gupta			1.06
O. Lassila			0.91	M. Kifer			0.57
I. Horrocks			0.85	B. Ludascher			0.52
T. Dammers			0.84	C. Neutermann			0.84

Y. Papakonstantino	0.40		P. McBrien	0.51	
T. Milo	0.38		E. Bertino	1.03	
D. Fensel	0.92		E. Damiani	0.90	
D. Brickley	0.92		A. Gupta		1.06
O. Lassila	0.91		M. Kifer		0.57
I. Horrocks	0.85		B. Ludascher		0.52
T. Berners-Lee	0.84		S. Nestorov		0.34
P. Biron	0.75				
S. Decker	0.75				

If major factors are interpreted as specialties, the results of the factor analyses presented in Tables 2 and 3 reveal structures of specialties within the XML research field and the associated authors' memberships in one or more specialties. A comparison between results in Tables 2 and 3 reveals that the major subfield structure of the XML research field is very similar for both first-author and all-author co-citation analysis, similar to an observation by Persson (2001) in the Library and Information Studies research field. This is probably due to the independent existence of the intellectual structure of a field. However, the number of subfields and authors' relative positions are different between the two sets of results.

Both types of co-citation analysis have identified four major specialties in the XML research field: (1) *XML or semi-structured databases*, (2) *Foundations of XML or semi-structured data management and processing*, (3) *programming for or processing of XML data*, and (4) *The Semantic Web*. This can be seen from the fact that almost all authors in these four factors who are common to both sets of top-ranked authors have been placed in the same factors in the two types of analysis.

The first two specialties are the most active ones, as indicated by the number of authors working in those areas (the size of the corresponding factors). They both deal with XML or semi-structured data management but with different emphases: design and implementation versus theory. *The Semantic Web* group has little to do with the other three specialties, which can be seen from the lack of overlap between the corresponding factor and other factors. In fact, this group has a high negative correlation in both types of analysis with the *XML or semi-structured databases* group (-0.638 and -0.31 respectively) and these two groups are located at opposite ends on the MDS maps (Figure 1 and Figure 2). It is different from the rest of the field in that it attempts to add semantics to the Web using technologies such as XML while other specialties essentially deal with a syntactic view of XML.

This structure is quite clear in the results from the all-author co-citation analysis depicted in Table 2, considering that the *XML and*

*Relational Databases* group is highly correlated with the *XML or semi-structured databases* group, as indicated by the correlation coefficient (0.545) given by the oblique rotation procedure. The small group of authors labeled as *XML and relational databases* represent the research focus of mapping of data between Relational Databases and XML, that is, the representation of XML data through relational database or of data in a relational database into XML format. It is closely related to *XML or semi-structured Databases* because they both apply database theory and technology to the management of XML data. The difference is just that one uses relational database and the other semi-structured database technology. This close relationship is confirmed by the same general location of these two groups on the MDS map (Figure 1), and by the merging of the *XML and relational databases* group into the *XML or semi-structured databases* group when different factor models with a smaller number of factors were tried.

This structure can also be clearly seen on the MDS map generated from the all-author co-citation analysis as shown in Figure 1. The group of authors who study *XML and databases* including XML or semi-structured databases and XML and relational databases are located on the right, the *Foundations of XML or semi-structured data management and processing* group along the bottom across from the left to the right, *The Semantic Web* group far away from the others in the upper left corner and the *Programming* group along the X axis on the left.

