Manfred Bonitz and the Matthew Effect: Quantitative Content Analysis of Citation Contexts

Loet Leydesdorff

Amsterdam School of Communication Research, University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands; loet@leydesdorff.net

Introduction

In this brief communication at the occasion of Manfred Bonitz' 80th birthday, I focus on citations of his work in the social sciences. Manfred is best known (to me) for his empirical testing of the *Matthew Effect* (e.g., Bonitz *et al.*, 1997 and 1999; Bonitz & Scharnhorst, 2001). Where is this contribution cited and in which citation contexts? Can a semantic map of these citation contexts inform us about the position of Manfred's work and how can this method be improved? (Leydesdorff & Welbers, in press) Could one perhaps automate citation context analysis in this way? (Small, 1982)

The Matthew Effect – formulated as follows: "For to all those who have, more will be given, and they will have an abundance; but from those who have nothing, even what they have will be taken away" –

was introduced to the sociology of science by Robert K. Merton. Merton (1968) argued that obtaining credit in science is a self-reinforcing process. Barabási (2002) later reintroduced this process as the mechanism of "preferential attachment" and formulated accordingly an algorithm in social network analysis (cf. Price, 1965).

Manfred is primarily a physicist. He published hitherto 212 papers (since 1975 and using the Web of Science in November 2010) of which 48 were attributed to the *Social Science Citation Index*. The 212 papers were cited in 1,007 unique documents contained in the (*Social*) *Science Citation Index*. The titles of these citing documents contain 7,257 non-trivial words. The frequency distribution of these title words shows the predominance of (nuclear) physics in this *œuvre*: "plasma" and "physics" lead the list which each 298 occurrences. "Matthew" follows only at the 116th position with 13 times; not so far behind "scientometrics" on the 76th position with 17 occurrences.

Let us focus on the 48 papers included in the *Social Science Citation Index*. These were cited 165 times by 119 unique documents of which I could retrieve (at the WoS interface) only 113. In these documents 55 words occurred more than twice. I use this set for a co-word analysis and the generation of a semantic map. The method is also further developed.

Results

The basic word-document matrix contains 113 cases (citing documents) and 55 variables (words occurring more than twice in titles of these documents). Factor analysis of the matrix suggests seven orthogonal dimensions (Varimax; SPSS) which cumulatively explain 36.8% of the variance in the matrix. Figure 1 shows the resulting cosine-normalized network data using the algorithm of Kamada & Kawai (1989) for the visualization in Pajek. The nodes are coloured in accordance with the seven factors distinguished on the basis of the so-called scree-plot of the eigenvalues (Figure 2).

The Matthew Effect is indicated by Factor 5 and colored pink in Figure 1. Other words which load on this factor (as variables) are "core," "countries," and "concentration. Other groups are also recognizable, such as a group of words colored green: "impact," "factor," "journal," "evaluation," "research," "researcher," "ranking," and "parameter." Additionally, "scientometric,"

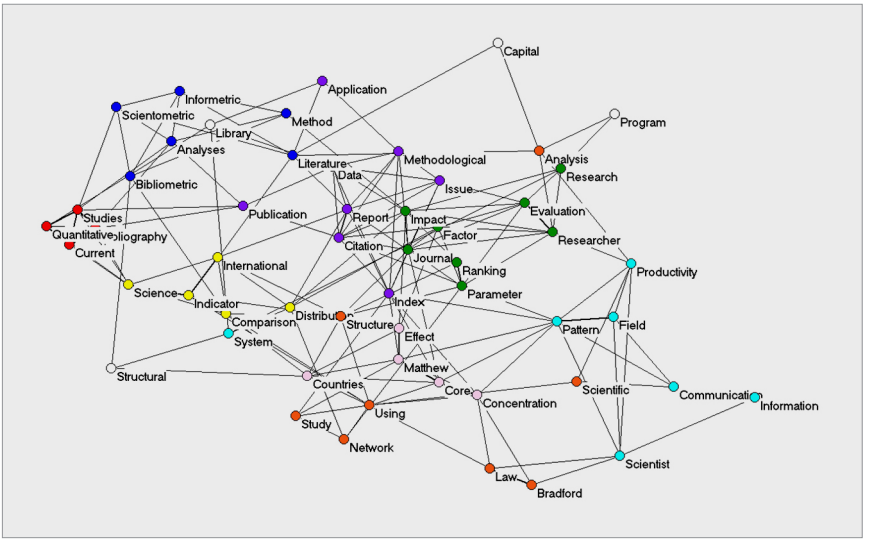

Figure 1: Cosine-normalized map of 55 co-words in 113 titles of citing documents; cosine > 0.2; Kamada & Kawai (1989).

