Mapping the asymmetrical citation relationships between journals by h-plots

Irene Epifanio[(1)]

*(*1*) Departament de Matemàtiques, Universitat Jaume I*

Corresponding address:

Irene Epifanio

Departament de Matemàtiques

Universitat Jaume I

Campus del Riu Sec. 12071 Castelló (SPAIN)

Tel: + 34 964728390 Fax: +34 964728429 email: epifanio@uji.es

Abstract

I propose the use of h-plots for visualizing the asymmetric relationships between the citing and cited profiles of journals in a common map. With this exploratory tool we can understand better the journal's dual roles of citing and being cited in a reference network. The h-plot is introduced and their use is validated with a set of 25 journals belonging to the statistics area. The relatedness factor is considered for describing the relations of citations from a journal "i" to a journal "j", and the citations from the journal "j" to the journal "i". More information has been extracted from h-plot, compared with other statistical techniques for modelling and representing asymmetric data, such as multidimensional unfolding.

Keywords: H-plot, Asymmetric data, Multidimensional unfolding, relatedness factor, bibliometric mapping.

Introduction

The aim of this paper is to introduce h-plot, a tool for mapping asymmetric proximities, and to examine their utility in bibliometric mappings. It is an exploratory technique that allows representing in a unique map the asymmetric relations. In concrete it will be used for mapping the citing and being cited relations between journals in a particular discipline, with the objective of discovering the relational structure from the selected journals. Note that journals can be cited by journals different from those that they cite, i. e., their being cited profile can be different from their citing profile. Despite the role as citing or being cited can be very different and it is interesting to study them, there are not many papers mapping these asymmetric relations. It is usual to symmetrize the relatedness measures, which allows using well-known dimensionality reduction and visualization methods (Klavan & Boyack, 2006). Although study of cross-references transactions is not novel (Tijssen, De Leeuw and Van Raan, 1987; Leydesdorff, 2006), recently, Schneider (2009) studied the cross-reference activity between journals by means of multidimensional unfolding, which maps journals simultaneously in both their citing and being cited roles.

The structure of the paper is as follows. The next section introduces the methodology and the data considered. Results are presented in the following section, and the paper ends with some conclusions and future work.

Material and methods

Matrices whose elements are intercitation counts are clearly asymmetric, since "i" does not necessarily cite "j" as "j" cites "i". Instead of raw frequencies, normalized frequencies give better results (Klavan & Boyack, 2006). I use the relatedness factor (RF), a normalized frequency, specific to journals, proposed by Pudovkin & Garfield (2002). This intercitation measure was designed to account for varying journal sizes, thus giving a more semantic or topic-oriented relatedness than other measures. The journal relatedness of "i" to "j", $R_{i>j}$, is $R_{i>j} = H_{i>j} * 10^6 / (Pap_j * Ref_i)$, where $H_{i>j}$ is the number of citations in the current year from journal "i" to journal "j" (to papers published in "j" in all years of "j"), Pap_j and Ref_i are the number of papers published and references cited in the j-th and i-th journals in the current year. This definition is quoted literally from Pudovkin & Garfield (2002). The higher the R values are, more related the journals are. The journal relatedness $R_{i>j}$ and $R_{j>i}$ are available in the Journal Citation Reports® (JCR). Note that data are available only for journals that have been cited more than 100 times. Also, R values per journal pair are calculated only if each journal cites the other at least two times.

I use the 2011 JCR edition and journals from the subject category "Statistics & Probability". I have chosen this subject since I am statistician and I know the journals of this field, so I can interpret more easily the results. Instead of using the 116 journals in this subject, I have selected only 25 for illustrating the methodology and comment the results more briefly. These 25 journals appear in JCR from several years. I have considered several journals belonging to different statistics subfields. In particular, three journals from each of the subcategories considered in Wikipedia (2013) have been selected, except for the smallest subcategories and the biggest one, where five journals have been selected. In each subfield, journals with different Impact Factors (belonging to different quartiles) have been selected when was feasible, in order to cover the whole spectrum of the subject. The journals are listed in Table 1, together with their abbreviated journal title, the ranking according to their alphabetical order, and the quartile in the category based on Impact Factor.

