

## BRIEF COMMUNICATION

# Empirical Evidence of Self-Organization?

**Peter van den Besselaar**

*Department of Social Sciences, Netherlands Institute of Scientific Information, Royal Netherlands Academy of Arts and Sciences (NIWI-KNAW), PO Box 95110, NL-1000 HC Amsterdam, The Netherlands. E-mail: peter.van.den.besselaar@niwi.knaw.nl*

**In a recent paper in this journal, Loet Leydesdorff and Gaston Heimeriks (2001, *Journal of the American Society for Information Science and Technology*, 52, 1262–1294.) argue that biotechnology develops in a self-organizational mode, through interaction between the intellectual structure and the institutional network of the research field. This claim is empirically supported by a multivariate analysis of documents from core biotechnology journals. One unexpected finding in this paper is the relationship between the title words of documents and the region of their origin. This claim requires examination because, as will be shown, it seems to be an artifact of the method used. If this is so, it undermines the authors' theoretical claim that the production of knowledge is a self-organizing process.**

### The Argument

Loet Leydesdorff and Gaston Heimeriks (L&H) argue that the development of the techno-sciences is based on the interaction of the intellectual organization and the institutional network of the field. They use this model to study the effect of Europeanization on knowledge production in the case of biotechnology. The mutual shaping (or 'selection') of the cognitive structure and the institutional network is assumed to result in a self-organizational (or 'autopoietic') mode of development. The aim of L&H is methodological: is there empirical evidence to support this claim? The empirical data are found in the set of scientific publications in a set of core biotechnology journals from 1996.

The basis of the approach is a multivariate analysis of the intellectual organization of the documents in terms of title

words, and of the institutional organization in terms of the documents' regions of origin. L&H try to identify the EU, US, and Japanese title word sets, as indicators of 'geographic-cognitive spaces' of the research field. These regional sets are then correlated with the total word set which indicates the 'global intellectual space' of biotechnology in 1996. This global intellectual space may have different dimensions, and the correlations between the regional sets and the dimensions of the global intellectual space are interpreted as indicators of self-organization of the system of knowledge production. The analysis is comprised of several steps.

First, the biotechnology field is delineated on the basis of journal-journal citations. This leads to defining the field of biotechnology in terms of five core journals. Second, the 245 most frequently occurring title words in the documents (articles, notes, letters and reviews) in these journals are used as a representation of the intellectual organization of the biotechnology field in 1996. This is accomplished through a factor analysis of the co-word matrix, which results in a weak factor structure. Nevertheless, L&H consider this sufficient to analyze the relation to the regional dimensions of the institutional network.

In the third and crucial step, the documents are classified according to region of origin: USA, European Union, and Japan. The authors apply discriminant analysis (DA) to predict the region of origin (the dependent 'grouping' variable) of the 778 documents from the 245 most frequently used title words (the independent variables). As this prediction is correct in 78% of the cases, the authors claim that the title words of the correctly classified articles form 'regional word sets.'

In the fourth step the correlation is analyzed between the factors representing the global intellectual structure and the three regional word sets (US, EU, Japan). The resulting patterns are interpreted in terms of the theory of self-

---

Received February 4, 2002; received April 9, 2002; accepted June 3, 2002

© 2003 Wiley Periodicals, Inc.

organization (using Kaufman's NK-model) and the authors claim to have found some traces of 'Europeanization as a process of self-organization.' In the fifth step, the results are tested in a more detailed analysis at the level of individual European countries. This does not support the claim of an emerging European dimension. Finally, the analysis is repeated for the 1997 data, which does not corroborate the results for the 1996 data. L&H conclude that the relation between the intellectual and the institutional structure is very unstable and that 'second order theorizing' is needed to explain this. Here I will not try to reiterate the complete argument, but only focus on the crucial third step: the identification of the regional word sets.<sup>1</sup>

