Coco at the Copacabana: Introducing Co-cited Author Pair Co-citation (Coco) Analysis

Richard Klavans[1], Olle Persson[2], and Kevin W. Boyack[3]

[1] *rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

[2] *olle.persson@soc.umu.se*
Department of Sociology, Umeå University, SE 901 87 Umeå (Sweden)

[3] *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM 87122 (USA)

Abstract
We present a novel method for generating an author-based map of the Information Science literature. This method allows each author to have multiple positions on a map (i.e. multiple identities). This is accomplished by using coco-citation analysis, where the rows and columns in the author matrix are co-cited author pairs rather than single cited authors. We compare our results to those found in White & McCain (1998) and Zhao & Strotmann (2008b) to gain some initial insights into the accuracy of this method. We then illustrate how this approach allows an author to have multiple positions on a map of science.

Introduction
One of the most commonly mapped fields is Information Science. This includes journal maps that show the relationship between information science and other disciplines (Leydesdorff, 2007; Marshakova-Shaikevich, 2005; Van den Besselaar & Leydesdorff, 1996), author maps that show the specialties within Information Science (Persson, 1994; White, 2003; White & McCain, 1998; Zhao & Strotmann, 2008a, 2008b, 2008c), and paper-level maps based on citations and/or text that show the finest degrees of differentiation in this field (Åström, 2007; Janssens, Leta, Glänzel, & de Moor, 2006; Persson, 1994; Van den Besselaar & Heimeriks, 2006).

One of the shortcomings of these maps is that they are all based on the common assumption that objects (journals, authors or papers) should have one and only one position on a map. This is not a serious shortcoming if one is mapping objects that rarely have multiple identities (such as limiting the analysis to journals that are not multidisciplinary). But this may be a serious problem if most objects have multiple identities, as in the case of authors. Specifically, we are concerned that maintaining this constraint will over-aggregate networks within the map and correspondingly create inaccurate representations of the role of a journal, author or paper.

This study explores a novel method for allowing objects in a map to have multiple positions. In this paper we introduce what we call "coco analysis" where "coco" means that the objects to be mapped are object pairs rather than single objects. Object pairs can be co-author pairs, co-cited author pairs, bibliographically coupled author pairs, or similar constructs of papers or journals. The particular example that we present here is that of co-cited author pair co-citation analysis of the Information Science field. This was chosen as the test case to introduce this method because several detailed author co-citation analyses of this field are available for comparison, particularly the classic work of White & McCain (1998, hereafter WM98) and the recent work of Zhao & Strotmann (2008b, hereafter ZS08).

Proceedings of ISSI 2009, pages 265–269.
Edited by B. Larsen and J. Leta.

The balance of the paper proceeds as follows. The first section gives a brief background on two author maps for Information Science (WM98 and ZS08). This is followed by a description of the data and methodology used in this study. We then provide a description of the map using descriptors derived from the 11 factors in the ZS08 map of the same corpus. This is followed by a micro-analysis of the seven clusters in the map associated with a single author, Henry Small. These clusters characterize different aspects or identities of Small's oeuvre. We conclude with a general discussion of the limitations and advantages of this new approach to mapping. In particular, we emphasize two advantages: greater articulation of specialties and a more accurate representation of an author's oeuvres.

Background

Both the WM98 and ZS08 studies used a similar methodology – they selected a sample of the 120 most highly cited authors. Each generated a 120x120 author-author matrix, and used factor analysis to reduce this to a 120x12 (WM98) or a 120x11 (ZS08) author-factor matrix. Authors were then assigned to factors. One can think of these assignments as 'identities' that the author can assume. Many of these authors had multiple identities.

WM98 uses multidimensional scaling to reduce the map to a 120x2 matrix. The fact that some authors have multiple identities is lost. ZS08 uses a more sophisticated visualization approach that simultaneously locates factors and authors. Authors only have one position on the map, while edges (links between authors and factors) are used to communicate the fact that authors have multiple identities. The ZS08 therefore approach maintains information about the multiple identities of authors. This approach appears to work well with sparse networks (only a few authors have two or three identities).