TABLE 1. Analyzed journals, organized by subcategories.

Number	Journals	Abbreviations	Quartile
	Introductory and outreach		
1	The American Statistician	AM STAT	Q2
	General theory and methodology		
3	Annals of Statistics, The	ANN STAT	Q1
11	Journal of the American Statistical Association	J AM STAT ASSOC	Q1
16	Journal of the Royal Statistical Society: Series B (Statistical Methodology)	J R STAT SOC B	Q1
21	Scandinavian Journal of Statistics	SCAND J STAT	Q2
24	Statistica Neerlandica	STAT NEERL	Q3
	Applications		
2	Annals of Applied Statistics	ANN APPL STAT	Q1
12	Journal of Applied Statistics	J APPL STAT	Q4
17	Journal of the Royal Statistical Society, Series C: (Applied Statistics)	J R STAT SOC C-APPL	Q3
	Biostatistics		
4	Biostatistics	BIOSTATISTICS	Q1
23	Statistical Methods in Medical Research	STAT METHODS MED RES	Q1
22	Statistics in Medicine	STAT MED	Q1
	Computational statistics		
7	Computational Statistics & Data Analysis	COMPUT STAT DATA AN	Q2
14	Journal of Computational and Graphical Statistics	J COMPUT GRAPH STAT	Q2
18	Journal of Statistical Computation and Simulation	J STAT COMPUT SIM	Q4
	Physical sciences, technology, and quality		
6	Chemometrics and Intelligent Laboratory Systems	CHEMOMETR INTELL LAB	Q1
13	Journal of Chemometrics	J CHEMOMETR	Q1
25	Technometrics	TECHNOMETRICS	Q2
	Social sciences		
5	British Journal of Mathematical and Statistical Psychology	BRIT J MATH STAT PSY	Q2

15	Journal of the Royal Statistical Society, Series A: (Statistics in Society)	J R STAT SOC A STAT	Q1
20	Multivariate Behavioral Research	MULTIVAR BEHAV RES	Q2
	Econometrics		
10	Econometrica	ECONOMETRICA	Q1
8	Econometric Reviews	ECONOMET REV	Q3
9	Econometric Theory	ECONOMET THEOR	Q3
	"Open access"		
19	Journal of Statistical Software	J STAT SOFTW	Q1

The nearness of two journals with the R values is an asymmetric measure ($R_{i>j}$ and $R_{j>i}$ are not necessarily equal, in fact, they can be very different). Furthermore, $R_{i>i}$ can be smaller than $R_{j>i}$ and/or $R_{i>j}$ (for instance, $R_{2>2} < R_{2>3}$ in step 1 of Figure 1), and triangle inequality does not hold: journals "i" and "j", and "j" and "k" can be more or less close, but "i" and "k" very distant. We can observe this fact with the three journals in step 1 of Figure 1 (journal 2 cites and is cited by 1 and 3, but 1 and 3 do not cite among them). Therefore, this measure does not correspond to a metric, neither as a consequence to the Euclidean distance.

	1	2	3
1	$R_{1>1}$ = 1142.25	$R_{1>2}$ = 118.46	$R_{1>3}$ = 0
2	$R_{2>1}$ = 166.89	$R_{2>2}$ = 754.09	$R_{2>3}$ = 1634.36
3	$R_{3>1}$ = 0	$R_{3>2}$ = 327.38	$R_{3>3}$ = 2676.99

Step 1

	1	2	3
1	7	3	2
2	4	6	8
3	2	5	9

Step 2

	1	2	3
1	2	6	7
2	5	3	1
3	7	4	0

Step 3

	Cited 1	Cited 2	Cited 3	Citing 1	Citing 2	Citing 3
1	2	6	7	2	5	7
2	5	3	1	6	3	4
3	7	4	0	7	1	0

Step 4

Figure 1. Illustration of the data processing, with: 1- BRIT J MATH STAT PSY (5), 2- J COMPUT GRAPH STAT (14), and 3- J R STAT SOC B (16). Step 1: Δ, original R values (3 x 3 = 9 values). Step 2: ranking the R values. Step 3: subtracting 9 from the values of the step 2. Step 4: D = [Δ | Δ']. (See text for details).

