Classification of Scientific Networks Using Aggregated Journal-Journal Citation Relations in the Journal Citation Reports

C.-M. Chen
Department of Physics, National Taiwan Normal University, Taipei, Taiwan, and Interdisciplinary Research in the Mathematical and Computational Sciences Centre, Simon Fraser University, Burnaby, British Columbia, Canada. E-mail: cchen@phy.ntnu.edu.tw

I propose an approach to classifying scientific networks in terms of aggregated journal-journal citation relations of the ISI Journal Citation Reports using the affinity propagation method. This algorithm is applied to obtain the classification of SCI and SSCI journals by minimizing intracategory journal-journal (J-J) distances in the database, where distance between journals is calculated from the similarity of their annual citation patterns with a cutoff parameter, *t*, to restrain the maximal J-J distance. As demonstrated in the classification of SCI journals, classification of scientific networks with different resolution is possible by choosing proper values of *t*. Twenty journal categories in SCI are found to be stable despite a difference of an order of magnitude in *t*. In our classifications, the level of specificity of a category can be found by looking at its value of $\overline{D}_{RJ}$ (the average distance of members of a category to its representative journal), and relatedness of category members is implied by the value of $\overline{D}_{J\text{-}J}$ (the average J-J distance within a category). Our results are consistent with the ISI classification scheme, and the level of relatedness for most categories in our classification is higher than their counterpart in the ISI classification scheme.

Introduction

The scientific knowledge of human beings is a complex and dynamic network. Due to its complexity, finding a perfect classification scheme for scientific networks remains an unsolved problem for scientists and librarians thus far. The original classification of journals was carried out in the early 1970s by combining subjective assignment and visual examination of cross-citation patterns among journals (Narin, 1976). Traditional classification methods (Glänzel & Schubert, 2003) are based on subjective analysis, whose output could vary from one person to another. In other words, these methods are more artistic than scientific. In practice, a subjective approach could lead to useful and flexible classification schemes as a result of human intelligence and expertise. Among existing subjective classification systems, the classification system developed by Institute of Scientific Information (ISI) is worthy of special attention. The ISI Journal Citation Reports (JCR) presents interjournal citation frequencies for thousands of journals annually. On the other hand, a quantitative approach to classification is usually constructed based on a set of simple rules, which offers robust classification schemes that do not rely on human interference. Although a classification scheme based on a quantitative approach can be comprehended more easily, so far, a complete and reliable classification scheme for most scientific databases is not yet available.

The aggregated journal-journal (J-J) citation data in JCR contain extensive information about interjournal citations, which could provide an understanding of the interaction among various scientific disciplines. Based on JCR citation data, Pudovkin and Garfield (2002) have used an intuitive criterion (relatedness factor) for finding semantically related journals. To avoid subjective analysis, various quantitative methods have been proposed to construct a robust classification system of scientific journals using JCR citation information. A variety of techniques for analyzing J-J citation relationships have been reported in the literature to cluster scientific journals (Doreian & Fararo, 1985; Leydesdorff, 1986; Tijssen, De Leeuw, & Van Raan, 1987). For example, by applying the notion of structure equivalence to analyze a small set of journals, Doreian and Fararo have delineated a set of blocks, which contain journals. These blocks have a very close correspondence to a categorization of the journals based on their aims and objectives. More recently Leydesdorff and Cozzens (1993) have developed an optimization procedure that stabilizes approximated eigenvectors of the scientific network from principal component analysis as representations of clusters. This principal component analysis has been

Received May 23, 2008; revised July 3, 2008; accepted July 3, 2008

© 2008 ASIS&T • Published online 13 August 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20935

further extended to rotated component analysis (Leydesdorff, 2006; Leydesdorff & Cozzens, 1993), which enables one to focus on specific subsets with internal coherence. Local optimizations have also been applied to cluster document sets in terms of other indicators. An alternative method of cocitation clustering has been investigated in constructing a *World Atlas of Sciences* for ISI (Garfield, Malin, & Small, 1975; Leydesdorff, 1987; Small, 1999). In spite of the recent progress in developing quantitative classification methods, it is still desirable to develop a fast, convenient, and operable classification scheme that can be applied to cluster a large set of journals based on quantitative citation data.

In this article, I propose a quantitative approach to classify the scientific network in terms of aggregated J-J citation relations of JCR using the affinity propagation method (Frey & Dueck, 2007). This method has been previously used to cluster images of faces, detect genes in microarray data, identify representative sentences of an article, and identify cities that are efficiently accessed by airline travel. Here we have applied the affinity propagation method to cluster journals in both the Science Citation Index (SCI) and the Social Science Citation Index (SSCI). We note that the number of journal categories, or their size, is a result, instead of an input, of the affinity propagation clustering method.