The following analysis suggests that there are far more 'identities' in information science than the 11 or 12 assumed in prior studies. In addition, we show that the networks are not sparse. Many authors are associated with four or more identities. We provide an example where an author (Henry Small) who had just one identity in the WM98 and ZS08 studies has at least four identities using this new method.

Data and Methodology

Data for this study were downloaded from the Web of Science for the years 2001-2005. The 12 journals (see Table 1) we used to define the field of information science were the same as those used by both WM98 and ZS08 with one exception. *Journal of Librarianship and Information Science* was used as a replacement for *Program—Automated Library and Information Systems*, (this journal is no longer indexed by Thomson Reuters).

Author coco-occurrence data is generated in much the same way that author co-citation data is generated. First, we selected a set of highly cited authors. The list of cited authors was generated from the bibliographies of 2,222 downloaded records. In order to somewhat reduce spelling variants of cited authors we keep only the first initial if an author has more than one first name initial. The 154 authors (first authors only) cited 30 or more times were included in the study. This is slightly higher number than the 120 authors used by both WM98 and ZS08. Co-occurrences between pairs of co-cited authors were calculated in the traditional method. The list of co-cited author pairs gets large very quickly. Thus, we limited the list of co-cited author pairs to those with a co-occurrence value of 5 or greater (2,805 pairs). Co-occurrences between these co-cited author pairs were then identified and summed over the entire corpus to generate a square matrix where the rows (and columns) corresponded to co-cited author pairs. Each cell in the matrix could thus have four authors.

Table 1. Journals used to define Information Science along with numbers of ALNR (articles, letters, notes, reviews) retrieved. Field is based on annotation in WM98.

Source	Field	#Records
Journal of the American Society for Information Science and Technology	Info Sci	557
Scientometrics	Info Sci	382
Information Processing & Management	Info Sci	258
Journal of Information Science	Info Sci	215
Electronic Library	Library	196
Journal of Documentation	Info Sci	150
Information Technology and Libraries	Library	126
Library & Information Science Research	Info Sci	104
Library Resources & Technical Services	Library	92
Journal of Librarianship and Information Science	Library	83
Annual Review of Information Science and Technology	Info Sci	59
Proceedings of the American Society for Information Science and Technology	Info Sci	0

Shifting from co-author to coco-author analysis significantly increases the size of the matrix to be analyzed. A normal co-author matrix would require the generation of a 154 by 154 co-occurrence matrix, whereas the author pair co-occurrence matrix is in this case 2,805 by 2,805 in size, and could be larger if we had not thresholded the list of co-cited author pairs. We limited our analysis to the 'top 15' cells for each author. Each author nominates 15 cells – those cells where the raw coco-occurrence is the highest and the author is part of the author pair. This approach allows an author to nominate 15 different positions in the corresponding solution set. The number of potential positions for an author is much higher in many cases because each author can be nominated by other authors. This approach also significantly reduces the numbers of rows and columns (from 2,805 to 631) and the number of non-zero cells in the matrix (from 1,012,139 to 1,499) that are needed to map the author space.

Before generating a cluster solution from these data, the raw coco-occurrence counts were normalized using a cosine index (Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2006). Layout of the graph of co-cited author pairs was done using VxOrd,[1] a force-directed placement algorithm, with maximum edge cutting, resulting in a solution space with 72 clusters. Each cluster consisted of a set of author pairs. We then aggregated coco-occurrence counts to the cluster level (72 by 72), calculated the corresponding cosine index values, and generated a map of the 72 clusters, this time using VxOrd with a no edge cutting setting.