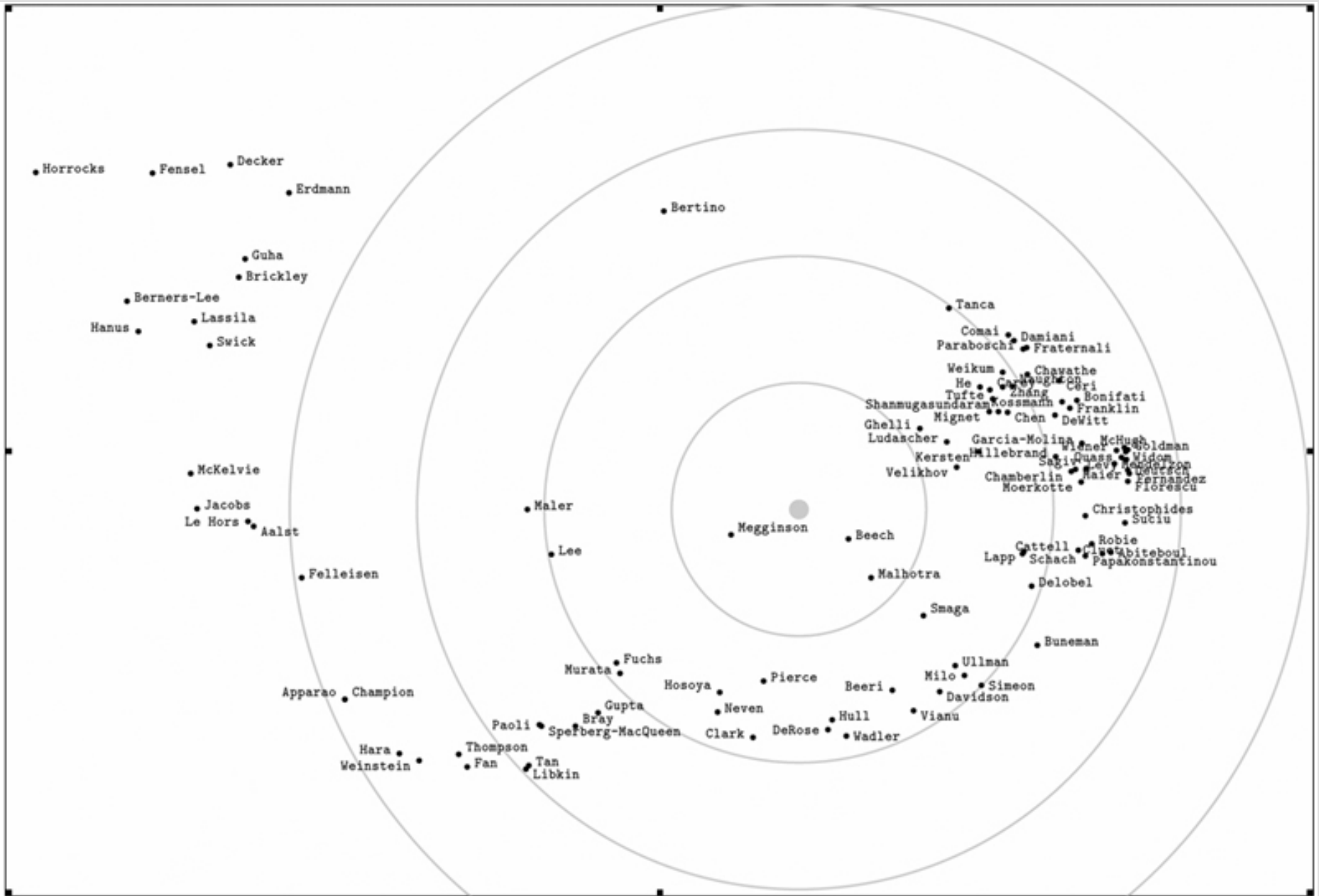


Figure 1: MDS map of top 100 authors in the XML research field (all-author co-citation)