Method

The method used by ISI in establishing journal categories for JCR is a heuristic approach, in which the journal categories have been manually developed initially. The assignment of journals was based upon a visual examination of all relevant citation data. As the number of journals in a category grew, subdivisions of the category were then established subjectively. Although this is a useful approach, a more robust, convenient, and automatic classification scheme is desired.

Here I delineate this quantitative approach to clustering scientific journals in terms of JCR citation data using the affinity propagation method. The citation data analyzed include the SCI of 2001 and the SSCI of 2005, which are directly computed from the extraction of the CD version of the ISI database. There are 2,195 journals of impact factor greater than 1 in the 2001 SCI. After removing 290 journals that did not publish any articles in 2001, there are 1,905 journals left in our data set, which contains 426,065 articles and 13,798,138 citations. For the 2005 SSCI, there are 1,583 journals in the database, of which 1,578 journals have nonzero contents. The SSCI database contains 66,051 articles and 2,437,389 citations. In principle, the dissimilarity between two journals can be visualized by the differences in their citation patterns. In other words, the citation pattern of each journal is represented by a normalized citation vector, and these vectors form a rescaled citation matrix. The dissimilarity (or similarity) in citation between two journals is related to the scalar product of their citation vectors.

To begin with, the citation matrix $\{N_{ij}\}$ (number of citations of journal j cited by journal i) is calculated from a dataset Ω. The similarity of two journals i and j in their citation patterns is defined as its cosine measure:

$$cs_{ij} = \frac{\sum_{k\in\Omega} c_{ik}c_{jk}}{\sqrt{\sum_{k\in\Omega} c_{ik}{}^2 \cdot \sum_{k\in\Omega} c_{jk}{}^2}}, \qquad (1)$$

where $c_{ik} \equiv N_{ik}/(\sum_{j\in\Omega} N_{ij})$ is the normalized citation-matrix element (Samoylenko, Chao, Liu, & Chen, 2006). Depending on the similarity in the citation patterns of journals i and j, the value of cs_{ij} ranges from 0 to 1. For mapping or visualization, coefficients of similarity are converted into distances such that closely related journals are short distances apart and remotely related journals are long distances apart. We express this conversion as

$$d_{ij} = \sqrt{\frac{1}{\max(t, cs_{ij})} - 1}, \qquad (2)$$

where t is a cutoff parameter for this distance conversion and the function $\max(a, b)$ chooses the bigger value from a and b. For the case of $t = 0$, the distance between two unrelated journals is infinite. In principle, a different value of t would correspond to a different approximated distribution of J-J distance ($D_{\text{J-J}}$) for the data set. In the Results and Discussion section, we will illustrate that clustering journals at different levels of resolution can be achieved by choosing proper values of t.

The affinity propagation method takes as input a collection of similarities between journals, where the similarity $s(i, j)$ measures how well journal j is suited to be the representative of a journal category for journal i. Since the goal is to minimize squared error, we set $s(i, j) = -d_{ij}$. Instead of requiring that the number of journal categories be prespecified, affinity propagation takes as input a real number $s(j, j)$ for each journal j so that journals with larger values of $s(j, j)$ are more likely to be chosen as representatives. There are two types of messages exchanged between journals, including the responsibility $r(i, j)$, which is sent from journal i to candidate representative journal (RJ) j, and the availability $a(i, j)$, which is sent from candidate representative journal j to journal i. Here the responsibility reflects the accumulated evidence for how well-suited journal j is to serve as the representative for journal i, and the availability shows the accumulated evidence for how appropriate it would be for journal i to choose journal j as its representative. Taking into account other potential representative journals for journal i, the responsibility is computed iteratively as

$$r(i, j) \leftarrow s(i, j) - \max_{j' \text{ s.t. } j' \neq j} \{a(i, j') + s(i, j')\}, \qquad (3)$$

where the initial value of $a(i, j)$ is set to zero in the first iteration. Similarly, taking into account the support from other journals that journal j should be a representative, the availability is updated by gathering evidence from journals as to whether each candidate representative would make a good representative journal:

$$a(i, j) \leftarrow \min\left\{0, r(j, j) + \sum_{i' \text{ s.t.} i' \neq \{i, j\}} \max\{0, r(i', j)\}\right\}. \tag{4}$$

To reflect accumulated evidence that journal j is a representative based on the positive responsibilities sent to candidate representative j from other journals, the self-availability is updated as

$$a(j, j) \leftarrow \sum_{i' \text{ s.t. } i' \neq j} \max\{0, r(i', j)\}. \tag{5}$$

During the process of affinity propagation, the sum of availability and responsibility can be used to identify the representative journal of emerging journal categories. In other words, for any journal i, the value of j that maximizes $a(i, j) + r(i, j)$ identifies that journal j is its representative. For further information on computing $a(i, j)$ and $r(i, j)$, we refer readers to the paper by Frey and Dueck (2007) in *Science*.

