Content-Based and Algorithmic Classifications of Journals: Perspectives on the Dynamics of Scientific Communication and Indexer Effects

Ismael Rafols

Science and Technology Policy Research (SPRU), University of Sussex, Brighton, BN1 9QE, England.
E-mail: i.rafols@sussex.ac.uk

Loet Leydesdorff

Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburg 48, 1012 CX Amsterdam, The Netherlands. E-mail: loet@leydesdorff.net

The aggregated journal-journal citation matrix—based on the Journal Citation Reports (JCR) of the Science Citation Index—can be decomposed by indexers or algorithmically. In this study, we test the results of two recently available algorithms for the decomposition of large matrices against two content-based classifications of journals: the ISI Subject Categories and the field/subfield classification of Glänzel and Schubert (2003). The content-based schemes allow for the attribution of more than a single category to a journal, whereas the algorithms maximize the ratio of within-category citations over between-category citations in the aggregated category-category citation matrix. By adding categories, indexers generate between-category citations, which may enrich the database, for example, in the case of inter-disciplinary developments. Algorithmic decompositions, on the other hand, are more heavily skewed towards a relatively small number of categories, while this is deliberately counteracted upon in the case of content-based classifications. Because of the indexer effects, science policy studies and the sociology of science should be careful when using content-based classifications, which are made for bibliographic disclosure, and not for the purpose of analyzing latent structures in scientific communications. Despite the large differences among them, the four classification schemes enable us to generate surprisingly similar maps of science at the global level. Erroneous classifications are cancelled as noise at the aggregate level, but may disturb the evaluation locally.

Introduction

The algorithmic decomposition of matrices of the size of the aggregated citation matrix among journals contained in the Journal Citation Reports (JCR; 7,611 journals in 2006) had hitherto remained a problem because of computational limitations, for example, in SPSS (Leydesdorff, 2006). In the summer of 2008, however, Vincent Blondel and his colleagues published a new algorithm for the decomposition of large networks at *arXiv* (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). This algorithm was quickly adopted in a follow-up study by Wallace, Gingras, and Duhon (2008) for the study of research communities in terms of citation networks. Others followed, as the algorithm was made publicly available at http://findcommunities.googlepages.com (e.g., Lambiotte & Panzarasa, in press). Another algorithm for the same purpose was recently made available by Martin Rosvall at http://www.tp.umu.se/~rosvall/code.html. Rosvall developed this program in the framework of a study (with Carol Bergstrom) of the aggregated journal-journal citation matrix using 6,128 journals included in the JCR of 2004 (Rosvall & Bergstrom, 2008).

At that time, we (Leydesdorff & Rafols, 2009) were deeply involved in testing the ISI Subject Categories of these same journals in terms of their disciplinary organization. Using the JCR of the Science Citation Index (SCI), we found 14 major components using 172 subject categories, and 6,164 journals in 2006. Given our analytical objectives and the well-known differences in citation behaviour within the social sciences (Bensman, 2008), we decided to set aside the study of the $(220 - 175 =)$ 45 subject categories in the social sciences for a future study.[1] Our findings using the SCI indicated that the ISI Subject Categories can be used for statistical mapping purposes at the global level despite being imprecise in terms of the detailed attribution of journals to the categories.

[1]Three (of the 175) subject categories in the SCI are no longer actively used for the indexing.

Might it be that the two newly available algorithms would provide us with a more reliable classification of the journals? Rosvall and Bergstrom (2008), for example, claimed to find 88 components using the same data for 2004. What effect would Blondel et al.'s (2008) new unfolding algorithm have on the JCR data? In this study, we compare the results of these two algorithms with the full set of 220 Subject Categories of the ISI. In addition to these three decompositions, a fourth classification system of journals was proposed by Glänzel and Schubert (2003) and increasingly used for evaluation purposes by the Steungroep Onderwijs and Onderzoek Indicatoren (SOOI) in Leuven, Belgium. These authors originally proposed 12 fields and 60 subfields for the SCI, and three fields and seven subfields for the Social Science Citation Index and the Arts and Humanities Citation Index. Later, one more subfield entitled "multidisciplinary sciences" was added.

Wolfgang Glänzel very kindly provided us with a copy of this classification of journals for the purpose of this research (Glänzel, personal communication, October 11, 2008). Based on our replication of Rosvall and Bergstrom's analysis using 2006 data and on our collaboration with Renaud Lambiotte—one of the authors of Blondel et al. (2008)—, we are able to compare below the four classification schemes using the same dataset of the combined JCRs for the Social Science Citation Index and the *SCI* 2006.

The Importance of Journal Classifications

Garfield's (1971) "law of concentration" can be considered as a corollary of Bradford's (1934) "law of scattering." Whereas Bradford's law formulated that "articles of interest to a specialist must occur not only in the periodicals specialising in his subject" (cited in Garfield, 1971, p. 1), Garfield stated that the tails of journals' citation distributions can be expected to encompass the cores of other sets (Garfield, 1972). Thus, because research topics are, on the one hand, thinly spread outside the core group and, on the other hand, the core groups are interwoven, one cannot expect that the aggregated journal-journal citation matrix matches one-to-one with substantive definitions of categories or that it can be decomposed in a single and unique way in relation to scientific specialties. The choice of an appropriate journal set can be considered as a local optimization problem (Leydesdorff, 2006).

Citation relations among journals are dense in discipline-specific clusters and are otherwise very sparse, to the extent of being virtually non-existent (Leydesdorff & Cozzens, 2003).[2] Although not completely decomposable, the matrix can be considered as heavily structured and nearly decomposable. The grand matrix of aggregated journal-journal citations is so heavily structured that the mappings and analyses in terms of

citation distributions have been amazingly robust despite differences in methodologies (e.g., Leydesdorff, 1987 and 2007; Tijssen, de Leeuw, & van Raan, 1987; Boyack, Klavans, & Börner, 2005; Moya-Anegón et al., 2007; Klavans & Boyack, 2009).

A decomposable matrix is a square matrix such that a rearrangement of rows and columns leaves a set of square sub-matrices on the principal diagonal and zeros everywhere else. In the case of a nearly decomposable matrix, some zeros are replaced by relatively small nonzero numbers (Simon & Ando, 1961; Ando & Fisher, 1963). Near-decomposability is a general property of complex and evolving systems (Simon, 1973 and 2002). The next-order units represented by the square sub-matrices—and representing in this case disciplines or specialties—are reproduced in relatively stable sets (of journals) that may change over time. The sets of journals are functional subsystems that show a high density in terms of relations within the center (that is, core journals), but are more open to change in relations at the margins. The organization among the subsystems can also change.

The decomposition into nearly decomposable matrices has no analytical solution. However, algorithms can provide heuristic decompositions when there is no single unique correct answer. Newman (2006a, 2006b) proposed using *modularity* for the decomposition of nearly decomposable matrices since modularity can be maximized as an objective function. Blondel et al. (2008) used this function for relocating units iteratively in neighbouring clusters. Each decomposition can then be considered in terms of whether it increases the modularity. Analogously, Rosvall, and Bergstrom (2008) maximized the probabilistic entropy between clusters by estimating the fraction of time during which every node is visited in a random walk (cf. Theil, 1972; Leydesdorff, 1991).

Methods and Data

The data were harvested from the CD-Rom version of the JCR of the SCI and Social Science Citation Index 2006, and then combined. In 2006, the JCR for the SCI contained 6,164 journals, and the Social Science Citation Index 1,768 journals. We corrected for the overlap of 321 journals that are contained in both databases. The resulting set of 7,611 journals and their citation relations otherwise precisely corresponds to the online version of the JCRs. This large data matrix of 7,611 times 7,611 citing and cited journals was stored conveniently as a Pajek (.net) file and used for further processing.