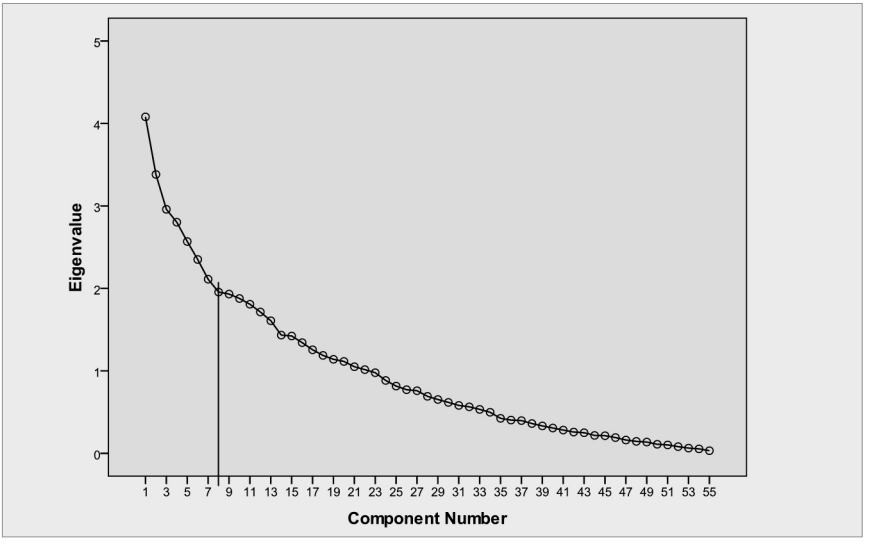

Figure 2: Screeplot of the eigenvalues of principal components using 55 title words in 113 citing documents.

"bibliometric," and "informetric" "analyses" of "literature" and "method" form a blue group at the top-left side of Figure 1. (This grouping corresponds to the negative factor loadings on Factor 6.)

Figure 2 indicates that after seven factors, the so-called "scree" of the hill begins in terms of the distribution of eigenvalues. However, the positive and negative loadings on Factor 6 provide two different groupings, colored blue and light blue in Figure 1, respectively, and on different sides of the figure. The light-blue colored factor covers words such as "productivity," "field," "scientist," "communication," information," and "pattern."

In my opinion, a problem with co-word analysis and semantic mapping is that each combination of words easily suggests an interpretation (Leydesdorff, 1991, 1997). In Leydesdorff (1995), I suggested and elaborated an algorithmic approach based on entropy statistics, but the visualizations are then less attractive and the reasoning is more difficult to follow. In a recent paper, Leydesdorff & Welbers (in press) reviewed the possibilities to improve the statistics, and suggested to use instead of observed frequencies the ratio of observed and expected frequencies of word occurrences. Let me apply this technique to this set and see whether the results can be improved and perhaps be more convincing.

Observed/Expected

A cell value in a matrix (or contingency table) can be measured against its expected value given the other values in this matrix. For example, if one has a matrix with four value 3, 5, 2, and 0 such as in:

3	5	*8*
2	0	*2*
5	*5*	*10*

One can add the margin totals and grand sum of this matrix and compute the expected value for each cell (e_{ij}) from the observed ones (o_{ij}) using

$$e_{ij} = \frac{\sum_i o_{ij} \sum_j o_{ij}}{\sum_i \sum_j o_{ij}} .$$

For example, the expected value of the first cell (e_{11}) above is (8 *5)/10 = 4. The observed/expected ratio consequently is 3/4. (Observed and expected ratios can

also be compared using the formula for χ^2 so that observed values can be tested for statistical significance. See Leydesdorff & Welbers (in press) for more details.)

In the meantime, the programs ti.exe (at http://www.leydesdorff.net/software/ti) and fulltext.exe (at http://www.leydesdorff.net/software/fulltext) for co-word analysis and semantic mapping of titles and texts, respectively, were extended with the option to choose for repeating the analysis with observed/expected ratios as cell values instead of (and after) the analysis with observed frequencies. The conceptual advantage is a normalization. (Other normalizations such as "term-frequency/inversed document frequency" or "tf-idf" are also possible, but in my opinion less easily connected to social-science statistics.)