Our classification scheme of scientific networks is based on the application of the affinity propagation method to cluster SCI and SSCI journals. Since different values of the cutoff parameter (t) lead to different levels of approximation in positioning journals in a high-dimensional space, in the Results and Discussion section we show that classifications with different resolutions are possible by choosing proper values of t. Statistical analysis of our classifications will be further performed and compared with that of ISI classifications. In our classifications, the level of specificity of a category can be found by looking at its value of $\overline{D}_{\text{RJ}}$ (the average distance of members of a category to its representative journal), and relatedness of category members is implied by the value of $\overline{D}_{\text{J-J}}$ (the average J-J distance within a category).

Results and Discussion

To demonstrate the applicability of the affinity propagation method in clustering a complete data set of journals, we first apply it to cluster journals in the 2005 SSCI database. Here the cutoff parameter t is set to 0.0001, implying that the maximal value of $D_{\text{J-J}}$ ($D_{\text{J-J}}{}^{\max}$) is 100. This choice of t is quite reasonable since the probability distribution (PD), or normalized histogram (bin size is 1), of $D_{\text{J-J}}$ in the unclustered SSCI journal database is mostly between 0 and 30, as shown in Figure 1. With a choice of $D_{\text{J-J}}{}^{\max} = 100$, the distance between unrelated journals is much larger than that between related journals. In other words, for any journal category, unrelated journals will not be located in the vicinity of its members (each journal is considered as a point in a high-dimensional space). Thus only correlated journals will be grouped together by the affinity propagation method. However, if $D_{\text{J-J}}{}^{\max}$ is too close to 30, the positions of unrelated journals are not well separated and the distortion to the journal positions due to the introduction of the cutoff would affect the clustering of journals. For the predicted SSCI classification, only those J-J distances within the same category are considered in calculating its PD of $D_{\text{J-J}}$. In Figure 1, there are two peaks observed from the statistical curves of PD in $D_{\text{J-J}}$, where the first peak shows the relatedness between journals within the database (or categories), while the second peak at $D_{\text{J-J}} = 100$ indicates the irrelevance between journals within the database (or categories). For the predicted SSCI classification, clearly its first peak in the PD of $D_{\text{J-J}}$ is much more prominent and the peak width is much more narrow than that of the unclustered SSCI database. On the other hand, its second peak of irrelevance is much smaller than that of the unclustered database. The probability distribution of the first peak is found to decrease exponentially with $D_{\text{J-J}}$, i.e., $P = P_0 \exp[-(D_{\text{J-J}} - d_0)/\Delta]$, where P_0 is the peak value, d_0 is the peak position, and Δ is the decay width. By fitting the statistical data, we find that $d_0 = 4$ and $\Delta = 9.08$ for the unclustered SSCI curve, while $d_0 = 2$ and $\Delta = 1.72$ for the clustered SSCI curve.

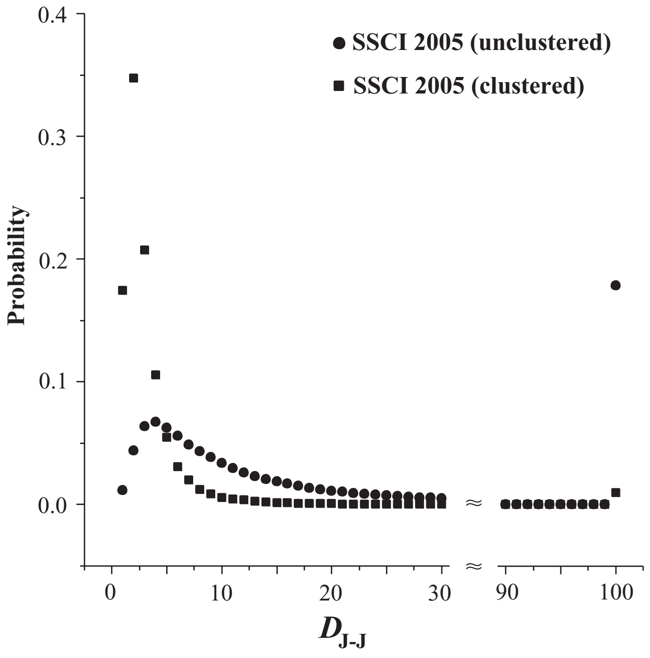

FIG. 1. Probability distributions of the journal-journal distance ($D_{\text{J-J}}$) for both the unclustered SSCI database of 2005 and its predicted classification. The data are shown selectively to focus on peaks at small or large $D_{\text{J-J}}$. The probability distribution for intermediate values of $D_{\text{J-J}}$ is essentially zero, and data within this range are omitted. For the predicted SSCI classification, only those J-J distances within the same category are considered in calculating the probability distribution.