When proximities are asymmetric, there are several specific techniques for representing them. Borg & Groenen (2005) discuss different models for asymmetric data, such as the Gower decomposition (it fits the skew-symmetric part separately), scaling the skew-symmetry (it fits the

symmetric and the skew-symmetric part separately) or unfolding (it fits asymmetric proximities directly), which is used in Schneider (2009).

In unfolding, data are usually conceived as dissimilarities between the elements of two sets, n_1 individuals and n_2 objects. It attempts to find a common quantitative scale that allows one to visually examine the relationship between the two sets. It is a special case of multidimensional scaling (MDS). Let Δ be an observed dissimilarity matrix of dimension n_1 x n_2 with elements δ_{ij}, indicating the observed dissimilarity from the object "i" to the object "j" .The objective is to find the $(n_1 + n_2)$ x p joint matrix of configurations X, minimizing the following stress function:

$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\delta_{ij} - d_{ij}(X_1, X_2))^2$ with $d_{ij}(X_1, X_2) = \sqrt{\sum_{s=1}^{p}(x_{1is} - x_{2js})^2}$ where X is partitioned into two matrices: X_1 of dimension n_1 x p, which is the individual's configuration, and X_2 of dimension n_2 x p, as the object's configuration matrix. In our case, $n_1 = n_2 = n$ coincide and they are equal to 25. This work has been done with the free software R (R, 2013). We use the library smacof developed by de Leeuw, J. & Mair, P. (2009), for computing unfolding. As usual, I have considered a 2D representation, i.e. p=2. However, local minima are more likely to occur in low-dimensional solutions (de Leeuw, J. & Mair, P., 2009), so I have restarted the algorithm 101 times, using 100 random starts and a rational start (which is the default value for initializing the algorithm) for the initial coordinates. The solution with the lowest stress is retained. The role of objects and individuals can be played by cited and citing profiles interchangeably, and if the global minimum was attained, the solution would be the same without regardless the roles.

Recently, in Epifanio (2013), h-plots were proposed for representing non-Euclidean dissimilarity matrices, including asymmetric data, successfully (it improved the representation obtained by the Gower decomposition and scaling the skew-symmetry). With h-plot, the dissimilarity matrix is treated as a data matrix, and the following variables are represented: the variable measuring the dissimilarity from "j" to other objects, and the variable measuring the dissimilarity from an object to "j", i.e., the citing profile of "j" and the being cited profile of "j". I consider the n x 2n matrix D = [Δ | Δ'] (| indicates that the matrices are combined by columns, and ' indicates the transposition). The variance-covariance matrix of D, S, is estimated. If we are interested in the h-plot in two dimensions, the two largest eigenvalues λ_1 and λ_2 (they are always positive, since S is always positive semi-definite), with corresponding unit eigenvectors q_1 and q_2 of S, can be computed. The matrix giving the configuration is $H_2 = (\sqrt{\lambda_1}\, q_1, \sqrt{\lambda_2}\, q_2)$. The Euclidean distance