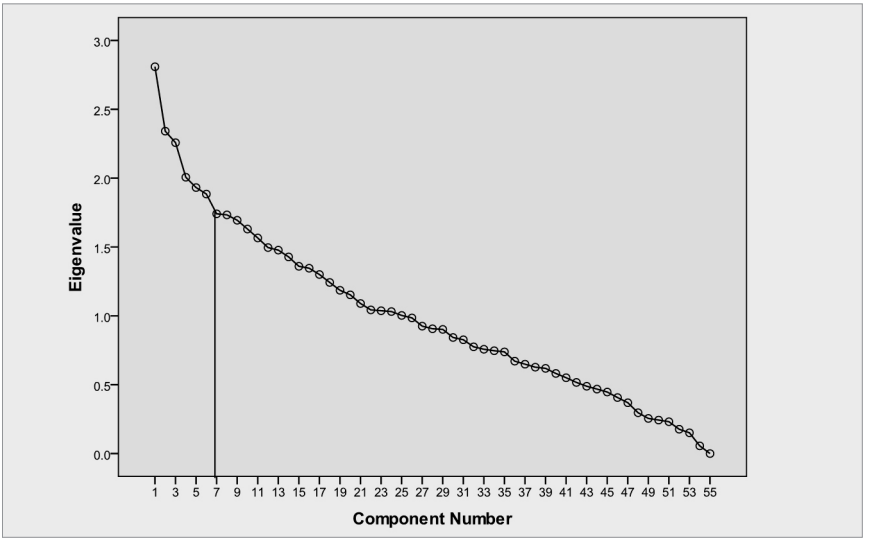

Figure 3: Screeplot of the eigenvalues of principal components using observed/expected values for 55 title words occurring in 113 citing documents.

The observed/expected matrix contains a structure different from the matrix based on observed values. Figure 3 shows the scree plot for precisely the same analysis as above (Figure 2), but now using the obs/exp matrix. Six instead of seven factors are indicated. These six factors explain only 24.1% of the variance. (Seven factors explain 27.2% of the variance.) Thus, the explained variance is lower; the normalization corrects for semantic structure that is incorrectly inferred from the raw data.

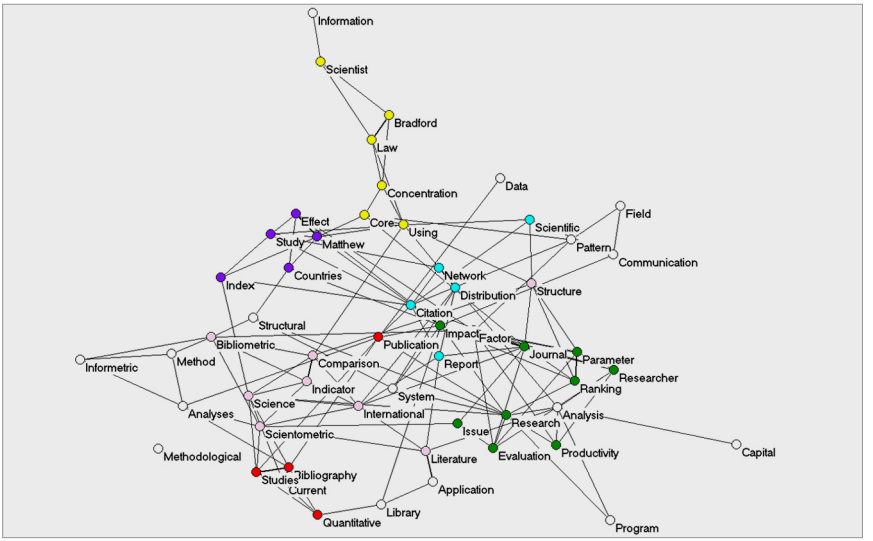

Figure 4: Cosine-normalized map of the observed/expected values of 55 co-words in 113 titles of citing documents; six factors; cosine > 0.1; Kamada & Kawai (1989).

In Figure 4 – coloured on the basis of the six-factor solution – the two groups colored pink and orange in Figure 1 are now clearly distinguished as a a group of words related to the study of the Matthew Effect as an "index" for "countries" (violet) versus a group (in yellow) focusing on the shape of the distribution. The fine-structure of Factor 6 is resolved in Figure 4.

The advantage of Figure 4 is that the coloring of different areas is more contingent than in Figure 1. Exceptions such as the word "Publication" are caused by differences between the use of the cosine for the normalization in the vector space and the Pearson correlation underlying the factor analysis. This problem can be circumvented by using the Pearson correlation also for the mapping (cf. Ahlgren *et al.*, 2003) or by using the factor matrix directly as input to Pajek. The (Varimax) rotated factor matrix is visualized as a 2-mode matrix in Figure 5.

Distances in Figure 5 are based on factor loadings. Factor loadings are equal to the Pearson correlation coefficients between the variable vector and the latent eigenvector or factor. Negative factor loadings are dashed, but also used in the spanning of the map. This is a major advantage over the representations in Figures 1 and 4. However, this representation may be more difficult to explain to a lay audience.