The results of our SSCI classification are presented in Table 1. The details of our journal classification can be found in the supporting information (Chen, 2008). The entire journal set of SSCI is decomposed into 23 journal categories. Among them, the categories Business, Finance and Business, Marketing contain business journals specialized in finance and marketing. The categories Psychology; Psychology, Clinical; Psychology, Developmental; and Psychology, Cognition and Experimental contain journals specialized

TABLE 1. Categories and their associated properties in the SSCI classification using $t = 10^{-4}$.

SSCI Category ($t = 10^{-4}$)	N_j	RJ	$\overline{D}_{RJ}$	$\overline{D}_{J\text{-}J}$
Anthropology	46	Am Anthropol	1.93	11.44
Business, Finance	29	J Bus	0.77	1.31
Business, Marketing	36	J Bus Res	1.02	1.75
Development	37	Dev Change	1.70	7.80
Economics	166	Rev Econ Stat	1.01	1.78
Education	70	Educ Stud	1.89	6.60
Ergonomics	22	Int J Hum Comput Int	1.51	6.42
Geography & Environmental Studies	65	Environ Plann A	1.69	5.07
Health & Social Science	87	Soc Sci Med	1.59	3.90
History & Family Studies	23	J Fam Hist	1.90	23.68
History & Philosophy of Science	23	ISIS	1.68	5.17
Information Science & Library Science	33	Libr Inform Sci Res	1.94	7.77
Law	78	New York U Law Rev	1.03	2.40
Linguistics	15	Appl Linguist	1.62	4.71
Management	61	Acad Manage Rev	1.14	2.02
Nursing, Aging, & Health	74	Western J Nurs Res	1.54	3.19
Political Science	95	Polit Sci Quart	1.56	8.29
Psychiatry	104	Curr Opin Psychiat	1.14	2.48
Psychology	158	J Psychol	1.25	2.53
Psychology, Clinical	92	Psychol Psychother T	1.56	4.94
Psychology, Developmental	79	Dev Psychol	1.11	1.91
Psychology, Cognition & Experimental	81	J Exp Psych Gen	1.43	2.75
Sociology	104	Soc Forces	1.38	4.24

in general psychology, clinical psychology, developmental psychology, and cognition and experimental psychology, respectively. The title of each category is given based on its core journals, i.e., the representative journal (RJ) and its members whose distance to RJ (D_{RJ}) is less than 1. The size (or number of journals, N_j) of SSCI categories varies from 22 to 166. The relatedness of journals within a category can be seen as the average value of $D_{J\text{-}J}$ within the category ($\overline{D}_{J\text{-}J}$), and the specificity of a category is related to the average distance of category members to its RJ ($\overline{D}_{RJ}$). For any category, a smaller value of $\overline{D}_{RJ}$ implies a higher level of specificity, and a smaller value of $\overline{D}_{J\text{-}J}$ implies that journals within a category are more closely related to each other.

In the predicted SSCI classification, the two categories of smallest $\overline{D}_{RJ}$ and $\overline{D}_{J\text{-}J}$ are Business, Finance and Economics, whose values of $\overline{D}_{RJ}$ and $\overline{D}_{J\text{-}J}$ are 0.77 and 1.31, and 1.01 and 1.78, respectively. For the category of Business, Finance, 23 journals (out of 29 journals) of our classification are consistent with those of the ISI classification. The calculated value of $\overline{D}_{J\text{-}J}$ for those journals in the category of Business, Finance in the ISI classification is 2.10, which is larger than its corresponding value of 1.31 in our classification. For the category of Economics, 138 journals (out of 166 journals) are consistent with those of the ISI classification. The calculated value of $\overline{D}_{J\text{-}J}$ for those journals in the category of Economics in the ISI classification is 3.12, which is larger than the corresponding value of 1.78 in our classification. In the predicted SSCI classification, the two categories of largest $\overline{D}_{RJ}$ and $\overline{D}_{J\text{-}J}$ are History and Family Studies, and Anthropology. For the category of Anthropology, 34 journals (out of 46 journals) of our classification are consistent with those of the ISI classification. The calculated value of $\overline{D}_{J\text{-}J}$ for those journals in the category of Anthropology in the ISI classification is 15.67, which is larger than the corresponding value of 11.44 in our classification. In general most categories in our classification scheme have a corresponding category in the ISI classification scheme, and their value of $\overline{D}_{J\text{-}J}$ seems to be smaller than that of their counterpart in the ISI classification scheme. Nonetheless, the category of History and Family Studies contains journals related to history, family studies, and gender studies, which can be expected from its large value of $\overline{D}_{J\text{-}J}$. We note that the above comparisons are based on the SSCI classification in the 2005 JCR. In Figure 2, we show the probability distribution of $D_{J\text{-}J}$ for the above mentioned four categories. There are no drastic differences for the first peak of these category curves ($d_0 = 1$, $\Delta = 1.51$ for Business, Finance; $d_0 = 2$, $\Delta = 0.90$ for Economics; $d_0 = 2$, $\Delta = 3.66$ for Anthropology; $d_0 = 3$, $\Delta = 2.02$ for History and Family Studies). However, the second peak for the category of History and Family Studies is rather significant (0.2), indicating that a large proportion of the J-J relationship is irrelevant in this category.