The 7,611 journals are attributed by the ISI with 11,856 subject classifiers. This is 1.56 (± 0.76) classifiers per journal. The ISI staff assign the 220 ISI Subject Categories on the basis of a number of criteria including the journal's title and its citation patterns (McVeigh, personal communication, March 9, 2006; Bensman & Leydesdorff, 2009). Pudovkin and Garfield (2002) state that the ISI assigns

---

[2] In 2006, the database contained only 1,315,143 of the 57,927,321 ($= 7611^2$) possible relations between two journals, or in other words a density of 2.27%.

journals to categories by "subjective, heuristic methods" (p. 1113):

> ...This method is "heuristic" in that the categories have been developed by manual methods started over 40 years ago. Once the categories were established, new journals were assigned one at a time. Each decision was based upon a visual examination of all relevant citation data. As categories grew, subdivisions were established. Among other tools used to make individual journal assignments, the Hayne-Coulson algorithm is used. The algorithm has never been published. It treats any designated group of journals as one macrojournal and produces a combined printout of cited and citing journal data. (Pudovkin and Garfield, 2002, at p. 1113n.)[3]

According to the evaluation of Pudovkin and Garfield (2002), in many fields these categories are sufficient, but the authors added that "in many areas of research these 'classifications' are crude and do not permit the user to quickly learn which journals are most closely related" (p. 1113). Boyack et al. (2005) estimated that the attribution is correct in approximately 50% of cases across the file (Boyack, personal communication, September 14, 2008). Leydesdorff and Rafols (2009) found that the ISI Subject Categories can be used for statistical purposes—the factor analysis for example can remove the noise—but not for the detailed evaluation. In the case of interdisciplinary fields, problems of imprecise or potentially erroneous classifications can be expected.

For the purpose of developing a new classification scheme of scientific journals contained in the SCIs, Glänzel and Schubert (2003) used three successive steps for their attribution. The authors iteratively distinguished sets *cognitively* on the basis of expert judgements, *pragmatically* to retain multiple assignments within reasonable limits, and *scientometrically* using unambiguous core journals for the classification. The scheme of 15 fields and 68 subfields is used extensively for research evaluations by the Steunpunt Onderwijs and Onderzoek Indicatoren (SOOI), a research unit at the Catholic University in Leuven, Belgium, headed by Glänzel.

The SOOI categories cover 8,985 journals. Using the full titles of the journals, 7,485 could be matched with the 7,611 journals under study in the JCR data for 2006 (which is 98.3%). These journals are attributed 10,840 classifiers at the subfield level. This is 1.45 ($\pm$0.66) categories per journal. One category ("Philosophy and Religion") is missing because the Arts & Humanities Citation Index is not included in our data. Thus, we pursued the analysis with the 67 SOOI categories.

Like the ISI Subject Categories, the SOOI categories are kept the same over time to make longitudinal comparisons possible. In other words, one can also consider the classification by Glänzel and Schubert (2003) as an update of the old (1974) ISI classification. Neither classification, however, can be reproduced by an outsider. Furthermore, they are not based on "literary warrant" like the classification of the Library of Congress (Chan, 1999). The LC has a policy of continuous revision to take current literary warrant into account, so that new areas are developed and obsolete elements are removed or revised (Leydesdorff & Bensman, 2006, p. 1473). The ISI or SOOI categories, however, are changed in terms of respective coverage but cannot themselves be revised from the perspective of hindsight (Leydesdorff, 2002).

Unlike the content-based classifications, the two algorithms under study provide unique assignments of journals to groups. Using Rosvall and Bergstrom's (2008) algorithm with 2006 data, we obtained findings similar to those of these authors on August 11, 2008. Like the original authors using 6,128 journals in 2004, we found 88 clusters using 7,611 journals in 2006. We used this attribution for the purposes of further analysis.[4] Lambiotte, one of the coauthors of Blondel et al. (2008), was so kind as to input the data into the unfolding algorithm and found the following results: 114 communities with a modularity value of 0.527708 and 14 communities with a modularity value of 0.60345. We use the 114 communities for the purposes of this comparison. These categories refer to 7,607 ($= 7611 - 4$) journals because four of the journals in the file were isolates.[5]

Results

To understand the differences and similarities among the four classifications investigated, we proceed with the following analyses. First, the statistical properties of each of the decompositions are investigated. Second, we look at the degree of agreement between the decompositions. Third, the visual maps of science obtained using the diverse classifications are compared. In the conclusions and discussion sections, we discuss the main differences between the four classification schemes and the implications for information retrieval and the sociology of science.

Statistical Properties of the Decompositions

Let us begin by providing the descriptive statistics of the different classifications used in this study. The number of categories in each decomposition is of the same order of magnitude of hundred—with a range from 67 (SOOI) to 220 (ISI) as the extremes—and, therefore, one can expect that the results will be roughly comparable. The number of journals per category is log-normally distributed in each of the four classifications. In other words, they all have a relatively small number of categories with a large number of journals and many categories with only a few journals. However, as shown in Figure 1, the classifications based on the random walk and unfolding algorithms are more skewed than the content-based classifications.

---

[3]Pudovkin and Fuseler (1995, p. 228) further specified the Hayne-Coulson algorithm as follows: "The number of citations each journal receives from different specialty core journals is obtained annually by a computer routine (Hayne-Coulson) that is used to create the JCR database."

[4]In response to our exchange, Rosvall uploaded an improved algorithm on August 13, 2008, which provides different numbers of clusters depending on a random seed number. This algorithm is no longer deterministic.

[5]These four journals are *Mer-Marine Engineers Review*, *Man in India*, *Nursing History Review*, and *Russian Politics and Law*.
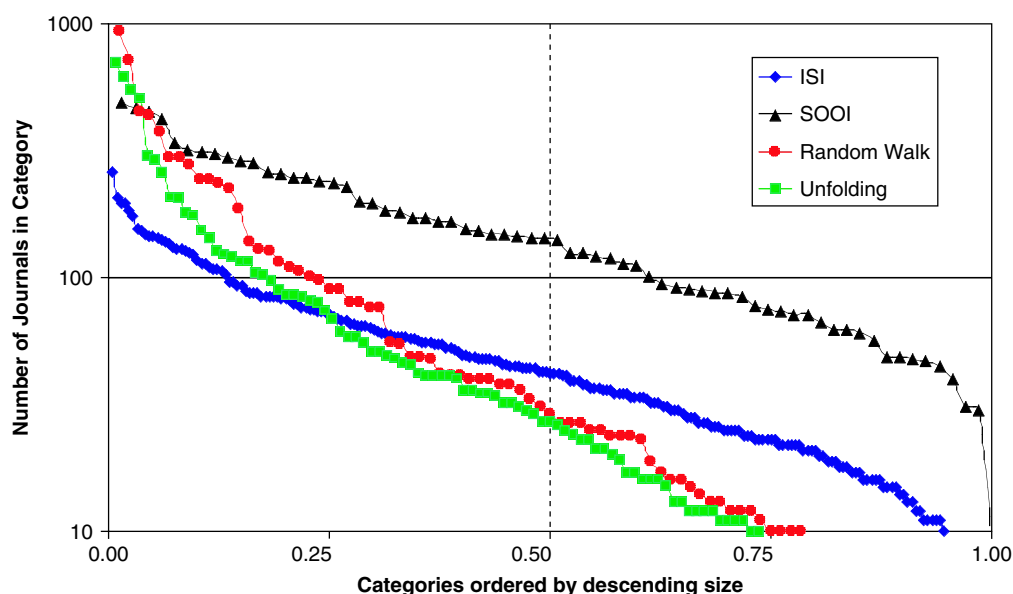
FIG. 1. Number of journals in each category ranked according to size. The horizontal axis is normalized so that the category with the highest number of journals appears in 0 and the one with the lowest value in 1.