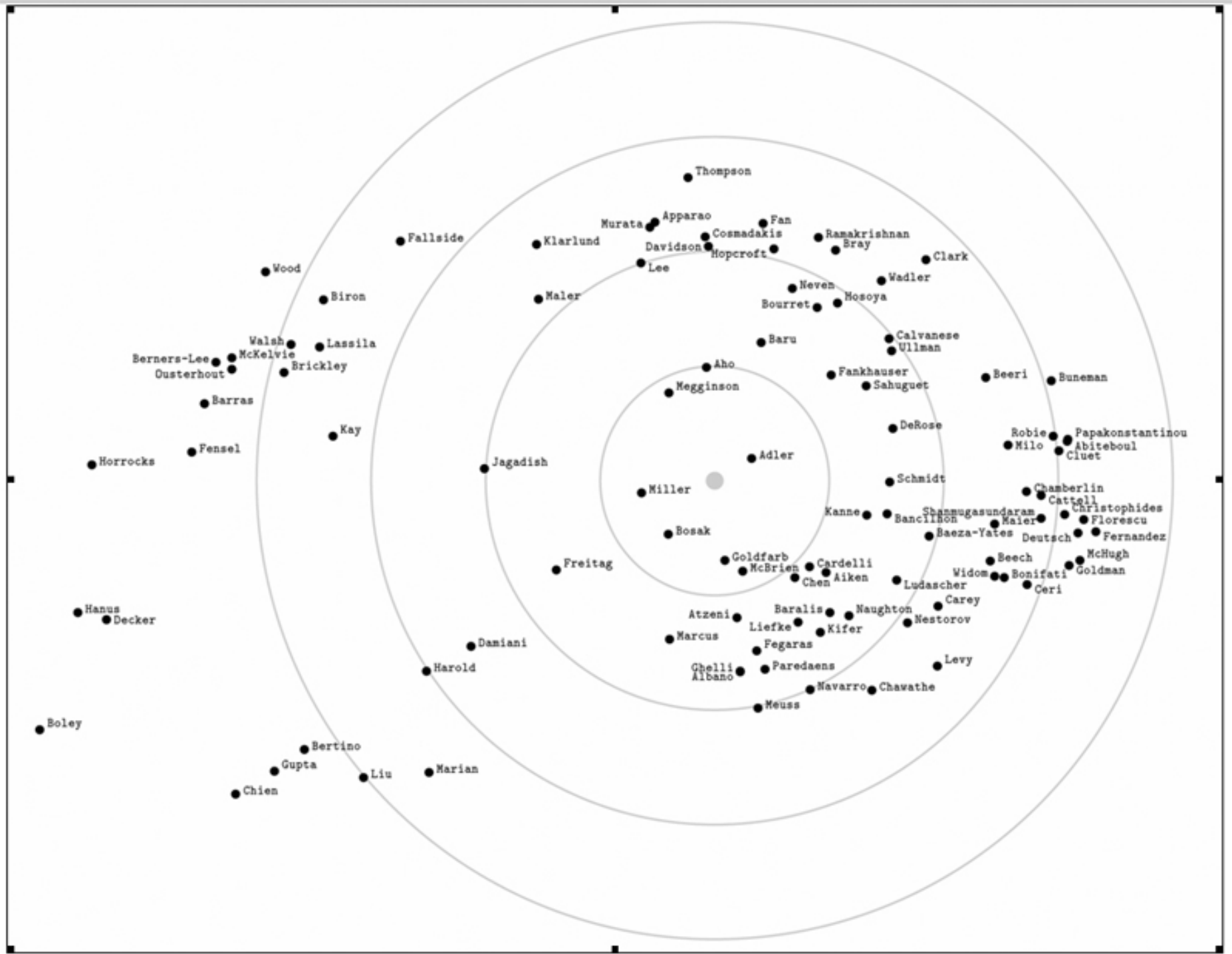
A light dot is placed at the origin and four circles around it to show more clearly the distance of author-points to the origin. The

distance between the first circle and the origin and between any two consecutive circles is the same, namely 0.5.

It appears that one of the two dimensions on the map (X axis) is the degree to which database technology is a research focus. From the left to the right, the importance of database technology becomes more pronounced. This can be seen from the group structure: from *The Semantic Web* group at the far left end that has little to do with databases to the *XML or semi-structured databases* group at the right end in which the database technology is the focus. It can also be seen from the structure inside the *Foundations of XML or semi-structured data management and processing* group at the bottom: the core database people (e.g. Buneman, Ullman, Davidson, and Vianu) are at the right and the XML-related standards and specifications that are not database related in their own right at the left (e.g. Apparao, Champion, Thompson, Paoli, Sperberg-MacQueen and Bray). This reinforces the significant difference between *The Semantic Web* specialty and the *XML or semi-structured databases* group indicated by the high negative correlation (-0.638) between them. The meaning of the second dimension (Y axis), however, is not as readily apparent.