### Can Title Words Be Used to Map Research Fields?

Research indicates that to a certain extent words are useful as indicators of intellectual structure: for example, for the topics studied and for the methods used within a restricted specialty (e.g., Bhattacharya & Basu, 1998; Van den Besselaar, 2000). However, in broad research fields like biotechnology, codification on the level of title words is generally not very strong: words often have different meanings, even within a restricted document set (Leydesdorff, 1997). Thus, if intellectual structure is defined as textual codification, title words are not an obvious indicator for this. The intellectual organization of a field should be operationalized as a multidimensional indicator and includes many more types of relations among documents than only title words. If that is so, the main problem is how to combine the indicators in a useful way. Consequently, the strong results of L&H's analysis are unexpected: ordinarily, this very strong codified use of title words in a regional-cognitive space would not be an automatic assumption. However, if L&H's results are correct and title words are strongly codified according to region, this would be an important finding that requires additional examination.

L&H determine the specific regional (US, EU, Japan) word sets by Discriminant Analysis (DA) of the *article by title words* matrix, with  $CELL(ij) = 1$  if word  $j$  is contained in title  $i$ ,  $CELL(ij) = 0$  if word  $j$  is *not* in the title of document  $i$ . The grouping variable (region of origin of the paper) is added as the last column in the matrix. The DA results in a 78% correct classification of cases in terms of their regional origin; the correctly classified titles are considered to form the (partly overlapping) US, EU, and Japanese word sets.

Discriminant analysis is useful for building a predictive model of group membership based on characteristics ob-

served in the individual cases. The procedure generates a discriminant function (or, for more than two groups, a set of discriminant functions) based on linear combinations of the independent (predictor) variables that provide the best discrimination between the groups. The discriminant functions (model) are generated from a sample of cases for which group membership is known. These functions can then be applied to new cases with known values for the predictor variables but unknown group membership (the dependent variable).

DA is meant for causal analysis in which the independent variables are on an interval scale and the dependent variable is nominal. Predictor variables should have a multivariate normal distribution, and covariance matrices should be equal across groups (Klecka, 1980, p.61-62; Norusis, 1992, p.41). The data used by L&H do not meet the conditions for DA, as the independent variables in the L&H analysis are nominal: a word is used or not used in the title of a paper. Although DA is considered to be relatively robust with respect to violation of these conditions (Norusis, 1992, p.41), we will show that this is not the case in L&H's analysis.

### A General Phenomenon?

Do other research fields exhibit similar patterns? To answer this question, we applied the same methodology to information science in 1996, and science and technology studies in the period 1992–1997.<sup>2</sup> The method used by L&H for delineating biotechnology was applied to the other two research fields. The threshold for including title words was set to ensure the same the ratio of variables and cases as in L&H's study. Table 1 gives the results of the DA. For these two research fields we also found a high percentage of correctly classified documents. Consequently, L&H's results seem independent of the field of study. The question now is: is the use of title words truly region-specific, or is it an artifact of the method used? A standard way for testing the quality of the DA may answer this question.

### Testing the Discriminant Analysis

To test the quality of a model derived by DA, the cases should be divided in two groups. Applying the DA to one of the two subsets results in a model that can be used to classify cases in the other subset. Dividing the data into subsets provides a technique to perform a validation of the model generated (Klecka, 1980, p.51; Norusis, 1992, p.46). We performed various DAs on (randomly selected) sets of 50% of the cases of the L&H data, which resulted (as expected) in high rates of correctly classified cases. However, upon applying the derived discriminant functions to

---

<sup>1</sup>Actually, the discriminant analysis is the only analysis in the paper with strong outcomes, whereas the other empirical findings are rather weak: the factor analysis in step 2, the correlations in step 4, the analysis of the European dimension in step 5, and the comparison of the 1996 and the 1997 results in step 6.

---

<sup>2</sup>We use a longer period for science and technology studies, as the number of papers per year in the core journals in this field is relatively low.

TABLE 1. Number of correctly classified documents according to region of origin (EU, US, Japan).