TABLE 1. Statistical properties of the number of journals per category and the within/between category citations for each decomposition.

7,611 journals	ISI	SOOI	Random walk	Unfolding
Number of categories	220	67	88	114
Journals/categories				
Mean	53	161	86	66
Median	42	141	28	26
Standard deviation	43	113	151	118
Citations within/between categories				
Within category citations	13,286,544	14,194,033	15,458,430	13,630,264
Between category citations	41,163,842	33,013,188	7,412,325	9,240,491
Grand-sum of citation matrix	54,450,386	47,207,221	22,870,755	22,870,755
Within/between categories ratio	3.10	2.33	0.48	0.68

Whereas the top-10 categories on the basis of a random walk comprise 57% of the journals (50% for unfolding), they cover only 15% in the ISI decomposition and 31% for the SOOI classification. In the case of skewed distributions, the characteristic number of journals per category can best be expressed by the median: the median is below 30 in the random walk or unfolding classifications, compared with 42 journals for the ISI classification and 141 for the SOOI classification (Table 1).

Another way to analyze the overlaps among categories for a given decomposition is to compare the number of aggregated citations between categories (off-diagonal) with the citations within categories (on the diagonal). As presented in the last rows of Table 1, the total numbers of citations in the aggregated matrices based on the ISI or SOOI classifications are much higher because the same citation can be attributed to two or three categories. Thus, whereas random walk and unfolding lead to matrices with most citations within categories (on the diagonal), matrices based on ISI and SOOI classifications lead to matrices with most citations between categories (off-diagonal).

Finally, to measure how similar the categories in the four decompositions are to each other, we computed the cosine similarities in the citation patterns between each pair of citing categories in the four aggregated category-category matrices (Salton & McGill, 1983; Ahlgren, Jarneving, & Rousseau, 2003). These cosine similarities provide an indicator between zero and one of the degree to which two categories in a decomposition cover the same set of journals. For example, in the ISI decomposition, the new Nanoscience and Nanotechnology categories have a similarity of cosine = 0.003 with Women's Studies (hence, they are completely distinct), but a cosine = 0.50 similarity with Electrochemistry (suggesting important overlap), and a cosine = 0.96 with Materials
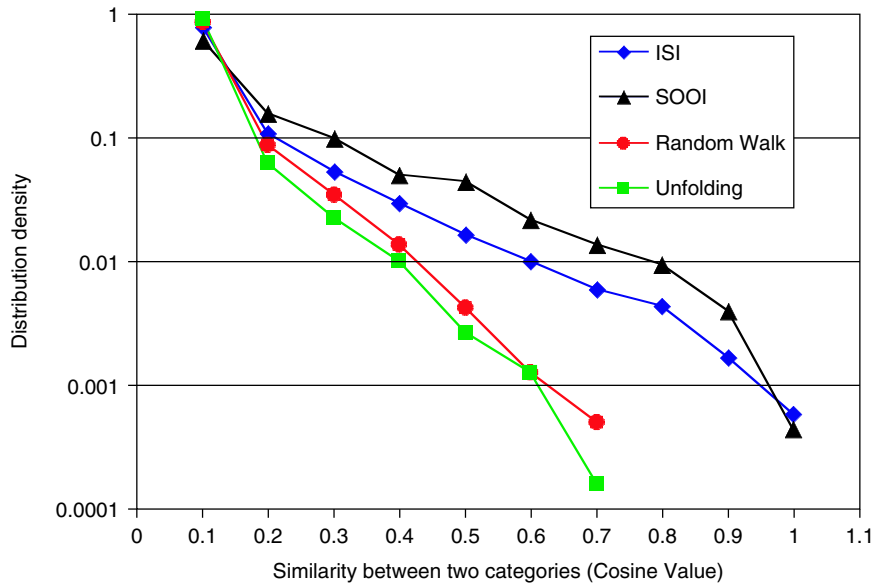
FIG. 2. Distribution density of cosine similarities between categories. ISI and SOOI's similarity values are much higher than those of a random walk or the unfolding algorithm.

TABLE 2. Statistical properties of the cosine similarities between categories and the within/between category citations for each decomposition.

Cosine similarity between categories	ISI	SOOI	Random walk	Unfolding
Number of categories	220	67	88	114
Mean	0.076	0.136	0.044	0.033
Median	0.020	0.066	0.009	0.007
Standard deviation	0.129	0.165	0.077	0.065

Science, Interdisciplinary. This high value raises the question of whether these two categories, as defined by the ISI, should be distinguished from each other.

The statistical analyses of the cosine similarities are presented in Table 2. Figure 2 shows the density distribution of the similarities between categories, i.e., the proportion of similarities that fall within a certain value bin (e.g., in the case of unfolding, the proportion of similarities between 0.3 and 0.4 is 0.01, or 1% of the similarities). We find again that all the distributions are highly skewed and that the random walk and unfolding algorithms exhibit a much lower median similarity value among categories. The lower medians indicate that the algorithmic decompositions produce a much "cleaner" cut between categories than the content-based classifications. As shown in Figure 2, the proportion of similarity values above 0.5 is below 0.01 (or 1%) in the algorithmic decompositions (0.6% in random walk and 0.4% in unfolding), whereas it is 10 times higher using the content-based classifications (3.9% for ISI and 9.3% in SOOI).

In conclusion, the analysis of the statistical properties of the different classifications teaches us that the random walk and the unfolding algorithms produce much more skewed distributions in terms of the number of journals per category, but these constructs are more specific than the content-based

classification of the ISI and SOOI. The content-based sets are less divided because the boundaries among them are blurred by the multiple assignments.

Degree of Agreement Between Different Decompositions

Let us now turn to the relations among the four decompositions. First, we tested whether the distributions are independent. To this end, the four classifications were cross-tabled, and the cross-tables tested for the zero-hypothesis of mutual independence by using chi-square (with Yule-correction because of the low expected values in many cells). This zero-hypothesis was rejected in all comparisons at the level of $p < 0.001$. We were not surprised that the four classification schemes are correlated.

Second, we focused on the diagonal values in the cross-tables. The Appendix provides an example of how the cross-tables are analyzed. If two classification schemes have a perfect match between them—which cannot be expected because of the different numbers of categories—, then the results of cross-tabling would yield a matrix with values that can be sorted on the main diagonal; all off-diagonal values would then be zero. Therefore, the proportion of journals on the diagonal of the cross-tables provides us with a measure

TABLE 3. Percentage of journals in one decomposition belonging to the matching category in another decomposition.

% journals given this decomposition...	... in top matching category of this decomposition			
	ISI	SOOI	Random walk	Unfolding
ISI	–	98.2	66.2	60.1
SOOI	58.4	–	51.5	47.0
Random walk	43.7	56.0	–	**71.3**
Unfolding	48.4	59.3	**85.7**	–

of the accuracy of the fit between matching categories from different decompositions.

Table 3 shows the level of matching. For example, given the categorization by the random walk algorithm (third row), only 43.7% of the journals are located in the main matching category of the ISI classification (first column). Because journals in the classification schemes of the ISI and SOOI are assigned to more than a single category, these percentages provide an over-estimation of the agreement because a larger chance of a match is generated by the multiple assignments. Furthermore, one can expect the agreement to be higher when categories from a finer decomposition are used as a predictor for a coarser one.