between the rows h_i and h_j is approximately the sample standard deviation of the difference between variables "j" and "i". Therefore, if the citing profiles of two journals are similar, they will be represented near to each other, and analogously for the being cited profiles. In one unique representation, the citing and being cited profiles are mapped, and can be compared. Two citing or being cited profiles with a big (respectively small) Euclidean distance between them in the h-plot, are different (respectively similar). A citing profile close to a being cited profile means that they are similar. Therefore, we can know the more (or less) asymmetric journals, in the sense that their citing and being cited profiles are different (or similar), computing the Euclidean distance between these profiles in the representation. H-plots have also the following advantages: 1) they have an explicit solution in terms of eigenvectors, so the local minima problem of unfolding does not exist and we find the same solution when the being cited and citing profiles roles as objects and individuals are interchanged; 2) With h-plot, if the scale of the dissimilarities is linearly modified, the resulting configuration does not change in the sense that the visual configuration will be the same as before; 3) H-plots can be computed for very large matrices (Saad, 2011); 4) The goodness-of fit can be easily assessed by (a high measure, close to 1, indicates a better fit): $(\lambda_1^2 + \lambda_2^2) / \sum_j \lambda_j^2$. See Epifanio (2013) for more details about the methodology.

The objective of the h-plot is not to preserve the interjournal dissimilarity exactly, as with unfolding or multidimensional scaling. Instead, h-plot aims to preserve relationships between dissimilarity variables (profiles). This point of view is especially interesting when non-metric dissimilarities are present, as in this case the dissimilarities cannot be represented exactly in an Euclidean space, because the proximities are not Euclidean. As explained before, our proximities are not Euclidean.

If we have in mind cluster and pattern detection, then an expansion or contraction of the configuration could be more useful (Seber, 1984). For this reason, instead of the R values, I consider the ranking (the first one is the one with the smallest R) of them as in Epifanio (2013) (this kind of relativization is used in plant biology, where asymmetric measures are most useful in analyzing community data) . When some values are equal, ties, we replace them by the maximum, in this way if there are many zeros, they would not be very far away from the rest of the data. Furthermore, I need dissimilarities to apply unfolding and h-plot, so I convert the similarities into dissimilarities by subtracting them from the number of elements of Δ.

This process is summarized in Figure 1. We apply unfolding to the matrix in the step 3 and h-plot to the matrix in the step 4.

Results

The results of the multidimensional unfolding analysis are displayed in Figure 2. On the left hand, the best solution when the role of objects and individuals are played by cited and citing profiles respectively. On the right hand, when these roles are exchanged. Note that they do no coincide due to the local minima problem. The value of the stress function for the map on the left hand is lower than for the map on the right hand, so we can consider the left map as the final unfolding solution. Figure 3 presents the results for the h-plot (the same configuration is obtained when objects-individuals roles are exchanged). The goodness-of-fit for the h-plot is 83.2% for two dimensions, which indicates that is a good representation. The cited positions of the journals are represented in black, whereas the citing positions are circled in red.