The resulting map is shown in Figure 1. The layout of the 72 clusters is relatively easy to interpret by using the author and factor assignments from ZS08. The lower left set of 27 clusters is mostly associated with areas related to citation analysis. The breakdown (using the labels suggested by ZS08) are webometrics (8), mapping (4), models/distribution (4), scientometrics (3), patent analysis (1), user judgement (1) and unassigned (5). The lower right set of 21 clusters is mostly information retrieval. This breaks down into information retrieval (14), undefined (5) and mapping (2). The upper right set of 21 clusters is mostly associated with information behavior. The breakdown of this group is information behavior (9), childrens information behavior (5), user judgment (3), unassigned (3) and information retrieval (1). The remaining 3 nodes (upper right) are mostly unassigned.

We have also selected four drill-downs (author networks within each cluster). These drill-downs illustrate how an author is allowed to have multiple positions on this map (and

[1] VxOrd is not commonly available. However, the open source code DrL is based on the same equations and technique, and can be used in place of VxOrd.

different positions in the corresponding author networks). We selected Henry Small because (a) he was only given one identity by WM98 and ZS08 and (b) we were sufficiently familiar with his work to make some tentative conclusions about the nature of the author networks.

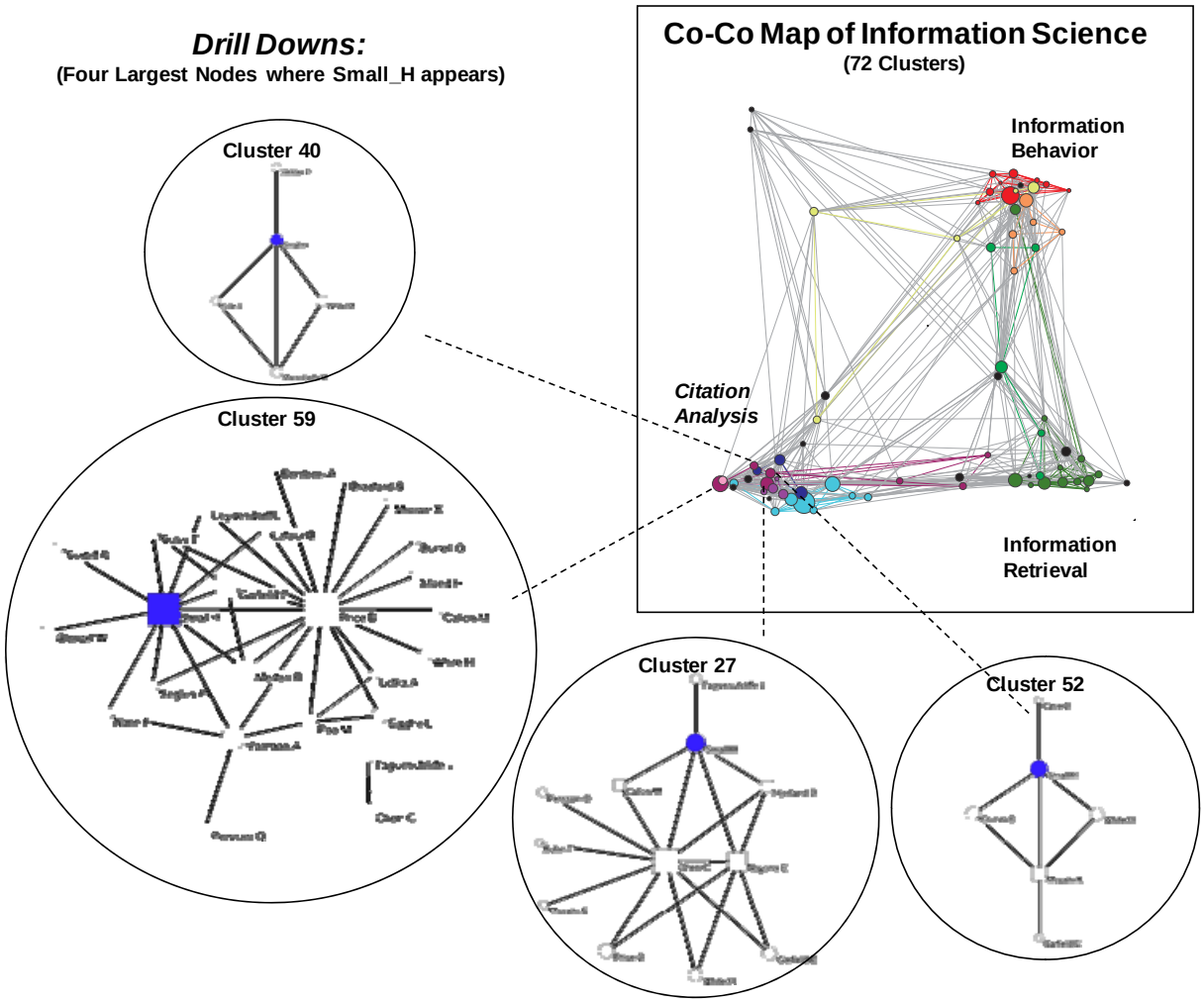

Figure 1. Coco author map of Information Science with four 'drill downs' showing author networks involving Henry Small (filled node).

The largest network (cluster 59) appears to be the major intellectual base for Scientometrics. Derek de Solla Price has the most central role. Henry Small has the second most important role in this network. The next largest network (cluster 27) includes those researchers mostly associated with science mapping; C. Chen occupies the most central position, followed by E. Noyons and H Small. Cluster 52 represents a different network dealing with science mapping focusing more on the work by White and McCain. Cluster 40 deals with the critical view of citation measures (MacRoberts, M. White and Katz). This network represents that portion of Small's work that has addressed these criticisms.

Summary and Implications

We have shown how one can generate an author map of Information Science using coco-analysis. This map shows that Information Science has three broad areas of research (citation analysis, information retrieval and information behaviour) that can be broken down into 72 specialities (far more than the 11 or 12 suggested by traditional approaches). In addition, authors are allowed to have multiple positions in the map (instead of the tradition of allowing an author to have one and only one position).