For example, when the 220 ISI Subject Categories are used as predictors for the 67 SOOI subfields, the ISI-to-SOOI matching is 98.2%, whereas the SOOI-to-ISI matching is only 58.0% despite the favorable effect of multiple assignations in both schemes. We were not surprised that the agreement among diagonal elements is highest (71.3% and 85.7%; bold-faced in Table 3) when comparing the results of the random walk and the unfolding algorithm because both of these are based on the maximization of an objective function.

Table 3 indicates that the concordance between the different classifications is modest: In general, in the 40%–60% range, which is in agreement with Boyack's estimation of 50% correct classifications for the ISI categories (Boyack, personal communication, September 14, 2008). However, when we expand the diagonal from the single most matching category to the top-three matching categories, we obtain agreements in the order of 70%–85%, that is, a 15%–25% increase in comparison to the figures in Table 3. For example, the ISI-to-random walk agreement is, in this case, 80.5%, ISI-to-unfolding is 84.3%, SOOI-to-random walk is 78.4%, and SOOI-to-unfolding is 74.0%.

In summary, although the correspondences among the main categories are sometimes as low as 50% of the journals, most of the mismatched journals appear to fall in areas within the close vicinity of the main categories. In other words, it seems that the various decompositions are roughly consistent but imprecise. We believe that this broad-range agreement, in spite of local inaccuracies, explains the unexpected similarities among the science maps that we present below.

Comparison Between Science Maps

Maps of science for each decomposition were generated from the aggregated category-category citation matrices using the cosine as similarity measure. The similarity matrices were visualized with Pajek (Batagelj & Mrvar, 1998) using Kamada and Kawai's (1989) algorithm. The method is kept similar to that used in Leydesdorff and Rafols (2009). The threshold value of similarity for edge visualization is pragmatically set at cosine > 0.01 for the algorithmic decompositions and cosine > 0.2 for the content-based decompositions to enhance the readability of the maps without affecting the representation of the structures in the data.

Because maps with many labels are difficult to read, the following simplified representations were prepared. For the ISI decomposition, the 220 categories (Figure 3) were clustered into 18 macro-categories (Figure 4) obtained from the factor analysis (cf. Leydesdorff and Rafols, 2009). The map of the SOOI classification was constructed with all is 67 subfields (Figure 5). Taking advantage of the concentration of journals in a few categories, in the case of random walk and unfolding only the top 30 and 35 categories were used, respectively. These top-segments comprise 85.9% and 81.4% of all journals, respectively.[6] The labels for the random walk and unfolding classifications were assigned on the basis of on an inspection of the journal titles in the category but as far as possible in accordance with the labels of the ISI Subject Categories.[7] Because both maps are extremely similar, only the unfolding is shown (Figure 6).

Each map shows the categories of each decomposition, their relative position according to their main inter-relations, and the size of each category (the area of a node/discipline is proportional to its number of outbound citations, that is, "total citing"). The overall structure of science obtained is astonishingly robust, given the major differences among the four decompositions shown in Table 3. These results do not solve the problems of decomposition, but they support the observation that 50% correspondence in the classification is enough for obtaining similar maps (Leydesdorff and Rafols, 2009).

Because the similarities used to make these maps depend on cross-category citations, a structural agreement in the

---

[6]The choice of 30 and 35 categories is arbitrary. By using this segment, we capture most of the journals (>80%), and yet the results remain readable.

[7]In about 10% of the cases, the ISI attributions suggested minor corrections to the labels. For example, we originally categorized the field of "clinical neurology" in the random walk classification as "neurosciences" after inspection of the titles of the journals involved, but then we relabeled it with the ISI Subject Category "clinical Neurology" after finding a 59% correspondence.
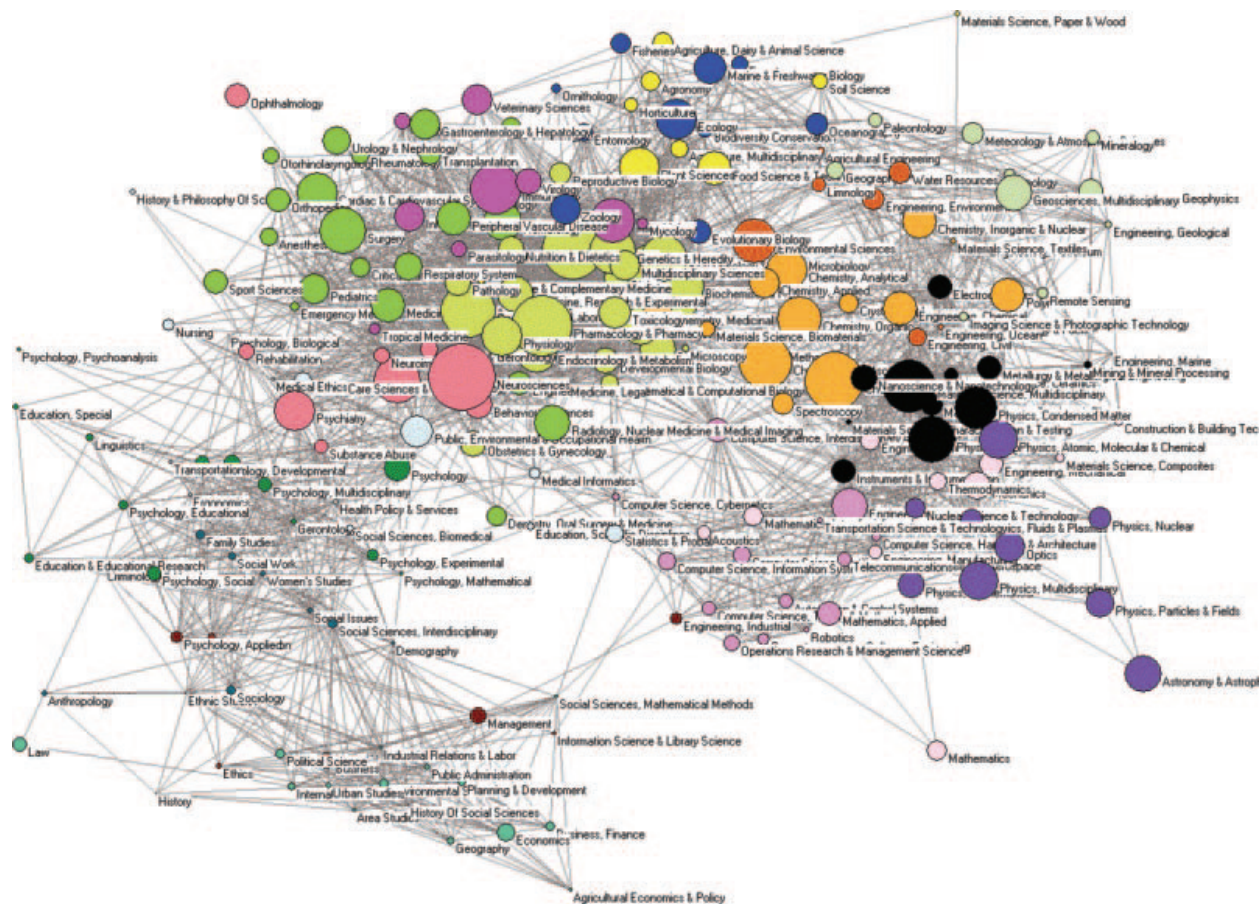
FIG. 3.   Similarity map among the 220 subject categories of the ISI; cosine > 0.2.

map would indicate concordance in the underlying relations among categories. In other words, the resemblance among these maps provides us with an assessment of the agreement in the off-diagonal elements in the aggregated citation matrices among categories. Indeed, the four maps correspond in displaying two main poles: a very large pole in the biomedical sciences and a second pole in the physical sciences and engineering. These two poles are connected via three bridging areas: chemistry, a geosciences-environment-ecology group, and the computer sciences. The social sciences are somewhat detached, linked via the behavioral sciences/neuroscience to the biomedical pole, and via the computer sciences and mathematics to the physics/engineering pole.[8] Although the objective of this study was to investigate the differences among decomposition, our findings indicate an unexpected robustness and convergence ("consensus") among the resulting maps (Moya-Anegón et al., 2007; Klavans & Boyack, 2009).