After demonstrating the application of the affinity propagation method in the classification of a complete journal database (SSCI), we further examine its applicability in classifying an incomplete journal database, for example, 1905 journals of impact factor greater than 1 in the SCI database. In addition, we would like to investigate the effects of the cutoff parameter t in clustering journals. The probability distributions of $D_{J\text{-}J}$ for the unclustered SCI and two predicted SCI classifications (using $t = 10^{-4}$ and $t = 10^{-3}$) are shown in Figure 3. It is found that the second peak of J-J irrelevance for the unclustered SCI is ten times smaller than that of the unclustered SSCI, showing the characteristic difference

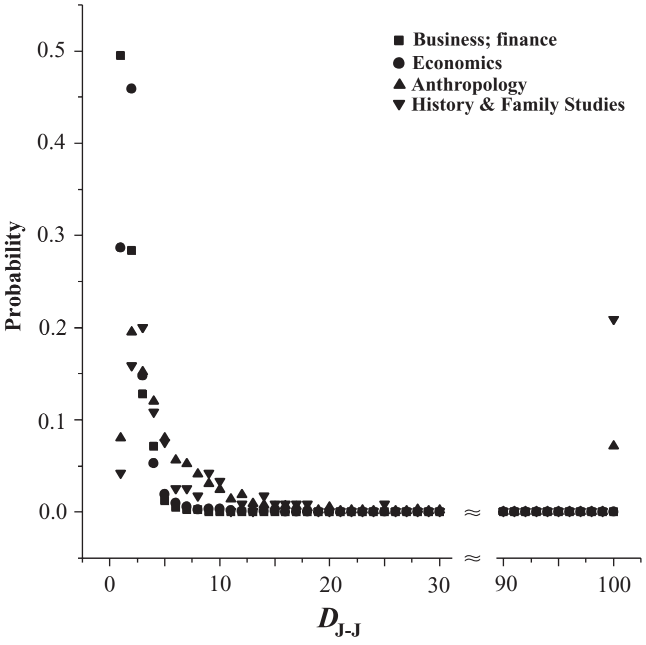

FIG. 2. Probability distributions of the journal-journal distance ($D_{\text{J-J}}$) for four predicted categories, including Business, Finance; Economics; Anthropology; and History and Family Studies. The data are shown selectively to focus on peaks at small or large $D_{\text{J-J}}$. The probability distribution for intermediate values of $D_{\text{J-J}}$ is essentially zero, and data within this range are omitted.

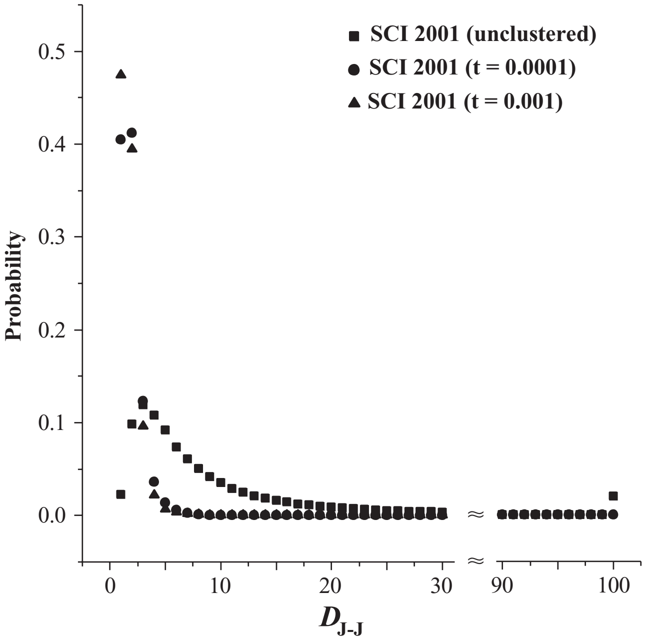

FIG. 3. Probability distributions of the journal-journal distance ($D_{\text{J-J}}$) for the unclustered SCI database of 2001 and two predicted ($t = 10^{-4}$ and $t = 10^{-3}$) SCI classifications. The data are shown selectively to focus on peaks at small or large $D_{\text{J-J}}$. The probability distribution for intermediate values of $D_{\text{J-J}}$ is essentially zero and data within this range are omitted. The maximal distance of $D_{\text{J-J}}$ is only 31.6 for the case of $t = 10^{-3}$. For the predicted SCI classifications, only those J-J distances within the same category are considered in calculating the probability distribution.