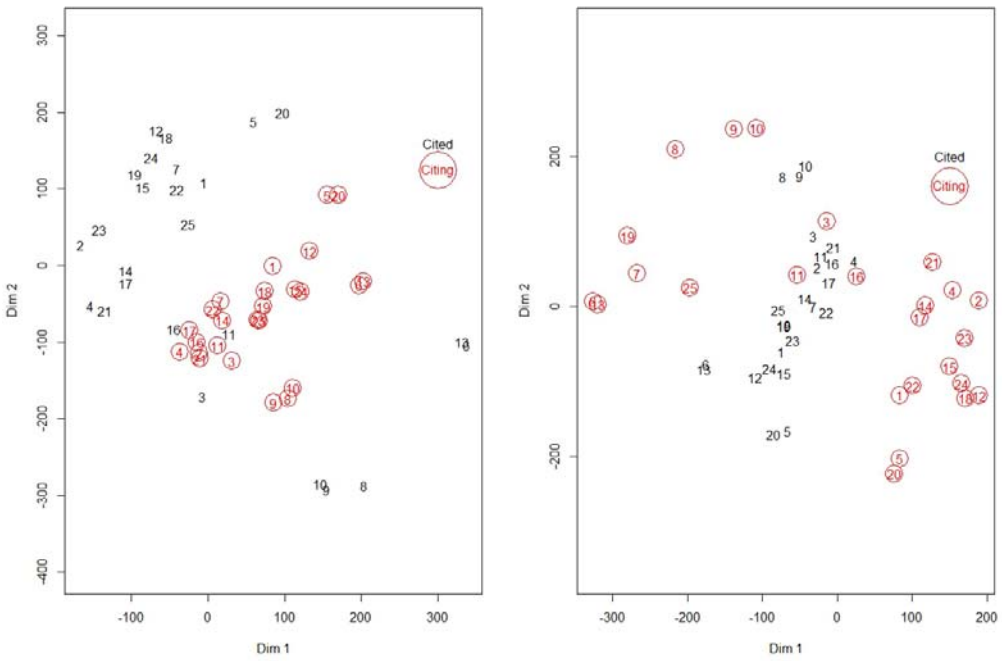

Figure 2. Unfolding solutions of 25 statistics journals. Table 1 gives the codes for the numbers. (See text for details).

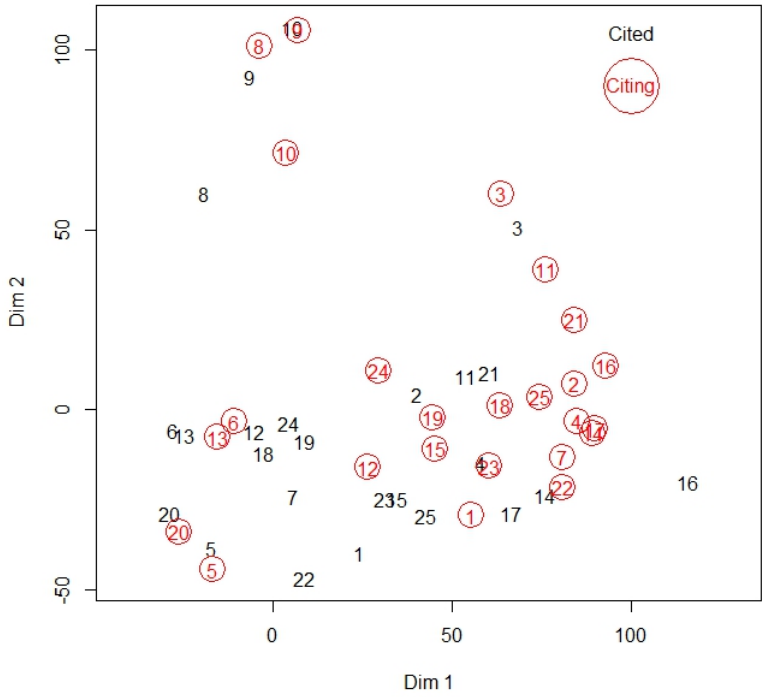

Figure 3. H-plot representation of 25 statistics journals. Table 1 gives the codes for the numbers. (See text for details).

In the plot on the left hand side in Figure 2, the citing positions are in the central part of the map, while the cited positions are in the periphery, except for ANNSTAT (3), JAMSTATASSOC (11) and JRSTATSOCB (16), which are the core set of cited journals. The cited positions for several groups of specialized journals are on the fringe of the map: economics journals (number 8, 9, and 10) are at bottom-right corner, chemistry journals (6 and 13) at three o'clock, and psychology journals (5 and 20) at twelve o'clock. With the unfolding solution we cannot know if a journal has a citing profile similar to the being cited profile of other journal, nor the details of similarities between citing profiles of the journals, since the citing profiles are concentrated in the same zone, except the more specialized journals previously commented. For the map on the right hand side in Figure 2, we find the same difficulties, although the configuration is 'opposite' to the previous one, in the sense that the cited journals are in the middle and the citing on the fringe of the map, except for the citing position of the same three journals (3, 11 and 16). In this solution, other journals besides specialized journals are found in the more extreme positions (for example, COMPUTSTATDATAAN (7)).