Another way to look at these maps is to use a circular ordering of disciplines as a first approximation. This perspective was recently proposed as overarching by Klavans and Boyack (2009). The circle of disciplines obtained in all four representations is consistent with their circular Consensus Map. Moving in the clockwise direction in Figures 3 to 8, one can identify the following categories: biomedical sciences, biology and environment, geosciences and chemistry, physics and engineering, mathematics and computer sciences, social sciences, psychology, neurosciences and health services, medical specialties and infectious diseases, to biomedical sciences again. Note that the use of the word "and" above does not designate a joint category but two distinct categories that occupy parallel positions in the clockwise reading of the maps.

Why is the overall structure of these science maps so robust despite the large differences in the decompositions? As noted above, although categories of different decompositions do not always match with one another, most "misplaced" journals are assigned into closely neighbouring categories. Therefore, the error in terms of categories is not large and is also unsystematic. The noise-to-signal ratio becomes much smaller when aggregated over the relations among categories. In summary, our results suggest that the robustness of the maps is due to a "pixelling effect": One can decompose a picture into different rather coarse subsets and yet keep an understanding of the original image if it is watched (i.e., "integrated") at a distance. The overall quality of a map at a high level of aggregation does not inform us about the quality of

---

[8]Because we use the JRC data, our data do not contain journals from the Arts and Humanities Citation Index.
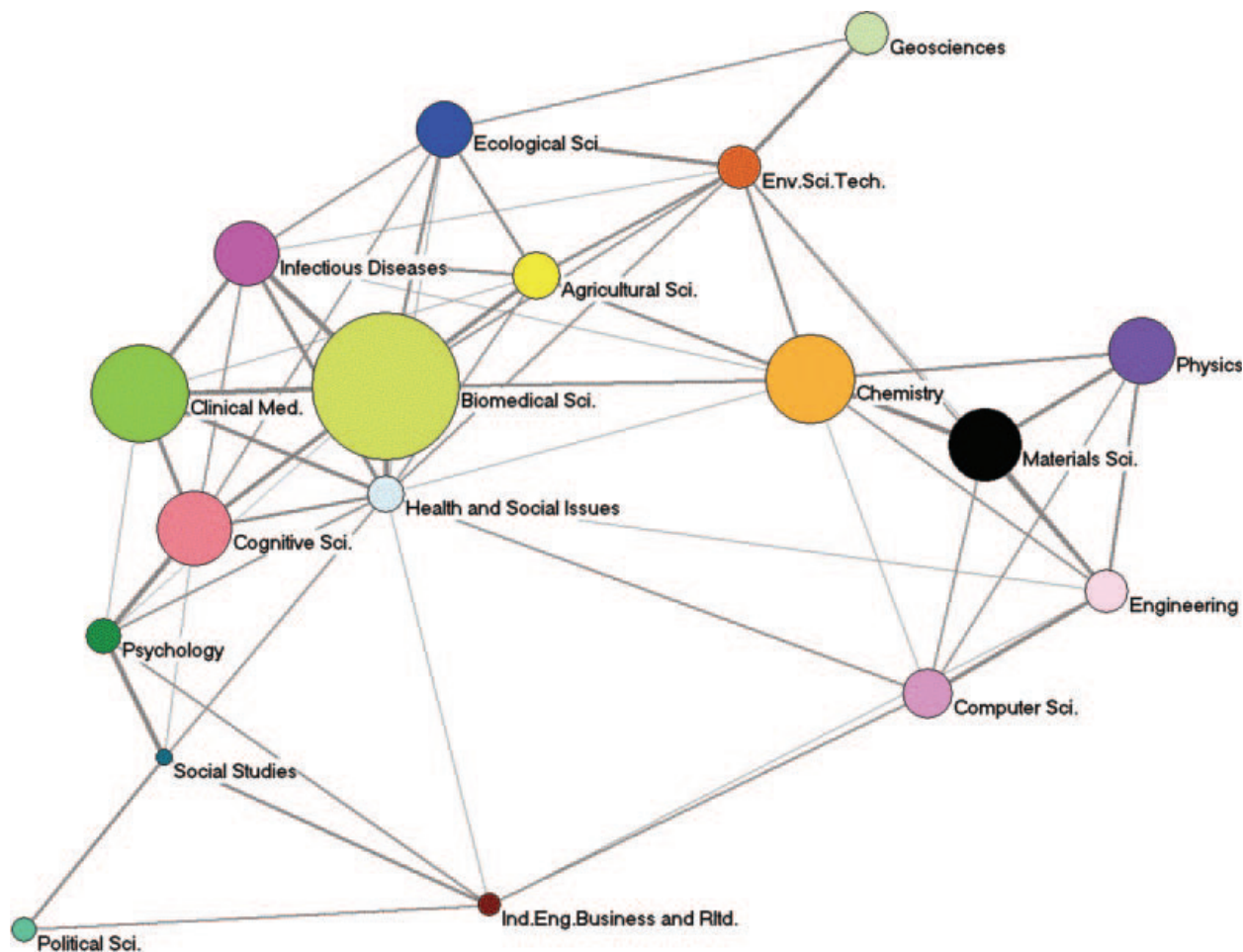
FIG. 4. Similarity map among the 18 macro-categories on the basis of factor analysis of the 220 ISI subject categories; cosine > 0.1.

the underlying micro-relations other than statistically. The "consensus" among these maps can thus be considered as an effect of the aggregation.

As a second important observation that can be made on the basis of these maps, we wish to point to the differences in category density between the content-based and the algorithm-based maps. In the case of the ISI and SOOI classifications (Figures 3 and 5, respectively), the three areas around biomedical sciences, chemistry, and physics/materials sciences contain a very high density of categories. The cosine similarity coefficients among these categories explain why: In the biomedical-science areas, the journals classified by the ISI as biochemistry and molecular biology have very high-similarity values with those in the biophysics (0.97), cell biology (0.95), and genetics and heredity (0.81) categories. In the SOOI classification, the biochemistry-biophysics-molecular biology category has a cosine of 0.93 with cell biology and 0.74 with genetics and developmental biology. However, in the random walk and unfolding classifications, all these categories are brought together under a single large biomedical sciences category. The highest cosine similarity values within these two decompositions are only 0.67 and 0.58, respectively.

This similarity among content-based categories is particularly relevant for the assessment of "interdisciplinarity" or more generally "diversity" (Morillo, Bordons, & Gómez, 2003; Stirling, 2007; Porter, Cohen, Roessner, & Perreault, 2007; Rafols & Meyer, 2009). Using the ISI or SOOI classifications, a collection of papers published in the biosciences could be counted as interdisciplinary because the set is retrievable under a number of categories in the biomedical sciences (e.g., cell biology, biophysics, and developmental biology).[9] However, if one takes into account that these categories have large overlaps among them, one reaches a different conclusion: This same document set can rather be considered as monodisciplinary within a single large disciplinary category as obtained by using classifications based on a random walk or the unfolding algorithm.