between these two databases. For the two predicted SCI classifications, the peak of J-J irrelevance completely vanishes. Moreover, both predicted classifications of SCI have a rather dominant first peak, showing that most journals within each category are highly related. It is found that $d_0 = 2$ and $\Delta = 0.83$ for the curve of $t = 10^{-4}$, $d_0 = 2$ and $\Delta = 0.71$ for the curve of $t = 10^{-3}$, and $d_0 = 3$ and $\Delta = 5.92$ for the curve of unclustered SCI. We note that, since the probability distribution of $D_{\text{J-J}}$ in the unclustered SCI journal set is mostly between 0 and 20, it is reasonable to choose $t \geq 10^{-3}$ such that ${D_{\text{J-J}}}^{\max} > 30$.

The results for the predicted classification of SCI using $t = 10^{-4}$ are presented in Table 2. Detailed information about the classification of journals is available at the supporting Web site (Chen, 2008). The SCI database contains 1,905 journals of impact factor greater than 1 and is decomposed into 56 journal categories. Among these categories, Chemistry, Chemistry, Analytical, and Chemistry, Physical contain journals in general chemistry, analytical chemistry, and physical chemistry, respectively. Other disciplines with subfields emerging from our classification scheme include Engineering (chemical, communication, and electrical and electronic), Mathematics (general and applied), Microbiology (general and medical), and Physics (general, applied, and nuclear and particle). Although the number of journals in this journal set is only 1.2 times of that of SSCI journal set, the number of its categories is 2.4 times of that of SSCI journal set. Moreover, the values of $\overline{D}_{\text{RJ}}$ and $\overline{D}_{\text{J-J}}$ of SCI categories are much smaller than those of SSCI categories. These discrepancies reflect the characteristic difference in their probability distributions of $D_{\text{J-J}}$. The size of SCI categories varies between 5 and 138. In the SCI journal set, both biological science and medical science publish a great number of journals. However, biological journals tend to form few large clusters, while medical journals tend to form many small clusters.

In this predicted SCI classification, the two categories of smallest $\overline{D}_{\text{RJ}}$ and $\overline{D}_{\text{J-J}}$ are Astronomy and Astrophysics and Ophthalmology, whose values of $\overline{D}_{\text{RJ}}$ and $\overline{D}_{\text{J-J}}$ are 0.32 and 0.46 as well as 0.47 and 0.71, respectively. In our classification scheme, the category of Astronomy and Astrophysics contains 13 journals, all of which are also included in the category of Astronomy and Astrophysics of the ISI classification scheme. The calculated value of $\overline{D}_{\text{J-J}}$ for those journals in the category of Astronomy and Astrophysics in the ISI classification is 2.39, which is larger than its corresponding value of 0.46 in our classification scheme. Note that, throughout this paper, only SCI journals with impact factor greater than 1 in 2001 are considered in our calculation. The category of Ophthalmology contains 15 journals in our classification, all of which are also included in the category of Ophthalmology of ISI classification. The calculated value of $\overline{D}_{\text{J-J}}$ for those journals in the category of Ophthalmology in the ISI classification is 1.24, which is larger than its corresponding value of 0.71 in our classification scheme. The category of Engineering, Communication, in our classification scheme contains 21 journals, most of which are under categories Engineering,

TABLE 2. Categories and their associated properties in the SCI classification using $t = 10^{-4}$.