For the h-plot in Figure 3, the patterns revealed are richer in details than those for the unfolding solution. First let us focus on the citing profiles. At the top-left we see the cluster of the econometrics journals (8, 9, and 10). At the bottom-left we see the cluster of the psychological journals (5 and 20), and above them the cluster of the chemical journals (6 and 13). At the top-right corner, we find a cluster of journals about general theory and methodology (3, 11, 16 and 21). Below them we find a group of applied and computational journals (2, 4, 7, 14, 15, 17, 18, 19, 22, 23), including journals with applications to all areas of applied statistics and also the biostatistical journals (note that among all the statistical application fields, biostatistics is the most studied due to their importance). Below them we find the introductory journal AMSTAT (1), which publishes "interesting and fun articles of a general nature about statistics and its applications, or the teaching of statistics". Moving more to the left, two journals with particular citing profiles appear: STATNEERL (24) (for example, it does not cite JRSTATSOCB (16), one of the most leading journals in statistical methodology, despite their common aims and scopes, although STATNEERL also publishes papers about probability and operations research) and JAPPLSTAT (12) (for example, it does not cite JCOMPUTGRAPHSTAT (14), one of the most leading journals in computational statistics, which is cited by the rest of journals devoted to applied statistics). Examining the being cited profiles, we see that for specialized journals in econometrics, psychology and chemistry, their cited and citing profiles are near, as it can be expected due to their specialization.

Journals JRSTATSOCB (16) and ANNSTAT (3) are isolated, they have particular profiles. They are leading journals in methodological and general statistics, with a very solid reputation among statisticians, together with JAMSTATASSOC (11) (these three journals have the highest ranking in the survey by Theoharakis & Skordia, 2003). However 3 and 16 are not cited by the majority of specialized or more applied statistics journals, unlike JAMSTATASSOC (11), which is cited by all the journals except the psychological journals. On the one hand, ANNSTAT (3) is not cited by the chemical and psychological journals nor AMSTAT (1) and JAPPLSTAT (12). On the other hand, JRSTATSOCB (16) is not cited by the chemical, psychological, and the majority of economical journals nor STATNEERL (24) and JSTATSOFTW (19).

The cited positions for the rest of journals are distributed at the bottom- central of the map. Some interesting findings about these positions are: 1) STATMED (22) is near the psychological journals, since it is highly cited by them; 2) JAPPLSTAT (12), JSTATCOMPUTSIM (18),

JSTATSOFTW (19) and STATNEERL (24) form a cluster. They are self-cited but scarcely cited by others journals. Curiously, the journals number 12, 18 and 24 have the lowest impact factor of the list, but JSTATSOFTW (19) has the highest impact factor. Note that JSTATSOFTW (19) is more cited by journals outside its category.

As regards to the relation between the citing and being cited profiles for each journal, the most symmetrical journals, i.e. those that cite and are cited in a similar way, are in this order: the psychological journals, MULTIVARBEHAVRES (20) and BRITJMATHSTATPSY (5), JCHEMOMETR (13) and ANNSTAT (3). While the most asymmetrical are in this order: STATMED (22), COMPUTSTATDATAAN (7) and JSTATCOMPUTSIM (18).

We can also examine the relation between the citing and being cited profiles between different journals. For example, the citing profile of ECONOMETTHEOR (9) is similar to the being cited profile of ECONOMETRICA (10), and the citing profile of STATMETHODSMEDRES (23) is similar to the being cited profile of BIOSTATISTICS (4).

The plots in Figures 2 and 3, with the journal abbreviated names, together with the code and data for reproducing this work are available at http://www3.uji.es/~epifanio/RESEARCH/asyhplot.rar .

The proximity matrix could be seen as a directed and weighted graph that could be represented using social analysis visualization techniques. The conceptual differences between these techniques and my MDS-like visualization can be seen in Leydesdorff & Rafols, 2012. However, in this case, the adjacency matrix is quite dense and the network representation is difficult to read (Ghoniem, Fekete & Castagliola, 2005). At http://www3.uji.es/~epifanio/RESEARCH/asyhplot.rar the representation with the 25% strongest citation links with a spring embedder as layout (Butts, 2008), as in Calero Medina & van Leeuwen, 2012, is available as figure 4.