The example discussed above suggests that, to mitigate the biases introduced by classifications on analysis, classifications should be interpreted together with their similarity

---

[9]Most common measures of interdisciplinarity are based on numbers or the balancing of categories (e.g., Van Raan & Van Leewen, 2002) or indicators that subsume both number and balance (such as the Shannon entropy, e.g., Grupp, 1990; Rafols & Meyer, 2009).
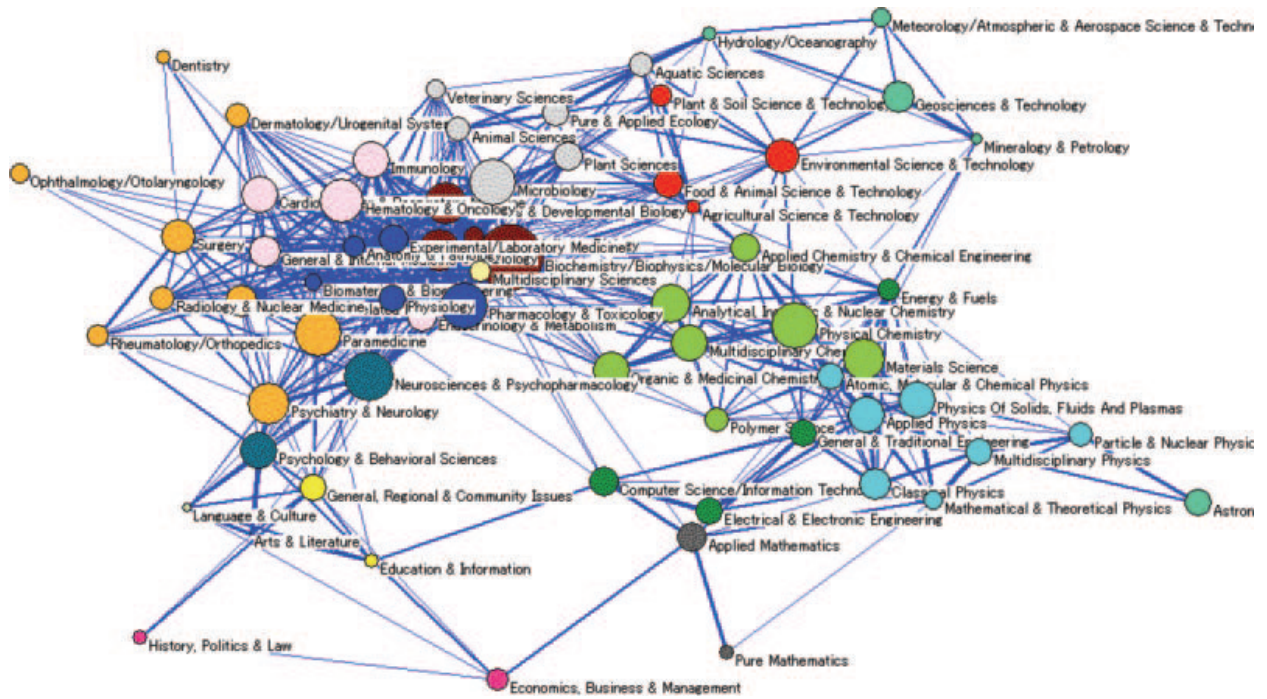
FIG. 5. Similarity map among the 67 categories of the SOOI; cosine > 0.2.
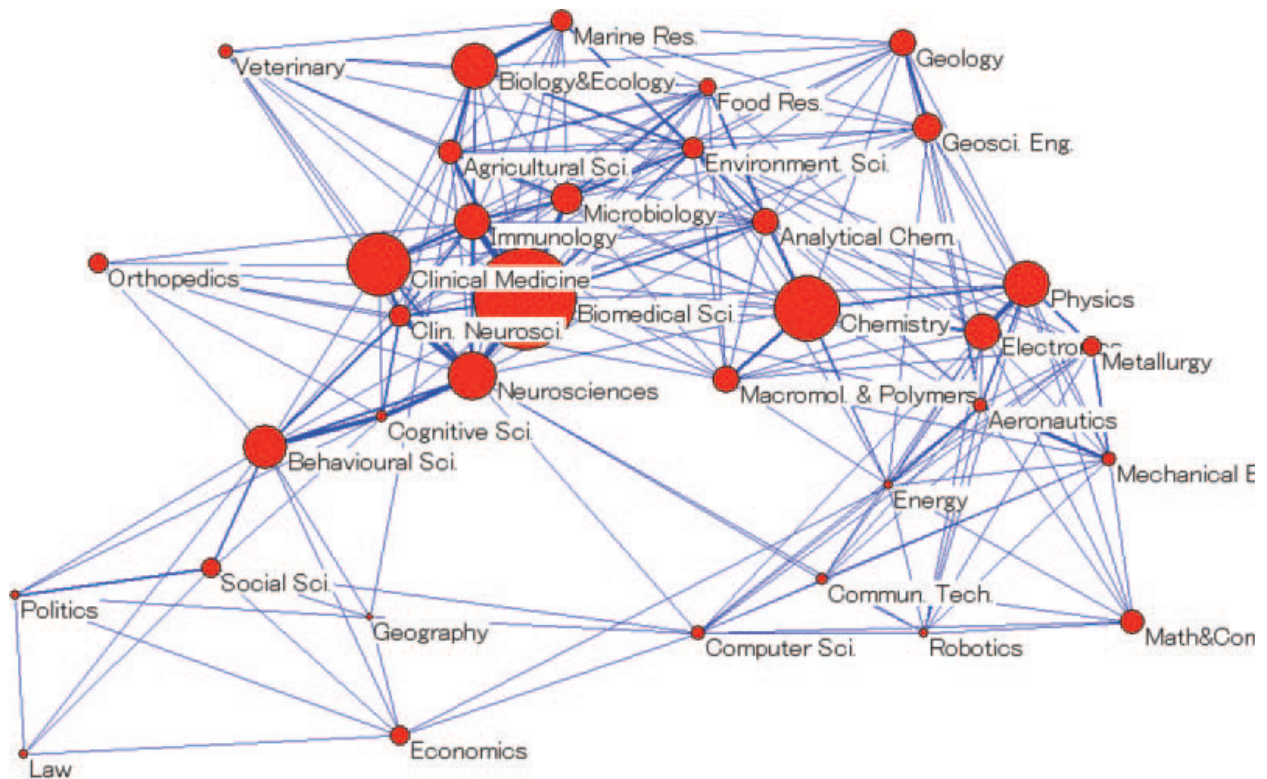

FIG. 6. Similarity map among the first 35 categories of the decomposition based on the unfolding algorithm; cosine > 0.02.

matrix and size distribution (as discussed in Tables 1 and 2). Although classifications can differ substantially, the similarity matrix and size distribution allow the user to contextualize the extent to which categories have overlaps, fuzzy areas or are very distinct. For example, this information can then be used to assess interdisciplinarity taking into account the degree of overlap between categories (Porter et al., 2007; Rafols & Meyer, 2009).