SCI Category ($t = 10^{-4}$)	N_{j}	RJ	$\overline{D}_{\mathrm{RJ}}$	$\overline{D}_{\mathrm{J\text{-}J}}$
Acoustics & Otology	9	Ear Hearing	1.23	2.13
Agriculture	17	Adv Agron	1.15	2.17
Astronomy & Astrophysics	13	Astrophys J	0.32	0.46
Biochemistry & Molecular Biology	138	Faseb J	0.67	1.03
Biology	50	P Roy Soc Lond B Bio	1.17	2.17
Biomedical Sciences	136	P Natl Acad Sci USA	0.76	1.22
Cardiac & Cardiovascular Systems	38	Curr Opin Cardiol	0.72	1.19
Chemistry	73	Chem Eur J	0.72	1.13
Chemistry, Analytical	40	Analyst	0.88	1.44
Chemistry, Physical	29	Annu Rev Phys Chem	0.59	0.90
Clinical Neurology	40	J Neurol Neurosur Ps	0.82	1.40
Computer Science	19	ACM Comput Surv	1.21	2.66
Dentistry, Oral Surgery, & Medicine	14	J Dent Res	0.91	1.49
Ecology	45	Ecoscience	0.98	1.67
Endocrinology & Metabolism	44	Endocr Rev	0.82	1.35
Engineering, Chemical	17	Ind Eng Chem Res	1.24	2.33
Engineering, Communication	21	Digit Signal Process	1.42	2.94
Engineering, Electrical & Electronic	10	Electron Lett	1.09	2.08
Environmental Sciences	21	Environ Sci Technol	1.14	1.83
Food Science & Technology	19	Trends Food Sci Tech	1.21	2.58
Gastroenterology & Hepatology	25	Gastroenterology	0.62	0.96
Genetics & Heredity	24	Eur J Hum Genet	0.84	1.36
Geochemistry & Geophysics	27	Rev Geophys	1.29	2.96
Geosciences	48	Earth Sci Rev	0.96	1.76
Hematology	29	Semin Hematol	0.56	0.86
Immunology	63	Clin Immunol	0.61	0.98
Marine & Freshwater Biology	22	Mar Ecol Prog Ser	0.87	1.35
Materials Science	20	Int Mater Rev	1.22	2.71
Mathematics	11	Ann Math	0.76	1.14
Mathematics, Applied	7	Siam J Sci Comput	0.94	1.32
Mechanics	10	Annu Rev Fluid Mech	0.83	1.46
Medicine	71	Am J Med Sci	1.15	2.19
Microbiology	33	Curr Microbiol	0.74	1.18
Microbiology, Medical	60	Clin Microbiol Rev	1.07	2.11
Mineralogy	21	Contrib Mineral Petr	0.87	1.36
Neurosciences	98	Neuroreport	0.73	1.16
Nutrition & Dietetics	22	Nutr Rev	0.69	1.11
Oncology	71	J Cancer Res Clin	0.77	1.32
Operations Research & Management Science	5	Math Oper Res	0.73	1.03
Ophthalmology	15	Graef Arch Clin Exp	0.47	0.71
Orthopedics	17	J Orthopaed Res	1.18	2.39
Pathology	17	Semin Diagn Pathol	0.98	2.15
Pharmacology & Pharmacy	54	Drug Safety	1.19	2.20
Physics	58	Europhys Lett	0.69	1.10
Physics, Applied	48	Jpn J Appl Phys 1	0.92	1.65
Physics, Nuclear & Particle	21	J Phys G Nucl Partic	0.49	0.80
Plant Sciences	33	Plant Physiol Bioch	0.64	1.03
Polymer Science	16	Macromol Chem Physic	0.78	1.22
Psychiatry	40	Am J Psychiat	0.70	1.02
Psychology	13	Psychol Rev	1.61	3.78
Public, Environmental, & Occupational Health	24	Epidemiology	1.00	1.53
Radiology, Nuclear Medicine, & Medical Imaging	20	Radiology	0.92	1.50
Statistics & Probability	9	J Roy Stat Soc B	0.94	1.77
Surgery	23	World J Surg	0.89	1.55
Toxicology	27	Toxicology	0.80	1.18
Water Resources	10	Water Resour Res	0.96	1.65

Electrical and Electronic; Telecommunication; and Computer Science in the ISI classification scheme. The Psychology category in our classification contains 13 journals, most of which belong to Psychology and Behaviour Science in the ISI classification. However, two journals (*Presence: Teleoperators and Virtual Environments* and *Journal of Product Innovation Management*) in this category seem to be incompatible with other journals in this category. These two journals are sorted into this category because they are even more incompatible with other categories. We note that the above comparisons are based on the SCI classification in 2001 JCR. A better classification is expected if more SCI journals are included. The probability distributions of $D_{\text{J-J}}$ for Astronomy and Astrophysics; Ophthalmology; Engineering, Communication; and Psychology are depicted in Figure 4.