Conclusions and future work

This paper shows the possibilities of the h-plots as a useful and straightforward technique for mapping and analysing asymmetric relations between journals, as regards their citing and being cited roles, in a unique representation. This technique, as well as unfolding, is able to model both roles at the same time in contrast to the majority of previous studies. However, more information has been extracted from the h-plot solution than that from unfolding, which suffers from some problems, such as local minima. We have also summarized some of the advantages of h-plots.

Note that although I have used R values for defining the asymmetric proximities, other asymmetric proximity could be used. In fact, I guess that a more robust measure would be to use more years

than the current year in the definition of $H_{i>j}$, since this estimation could be very variable with journals with low publication activity due to they have few references. The precision will be greater with a bigger sample size. This is an open question to investigate. As future work more ideas are: 1) the application of h-plot to study the journals of other fields or several fields (or all fields); 2) the application to other bibliometric entities, such as a set of authors.

ACKNOWLEDGEMENTS

I wish to thank ISI® for permission to use the JCR data for this study. This work has been partially supported by grants CICYT TIN2009-14392-C02-01, MTM2009-14500-C02-02, and Bancaixa-UJI P11A2009-02. I also thank my family for their support.

References

Borg, I. & Groenen, P.J.F. (2005). Modern multidimensional scaling: Theory and applications, 2nd edition, New York: Springer Science.

Butts, C.T. (2008). Social Network Analysis with sna. Journal of Statistical Software, 24(6): 1-51.

Calero Medina, C.M. & van Leeuwen, T.N. (2012). Seed Journal Citation Network Maps: A Method Based on Network Theory. Journal of the American Society for Information Science and Technology, 63(6):1226–1234.

de Leeuw, J. & Mair, P. (2009). Multidimensional scaling using majorization: The R package smacof. Journal of Statistical Software, 31(3), 1-30.

Epifanio, I. (2013). h-plots for displaying nonmetric dissimilarity matrices. Statistical Analysis and Data Mining, 6 (2), 136–143.

Ghoniem, M., Fekete, J.-D., and Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. Information Visualization, 4(2): 114-135.

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.

ISI Journal Citation Reports. http://www.isinet.com/isi/products/citation/jcr/index.html

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? Journal of the American Society for Information Science & Technology, Vol. 57(5), 601-613.

Leydesdorff, L., & Rafols, I. (2012). Interactive Overlays: A New Method for Generating Global Journal Maps from Web-of-Science Data. Journal of Informetrics, 6(3), 318-332.

Pudovkin, A.I., & Garfield, E. (2002). Algorithmic procedure for finding semantically related journals. Journal of the American Society for Information Science and Technology, 53(13), 1113-1119.

R Development Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Saad, Y (2011). Numerical Methods for Large Eigenvalue Problems: Revised Edition. SIAM - Society for Industrial & Applied Mathematics.

Schneider, J.W. (2009). Mapping of cross-reference activity between journals by use of multidimensional unfolding: Implications for mapping studies. Proceedings of 12th International Conference on Scientometrics and Informetrics (ISSI 2009), 443-454.

G. A. F. Seber. Multivariate observations. John Wiley, 1984.

Theoharakis, V., Skordia, M. (2003). How do statisticians perceive statistics journals? American Statistician, 57 (2), 115-124.

Tijssen, R.J.W., De Leeuw, J. & Van Raan, A.F.J. (1987). Quasi-correspondence analysis on scientometric transaction matrices. Scientometrics, Vol. 11(5-6), p. 351-366.

Wikipedia contributors, "List of statistics journals," Wikipedia, The Free Encyclopedia. Retrieved April 20, 2013, from http://en.wikipedia.org/w/index.php?title=List_of_statistics_journals&oldid=551166550