In summary, we were surprised to find that the different science maps are similar except that they differ in the density of categories within groups. The content-based classifications achieve a more balanced coverage of the disciplines at the expense of distinguishing categories that may be highly similar in terms of journals. These categories could have been collapsed without losing analytical quality. We appreciate that the ISI Subject Categories were primarily designed for information retrieval and not for scientometric purposes. However, the SOOI categories, which were developed for the purpose of scientometric evaluations, are even more densely packed than the ISI Subject Categories.

Summary of Findings

In this article, we have analyzed the results of four different decompositions of the aggregated journal-journal citation matrix. The first finding is that the algorithmic decompositions have very skewed and clean-cut distributions, with large clusters in a few scientific areas, whereas indexers maintain more even and overlapping distributions in the content-based classifications. Second, the different classifications show a limited degree of agreement in terms of matching categories.[10] In spite of this lack of agreement, however, the science maps obtained are surprisingly similar; this robustness is due to the fact that although categories do not match precisely, their relative positions in the network among the other categories is based on distributions that match sufficiently to produce corresponding maps at the aggregated level.

Regarding the first finding, we suggest that because of the specific shape of the citation distributions among journals—with dense citation traffic within clusters and weak citation traffic among them—the algorithmic decompositions can be highly successful in decomposing the ISI journal set using maximization of an objective function. One would expect the results of the two algorithmic decompositions to be very similar because of the common ideas behind them (Newman, 2006a and b). Because of the log-normal shape of the resulting distribution, the number of clusters distinguished by each algorithm can be expected to determine the skewness (Figure 1). In other words, the results can be compared with hierarchical clustering when one changes the cut-off level: the lower the level, the more groups are distinguished, and larger groups can be expected to split up first when lowering the threshold level.

On the other hand, we found that the differences between the two content-based classifications were marginal in statistical terms. The largest categories in the ISI classification subsume 200+ journals, while the largest categories in the SOOI classification subsume 400+ journals. These differences are again an effect of the granularity of the classification: the ISI distinguishes 220 categories and the SOOI 67 subfields.

Discussion

Two strategies are available to the content-based indexer that cannot so easily be left to the machine algorithm. First, the indexer can be expected to be aware that categories should not be unbalanced in terms of size. Thus, there will be a tendency to split up large categories if possible, and to merge extremely small ones. This human intervention can be considered as a legitimate indexer effect. After all, indexing aims at facilitating the user to retrieve relevant information, and neither extremely large nor extremely small categories are functional to this end.

Second, the indexer can add cross-connections between categories, which cannot be legitimated in terms of the nearly decomposable citation matrix. Both the ISI and the SOOI classifications make use of this option and, thus, generate off-diagonal elements. One should be aware that these off-diagonal elements cannot be legitimated by the citation patterns generated by the scientists themselves but are added by indexers.

Of course, indexers may consult individual experts— Glänzel and Schubert (2003) iteratively involve experts in each round of generating the classifications—, but the communities of experts themselves neither produce nor reproduce these externally identified cross-connections. The cross-connections may provide a help function to the user of the database by assuming an external user's perspective to the underlying structures. Because these content-based classifications are generated by indexers, they no longer provide a systematic representation of the dynamics of scientific communication itself.

For the purposes of more finer-grained evaluations, the indexer effects can be expected to disturb maps at lower levels of aggregation and lead to deceptive interpretations, for example, of interdisciplinarity. At the aggregated level, however, the content-based categorizations can be used for the mapping because these maps use information regarding the relative proximity of categories. This suggests that the providers of these categorizations could add information regarding the size distribution of the categories and the similarities among them (underlying the mapping) to control for the indexer effects (Porter et al., 2007; Rafols & Meyer, 2009).

In our opinion, the sociology of science—and along with it the policy analysis of science—should remain as close as possible to the representations of science that are reproduced in terms of aggregated citations by the scientists themselves, rather than those that are based on a reflexive representation of these data by indexers (or bibliometricians), who operate

---

[10]These results are in accordance with previous findings comparing other classifications, as illustrated in Figure 12 in Leydesdorff (2006, p. 612).

from different premises. However puzzling the results of the algorithmically generated classifications may be, these structures provide us with the *explanandum* of a theory of citations or a sociology of science. The reduction of uncertainty in hierarchical classifications by indexers cannot be taken for granted by the emerging "science of science policy" (Marburger, 2005; National Science and Technology Council / Office of Science and Technology Policy, 2008). In other words, we do not wish to deny the pragmatic usefulness of the content-based categories, but their validity.

Acknowledgments

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.

Ando, A., & Fisher, F.M. (1963). Near-decomposability, partition and aggregation, and the relevance of stability discussions. International Economic Review, 4(1), 53–67.

Batagelj, V., & Mrvar, A. (1998). Pajek. Program for Large Network Analysis. Connections, 21(2), 47–57.

Bensman, S.J. (2008). Distributional differences of the impact factor in the sciences versus the social sciences: An analysis of the probabilistic structure of the 2005 journal citation reports. Journal of the American Society for Information Science and Technology, 59(9), 1366–1382.

Bensman, S.J., & Leydesdorff, L. (2009). Definition and identification of journals as bibliographic and subject entities: Librarianship vs. ISI journal citation reports (JCR) methods and their effect on citation measures. Journal of the American Society for Information Science and Technology.

Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of community hierarchies in large networks. Retrieved December 17, 2008, from http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.0476v2.pdf

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

Chan, L.M. (1999). A guide to the library of congress classification (5th ed.). Englewood, CO: Libraries Unlimited.

Garfield, E. (1971). The mystery of the transposed journal lists—wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. Current Contents, 3(33), 5–6.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science 178 (Number 4060), 471–479.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics, 56(3), 357–367.

Grupp, H. (1990). The concept of entropy in scientometrics and innovation research. An indicator for institutional involvement in scientific and technological developments. Scientometrics, 18(3–4), 219–239.

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

Klavans, R., & Boyack, K. (2009). Towards a consensus map of science. Journal of the American Society for Information Science and Technology (in print, DOI: 10.1002/asi.20991).

Lambiotte, R., & Panzarasa, P. (in press). Communities, knowledge creation, and information diffusion. Journal of Informatics.

Leydesdorff, L. (1987). Various methods for the mapping of science. Scientometrics 11, 291–320.

Leydesdorff, L. (1991). The Static and dynamic analysis of network data using information theory. Social Networks, 13, 301–345.

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. (2002). Dynamic and Evolutionary updates of classificatory schemes in scientific journal structures. Journal of the American Society for Information Science and Technology, 53(12), 987–994.

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, 57(5), 601–613.

Leydesdorff, L. (2007). Visualization of the citation impact environments of scientific journals: An online mapping exercise. Journal of the American Society of Information Science and Technology, 58(1), 207–222.

Leydesdorff, L., & Bensman, S.J. (2006). Classification and powerlaws: The logarithmic transformation. Journal of the American Society for Information Science and Technology, 57(11), 1470–1486.

Leydesdorff, L., & Cozzens, S.E. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the science citation index. Scientometrics, 26, 133–154.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 60(2), 348–362.

Lotka, A.J. (1926). The frequency distribution of scientific production. Journal of the Washington Academy of Sciences 16, 317–324.

Morillo, F., Bordons, M., & Gómez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. Journal of the American Society for Information Science and Technology, 54(13), 1237–1249.

Moya-Anegón, F. de, Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., Munoz-Fernández, F.J., & Herrero-Solana, V. (2007). Visualizing the marrow of science. Journal of the American Society for Information Science and Technology, 58(14), 2167–2179.

Marburger J.H. (2005). Wanted: Better benchmarks. Science, 308(20 May 2005), 1087–1087.

National Science and Technology Council, Office of Science and Technology Policy (2008). The science of science policy: A federal research roadmap. Washington, DC.