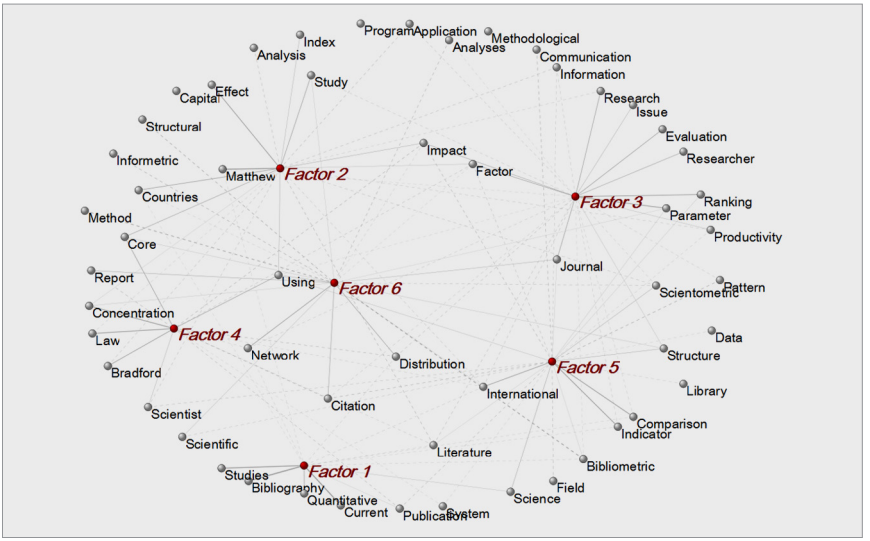

Figure 5: Visualization of the six-factor solution of observed/expected values in the word-document matrix of 55 title words in 113 citing documents; factor loadings > 0.1 or < −0.1 included; Fruchterman & Rheingold (1991).

Conclusion

Using different co-word maps, I explored the fruitfulness of Manfred Bonitz' scientometric contribution for the social sciences and the further development of theorizing about the Matthew Effect. This research question provided me with an opportunity to demonstrate some recent advances in quantitative content analysis (Danowski, 2009). These techniques and methods, among other things, enable us to objectify and automate citation context analysis (Amsterdamska & Leydesdorff, 1989; Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975; Small, 1982).

References:

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirement for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550-560.

Amsterdamska, O., & Leydesdorff, L. (1989). Citations: Indicators of Significance? *Scientometrics 15*(5-6), 449-471.

Barabási, A.-L. (2002). *Linked: The New Science of Networks*. Cambridge, MA: Perseus Publishing.

Bonitz, M., Bruckner, E., & Scharnhorst, A. (1997). Characteristics and impact of the Matthew effect for countries. *Scientometrics, 40*(3), 407-422.

Bonitz, M., Bruckner, E., & Scharnhorst, A. (1999). The Matthew Index – concentration patterns and Matthew core journals. *Scientometrics, 44*(3), 361-378.

Bonitz, M., & Scharnhorst, A. (2001). Competition in science and the Matthew core journals. *Scientometrics, 51*(1), 37-54.

Chubin, D. E., & Moitra, S. D. (1975). Content analysis of references: Adjunct or alternative to citation counting? *Social studies of science, 5(4),* 423-441.

Danowski, J. A. (2009). Inferences from word networks in messages. In K. Krippendorff & M. A. Bock (Eds.), *The content analysis reader* (pp. 421-429). Los Angeles, etc.: Sage.

Fruchterman, T., & Reingold, E. (1991). Graph drawing by force-directed replacement. *Software--Practice and Experience, 21*, 1129-1166.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7-15.

Leydesdorff, L. (1992). A Validation Study of "LEXIMAPPE". *Scientometrics, 25*, 295-312.

Leydesdorff, L. (1995). *The Challenge of Scientometrics: The development, measurement, and self-organization of scientific communications*. Leiden: DSWO Press, Leiden University; at http://www.universal-publishers.com/book.php?method=ISBN&book=1581126816.

Leydesdorff, L. (1997). Why Words and Co-Words Cannot Map the Development of the Sciences. *Journal of the American Society for Information Science, 48*(5), 418-427.

Leydesdorff, L., & Welbers, K. (in press). The semantic mapping of words and co-words in contexts. *Journal of Informetrics*.

Merton, R. K. (1968). The Matthew Effect in Science. *Science, 159*, 56-63.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social studies of science, 5(1),* 86-92.

Price, D. J. de Solla (1965). Networks of scientific papers. *Science, 149* (no. 3683), 510- 515.

Small, H. (1982). Citation context analysis. *Progress in communication sciences, 3*, 287–310.