The results, however, are still preliminary. The most serious shortcoming to this approach is the lack of any threshold that corresponds to role that eigenvalues play in factor analysis. Eigenvalues are used to identify the most important factors (factors with eigenvalues less than 1 are not reported). The corresponding issue in this new approach is to set some type of threshold on the size or structure of the 72 clusters. Stated differently, we suspect that the number of specialties in Information Science is much more than 11 or 12 reported in WM98 and ZS08, but significantly less than the 72 reported in this study.

Despite these, and other, shortcomings, our example illustrates two shortcomings of the traditional approach to author maps of a field: possible over-aggregation of structure and misspecification of the location of authors on a map of science.

References

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. *Journal of the American Society for Information Science and Technology, 58*(7), 947-957.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351-374.

Janssens, F., Leta, J., Glänzel, W., & de Moor, B. (2006). Towards mapping library and information science. *Information Processing and Management, 42*(6), 1614-1642.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57*(2), 251-263.

Leydesdorff, L. (2007). Mapping interdisciplinarity at the interfaces between the Science Citation Index and the Social Science Citation Index. *Scientometrics, 71*(3), 391-405.

Marshakova-Shaikevich, I. (2005). Bibliometric maps of field of science. *Information Processing & Management, 41*, 1534-1547.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science, 45*(1), 31-38.

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

Van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics, 68*(3), 377-393.

Van den Besselaar, P., & Leydesdorff, L. (1996). Mapping change in scientific specialties: A scientometric reconstruction of the development of Artificial Intelligence. *Journal of the American Society for Information Science, 47*(6), 415-436.

White, H. D. (2003). 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., & 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*(4), 327-356.

Zhao, D., & Strotmann, A. (2008a). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics, 2*, 229-239.

Zhao, D., & Strotmann, A. (2008b). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070-2086.

Zhao, D., & Strotmann, A. (2008c). Information science during the first decade of the web: An enriched author cocitation analysis. *Journal of the American Society for Information Science and Technology, 59*(6), 916-937.