Newman, M.E.J. (2006a). Finding community structure in networks using the eigenvectors of matrices. Physical Review E, 74(3), 36104.

Newman, M.E.J. (2006b). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23), 8577–8582.

Porter, A.L., Cohen, A.S., Roessner, J.D., & Perreault, M. (2007). Measuring researcher interdisciplinarity. Scientometrics, 72(1), 117–147.

Pudovkin, A.I., & Fuseler, E.A. (1995). Indices of journal citation relatedness and citation relationships among aquatic biology journals. Scientometrics, 32(3), 227–236.

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.

Rafols, I., & Meyer, M. (2009). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. Scientometrics, 81(2), in print. Retrieved December 18, 2008, from http://www.sussex.ac.uk/spru/documents/sewp167.pdf

Rosvall, M., & Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(4), 1118–1123.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Simon, H.A. (1973). The organization of complex systems. In H.H. Pattee (Ed.), Hierarchy theory: The challenge of complex systems (pp. 1–27). New York: George Braziller Inc.

Simon, H.A. (2002). Near decomposability and the speed of evolution. Industrial and Corporate Change, 11(3), 587–599.

Simon, H.A., & Ando, A. (1961). Aggregation of variables in dynamic systems. Econometrica, 29(2), 111–138.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of The Royal Society Interface, 4(15), 707–719.

Theil, H. (1972). Statistical decomposition analysis. Amsterdam/ London: North-Holland.

Tijssen, R., de Leeuw, J., & van Raan, A.F.J. (1987). Quasi-correspondence analysis on square scientometric transaction matrices. Scientometrics 11, 347–361.

van Raan, A.F.J., & van Leeuwen, T.N. (2002). Assessment of the scientific basis of interdisciplinary, applied research. Application of bibliometric methods in Nutrition and Food Research. Research Policy 31, 611–632.

Wallace, M.L., Gingras, Y., & Duhon, R. (2008). A new approach for detecting scientific specialties from raw cocitation networks. Retrieved December 17, 2008, from http://arxiv.org/ftp/arxiv/papers/0807/0807.4903.pdf

Appendix

Example of the Cross-Tables Between Classifications

For the purpose of illustrating the process of matching categories, Table A1 shows the cross-table between the seven largest categories in a random walk versus the ISI Subject Categories. This table can be used to identify those categories that do not match between classifications and that, therefore, may be of policy interest. The mismatches may be indexer effects or reflect an interdisciplinary process. By analyzing the off-diagonal elements of the matrices, one can learn to what extent the delineations are fuzzy and which categories can be considered as complementary disciplinary sources for the fuzzy sets.

For example, 199 journals are subsumed under the ISI subject category of *Pharmacology and Pharmacy* of which only 89 are in the biomedical sciences according to the random walk, while 12 are organized in the behavioural sciences, and 40 in clinical medicine. At the other extreme, we find 100% agreement between the ISI Subject Categories and the random walk in the case of 36 *Palenteology* journals. This suggests that this is a very clearly defined research field. As noted, the overlap in the case of the biomedical sciences may be a spurious effect from the overlap among the ISI Subject Categories.

TABLE A1.   Number of shared journals between categories of the random walk decomposition (columns) and the ISI Subject Categories (rows).

Rows → random walk category ← Columns ISI subject categories	Biomed. sci.	Behav. sci.	Clinical med.	Biology & ecology	Phys. & materials	Geosci.	Econ. & geogr.	Total journals	% in top categ.
Biochem. mol. biol.	206	1	5	8	0	0	0	262	79%
Cell biology	140	0	1	1	0	0	0	156	90%
Neurosciences	115	38	0	1	0	0	0	199	58%
Genetics & heredity	97	4	3	13	0	0	0	131	74%
Pharmacol. & pharm.	89	12	40	0	0	0	0	199	45%
Biotech. & microbiology	86	0	4	5	0	0	0	140	61%
Immunology	83	0	0	0	0	0	0	117	71%
Microbiology	54	1	0	4	0	0	0	88	61%
Physiology	50	3	2	6	0	1	0	79	63%
Med., res. & exp.	49	0	17	0	0	0	0	76	64%
Endocrin. & metabolism	46	1	21	0	0	0	0	93	49%
Biophysics	46	1	0	0	0	1	0	66	70%
Oncology	44	0	1	0	0	0	0	127	57%
Toxicology	43	0	5	0	0	0	0	76	57%
Biochem. res. methods	35	0	0	0	1	0	0	56	63%
Biology	29	0	0	16	0	2	0	64	45%
Hematology	25	0	16	0	0	0	0	61	41%
Developmental biology	25	0	0	0	0	0	0	34	74%
Psychiatry	3	118	0	0	0	0	0	133	89%
Psych., multidisciplinary	0	80	0	0	0	0	1	84	95%
Education	0	74	0	0	0	0	3	93	80%
Psychology, clinical	0	68	0	0	0	0	0	70	97%
Psychology, developmental	0	47	0	0	0	0	0	47	100%
Psychology	2	46	0	0	0	0	0	61	75%
Rehabilitation	0	45	1	0	0	0	0	72	63%
Psychology, social	0	42	0	0	0	0	0	46	91%
Psychology, experimental	0	42	0	0	0	0	0	44	95%
Psychology, educational	0	38	0	0	0	0	0	38	100%
Medicine, general & internal	6	2	86	0	0	0	0	103	83%
Cardiovascular systems	5	0	65	0	0	0	0	74	88%
Public, env.& occ. health	3	32	52	0	0	0	0	141	37%
Health care sci.	0	2	46	0	0	0	0	56	82%
Periph. vascular disease	9	0	39	0	0	0	0	52	75%
Pediatrics	2	5	32	0	0	0	0	74	43%
Nutrition & dietetics	8	1	30	0	0	0	1	55	55%
Surgery	2	0	12	0	0	0	0	138	28%
Med. lab. technology	7	0	11	0	0	0	0	25	44%
Geriatrics & gerontology	8	6	10	0	0	0	0	30	33%
Medical informatics	1	0	10	0	0	0	0	20	50%
Ecology	0	0	0	104	0	2	1	114	91%
Zoology	9	0	0	85	0	0	0	114	75%
Marine biology	1	0	1	73	0	0	0	79	92%
Entomology	5	0	0	58	0	0	0	69	84%
Mat. sci., multidiscipl.	0	0	0	0	94	1	0	175	54%
Physics, applied	0	0	0	0	69	1	0	84	82%
Metal. & metal. eng.	0	0	0	0	51	0	0	65	78%
Geosciences, multidiscipl.	1	0	0	4	0	101	0	131	77%
Geochem. & geophysics	0	0	0	0	0	58	0	59	98%
Meteorology	0	0	0	0	0	46	0	48	96%
Paleontology	0	0	0	0	0	36	0	36	100%
Geology	0	0	0	0	0	35	0	36	97%
Oceanography	0	0	0	18	0	24	0	48	50%
Mineralogy	0	0	0	0	1	23	0	26	88%
Geography, physical	0	0	0	7	0	22	0	30	73%
Economics	0	0	0	0	0	0	145	154	94%
Business, finance	0	0	0	0	0	0	37	39	95%
Geography	0	0	0	0	0	0	36	36	100%
Environmental studies	0	4	0	0	0	0	35	45	78%
Planning & development	0	1	0	0	0	0	29	37	78%
Urban studies	0	2	0	0	0	0	23	27	85%
Total # of journals in category	942	722	456	437	375	301	301		
% in top category	22%	16%	19%	24%	25%	34%	48%		