On the MDS map generated from the first-author co-citation analysis (Figure 2), however, the author grouping is not as clear, although the four specialties are identified from the factor analysis results (Table 3). Nevertheless, it is still quite visible on this map that the X dimension represents the degree of database technology being a focus of research.

In addition to these four major specialties that are common to both sets of results discussed above, the first-author co-citation analysis identified other areas of research within the field. The largest two are *Natural Language Processing (NLP)*, which focuses on techniques for representing natural text in XML, and *General database and information retrieval foundations*, in which authors, rather than working with XML per se, have tended to discuss general database and information retrieval technologies that can be applied in XML research. The remaining factors have captured some tightly focused small groups including *Version management*, *Data integration*, *Functional and Logic Programming in XML*, *Knowledge management (KM)*, and *Access Control*.



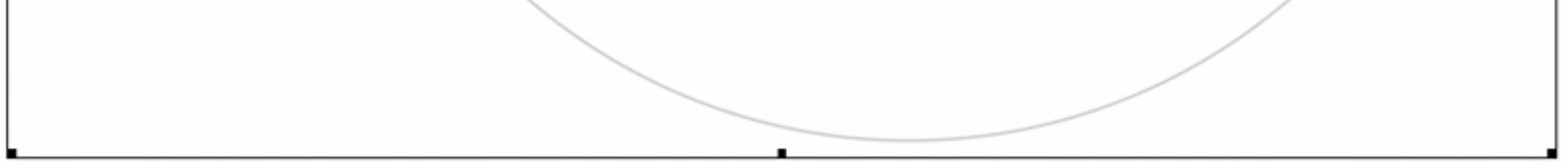


Figure 2: MDS map of top 100 authors in the XML research field (first-author co-citation)

The dot at the origin and the circles around it are drawn in the same way and for the same purpose as Figure 1.

It appears that the major differences between the structure revealed through all-author co-citation analysis and that found through first-author co-citation analysis are (1) the number of specialties identified and (2) the cohesion level of specialties identified. The picture produced through all-author co-citation analysis contains author groups that are more coherent, and is therefore considerably clearer. However, the same picture represents fewer specialties in the research field being studied than that produced through first-author co-citation analysis.

The first of these differences is probably due to the different methods used in selecting the representative authors to be included in the analyses. Since complete counts are used in the all-author co-citation analysis, authors who often publish as co-authors have the same chance as first authors to be selected into the analysis. These authors, however, are not likely to be able to make the cut to the top ranked authors in the first-author co-citation analysis that uses straight counts to select authors. Co-authors usually have been working on the same general topics and their being included in the analysis tends to push out from the list of top ranked authors other authors who may represent smaller, unrelated research areas. As a result, the same number of top ranked authors is likely to represent more research topics in first-author co-citation than in all-author co-citation analysis. In other words, when the same number of highly cited authors is selected, all-author co-citation analysis appears to reveal a picture of the most active research specialties while first-author co-citation analysis can cover more specialties in the research field being studied. This is evidenced by the observation that authors in the all-author co-citation analysis results are very concentrated: the first two large specialties include about 75% of the authors while those in the first-author co-citation analysis results are scattered into many specialties. It would be interesting to test whether more specialties will come out if we include more highly cited authors in the all-author co-citation analysis and what pattern it may follow.