	Number of documents (cases)	Number of words (variables)	Ratio of variables and cases	Percentage correctly classified documents
Biotechnology 1996*	778	245	0.31	75%
Information Science 1996**	305	117	0.38	87%
Science & Technology Studies 1992–1997**	640	250	0.39	84%

\* Leydesdorff, Heimeriks, 2001.

\*\* Own calculations.

the other half of the cases as validation, we found that the average correct classification was no better than the a priori probabilities (Table 2). This test was also applied on the information science data and the science and technology studies data, with identical outcomes. The implication of this test is that every sample results in a different model for the relationship between the independent and dependent variables, and consequently, none of the models is significant.<sup>3</sup> In other words, the hypothesis that word use is related to the region of origin must be rejected.<sup>4</sup>

If none of the derived models is correct, what is the meaning of the high percentage of correctly classified documents? To obtain clarification, we grouped the biotechnology documents randomly in three groups. Running the DA with the random grouping resulted in as ‘correct’ a classification of the titles for the three fictional ‘regions’ as did the classification for the real regions. Various other trials with random groupings had similar outcomes. Again we obtained identical results for information science and for science and technology studies.

This implies that every grouping can be predicted using the title words. For example, if one classifies the papers in

three categories according to length (long, medium, short), or first letter of author name (A–K; L–S; T–Z), Discriminant Analysis will show that title word use is ‘length specific’ or ‘author name specific.’

### Conclusion

Thus we see that the relationship between region of origin and use of title words is an artifact of the method used with this type of *nominal* data. Although DA does relate characteristics of the cases to grouping, if every case has *unique* characteristics, deriving *ex ante* a strong relationship between these characteristics and *any* classification becomes trivial. This is also the case here, as the combinations of title words in the titles are almost unique.

L&H’s theoretical argument is based on their finding that the use of title words is systematically related to the country of origin of documents: “The test of a European vocabulary in these title words, however was significant” (p.1272). As we see from the above analysis, however, this does not seem to be the case. So not only the “interpretations of the results remain speculative,” as L&H realize (p.1272), but the same holds for the empirical statements. Consequently, the argument that L&H found empirical traces of self-organizing development of mode-2 research fields such as biotechnology does not stand up to examination, which leaves their theoretical argument equally unfounded. In our view other empirical strategies are required to do this job.

<sup>3</sup>A reviewer asked about statistics related to the quality of the discriminant functions, such as the canonical correlation and Wilks’ Lambda. These statistics and their levels of significance are very similar for the ‘real’ cases and the random cases.

<sup>4</sup>An additional problem is that in the DA ‘region of origin’ is the dependent variable, but in the L&H design ‘region of origin’ is the independent variable and ‘title words’ are the dependent ones.

TABLE 2. Testing the discriminant analysis: Biotechnology 1996; two examples.

	Prior probability	Percentage of correctly classified titles			
		First run		Second run	
		Random 50% cases used for the DA	Using the DF for the other 50% of cases	Random 50% cases used for the DA	Using the DF for the other 50% of cases
EU	0.50	92	51	93	49
US	0.43	91	53	94	39
Japan	0.08	93	28	85	9

## Acknowledgments

This research was partly funded by the European Commission (the SOEIS project TSER-SOE1-CT97-1060 and the EIC-STES project IST-1999-20250). The paper was written during a sabbatical leave from the University of Amsterdam, Department of Psychology, Social Informatics Program.

## References

- Battacharya, S., & Basu, P.K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, 43, 359–372.
- Klecka, W.R. (1980). *Discriminant Analysis*. London, Sage.
- Leydesdorff, L. (1997). Why words and co-words cannot measure the development of the sciences. *Journal of the American Society of Information Science*, 48, 418–27.
- Leydesdorff, L. & Heimeriks, G. (2001). The self-organization of the European information society: the case of Biotechnology. *Journal of the American Society of Information Science and Technology*, 52, 1262–1274.
- Norusis, M.J. (1992). *SPSS-PC+, professional statistics*. Chicago, SPSS.
- Van den Besselaar, P. (2000). Communication between science and technology studies journals: A case study in differentiation and integration in scientific fields. *Scientometrics*, 47, 169–193.