The second difference appears to be due to the co-citation counting methods used in the two types of analysis. All-author co-citation takes into account co-citations received by scholars as authors other than first authors in addition to those as first-authors. It also



counts co-authors of cited articles as being co-cited. In all, considerably more links between scholars are considered in all-author co-citation analysis. As a result, related authors tend to get higher co-citation counts in all-author co-citation than in first-author co-citation analysis, which ties authors in the same groups closer and pulls authors in different groups farther away from each other, resulting in a clearer picture. This can be easily seen from the MDS maps: the four groups in Figure 1 are quite clear-cut whereas those in Figure 2 show considerable overlap even between the four major groups common to these two maps.

## **Conclusion**

We have examined one of the fundamental aspects of author co-citation analysis (ACA) that has rarely been touched since its introduction in 1981, namely the way that the co-citation counts are defined which provide the raw data on which all subsequent statistical analyses and mapping are based. A comparison between first-author and all-author co-citation analyses of the XML research field has indicated that an all-author co-citation analysis, which takes into account more links between related authors, results in a considerably clearer picture of the intellectual structure of a research field than the classic first-author co-citation analysis. Although the same number of authors selected by citedness when counting all authors tends to represent fewer specialties than counting only first authors, this should not be a problem if future studies can confirm that including a larger number of authors in the analysis will increase the number of specialties identified because more recent techniques such as Pathfinder Networks (PFNETs) certainly have this capability, unlike SPSS, which only allows a limited number of variables. For example, PFNETs when applied to ACA allow about 200 author names to be mapped and have shown considerable advantages for ACA (White, 2003).

ACA has been shown to be a powerful approach to the study of scholarly communication. However, as collecting co-citation counts in the print world is nearly impossible without the aid of citation indexes, ACA studies have been relying heavily on the ISI databases, and consequently have been limited to first-author co-citation. As full-text scholarly publications and tools for searching them are becoming increasingly available on the Web, there are now alternatives to the ISI databases for collecting co-citation data that allow us to go beyond first-author co-citation and thus to get a clearer picture of scholarly communication patterns. The present study has offered an example. We are confident that future studies will combine recent advanced information visualization techniques with various co-citation counting methods to produce even more interesting and revealing ACA results.

## Acknowledgements

The author wishes to thank Dr. Andreas Strotmann, Dr. Elizabeth Logan, and Dr. Gary Burnett for their many helpful insights.

## 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*, 54, 550-560

Hair, J.F. Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate Data Analysis* (5th edition). Upper Saddle River, NJ: Prentice Hall.

Kreuzman, H. (2001). A co-citation analysis of representative authors in philosophy: examining the relationship between epistemologists and philosophers of science. *Scientometrics*, 51, 525-539.

Lawrence, S., Giles, C. L. & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6): 67-71.

Marshakova, I. V. (1973). A system of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6, 3-8.

McCain, K. W. (1990). Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science*, 41, 433-443.

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

Small, H. (1973). Cocitation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265-269.

- Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50, 799-813.
- White, H. D. & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49, 327-355.
- White, H. D. (2003). Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science*, 54, 423-434.
- White, H. D. & Griffith, B.C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32, 163-171.
- Zhao, D. (2003). *A comparative citation analysis study of Web-based and print journal-based scholarly communication in the XML research field*. Dissertation, Florida State University. Retrieved Jan. 20, 2005, from [http://etd.lib.fsu.edu/theses/available/etd-09232003-012028/unrestricted/DangzhiZhao\\_dissertation\\_summer03.pdf](http://etd.lib.fsu.edu/theses/available/etd-09232003-012028/unrestricted/DangzhiZhao_dissertation_summer03.pdf)
- Zhao, D. (in press). Challenges of scholarly publications on the web to the evaluation of science –a comparison of author visibility on the web and in print journals. *Information Processing & Management*.
- Zhao, D. & Logan, E. (2002). Citation analysis of scientific publications on the Web: A case study in XML research area. *Scientometrics*, 54, 449-472.
- Zhao, D. & Strotmann, A. (2004). Towards a Problem Solving Environment for Scholarly Communication Research. In *Proceedings of the Canadian Association for Information Science 2004 Annual Conference*, June 3-5, 2004, Winnipeg, Canada.