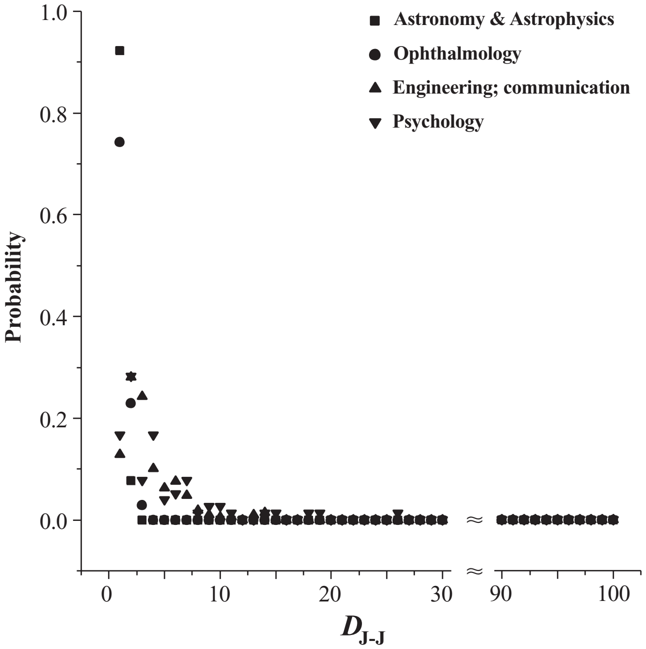

FIG. 4. Probability distributions of the journal-journal distance ($D_{\text{J-J}}$) for four predicted categories using $t = 10^{-4}$, including Astronomy and Astrophysics; Ophthalmology; Engineering, Communication; and Psychology. The data are shown selectively to focus on peaks at small or large $D_{\text{J-J}}$. The probability distribution for intermediate values of $D_{\text{J-J}}$ is essentially zero and data within this range are omitted.

The probability distributions of $D_{\text{J-J}}$ of Astronomy and Astrophysics and Ophthalmology highly populate at $D_{\text{J-J}} = 1$, while the distribution of $D_{\text{J-J}}$ in Psychology has a nonzero value for $D_{\text{J-J}} > 20$.

When a larger value of the cutoff parameter is used, the maximal distance of $D_{\text{J-J}}$ becomes smaller. For $t = 10^{-3}$, we have $D_{\text{J-J}}{}^{\max} = 31.61$. Since the high-dimensional J-J distance space is now approximated by a high-dimensional sphere of smaller radius, the resolution in clustering journals is higher in this case. Thus the SCI database is expected to be decomposed into more clusters for $t = 10^{-3}$, compared to the case of $t = 10^{-4}$. The results of the SCI classification using $t = 10^{-3}$ are presented in Table 3. Detailed information about the classification of journals is given at the supporting Web site (Chen, 2008). There are 71 categories in this predicted SCI classification. Several specific disciplines emerge from reorganizing more coarse-grained categories that are predicted by using $t = 10^{-4}$, such as Anesthesiology; Chemistry, Catalysis; Dermatology; Nephrology; Rheumatology; Urology; Veterinary Sciences; and Virology. Journals in Anesthesiology are separated from other journals in Pharmacology and Pharmacy and form a specific subfield. Journals in the category of Chemistry, Catalysis are recombined from Chemistry and Engineering, Chemical. Journals in Dermatology, Nephrology, Rheumatology, and Urology were originally distributed in Medicine, Pathology, Biochemistry and Molecular Biology, Cardiac and Cardiovascular Systems, and Immunology. Most journals in Veterinary Sciences and Virology were originally grouped into the category of Microbiology, Medical. Therefore, from comparing clustering results with different values of the cutoff parameter, the relationship among various disciplines can be revealed. We note that our classification scheme based on the affinity propagation method is still valid when an incomplete database is considered. For example, only 11 journals

TABLE 3. Categories and their associated properties in the SCI classification using $t = 10^{-3}$.

SCI Category ($t = 10^{-3}$)	N_{j}	RJ	$\overline{D}_{\text{RJ}}$	$\overline{D}_{\text{J-J}}$
Acoustics & Otology	8	J Acoust Soc Am	0.95	1.42
Agriculture	17	Adv Agron	1.15	2.17
Anesthesiology	6	Can J anaesth	0.28	0.33
Astronomy & Astrophysics	14	Astrophys J	0.37	0.56
Behavioral Sciences	16	Anim Behav	0.92	1.50
Biochemistry & Molecular Biology	123	Faseb J	0.64	0.97
Biomedical Sciences	104	P Natl Acad Sci USA	0.72	1.12
Biology	37	Biol J Linn Soc	1.02	1.70
Biophysics & Biochemistry	42	Annu Rev Bioph Biom	0.63	0.94
Cardiac & Cardiovascular Systems	29	Z Kardiol	0.49	0.76
Chemistry	65	Chem Eur J	0.68	1.08
Chemistry, Analytical	40	Analyst	0.88	1.44
Chemistry, Catalysis	8	Catal Lett	0.24	0.32
Chemistry, Physical	28	Annu Rev Phys Chem	0.55	0.81
Clinical Neurology	39	J Neurol Neurosur Ps	0.80	1.36
Computer Science	19	ACM Comput Surv	1.21	2.66
Dentistry, Oral Surgery, & Medicine	14	J Dent Res	0.91	1.49
Dermatology	12	Brit J Dermatol	0.69	0.99
Ecology	40	Ecoscience	0.92	1.57

(*Continued*)

TABLE 3. (*Continued*)

SCI Category ($t = 10^{-3}$)	N_{j}	RJ	$\overline{D}_{\mathrm{RJ}}$	$\overline{D}_{\mathrm{J\text{-}J}}$
Endocrinology & Metabolism	46	Endocr Rev	0.85	1.44
Engineering, Chemical	11	Ind Eng Chem Res	1.09	1.98
Engineering, Communication I	16	Digit Signal Process	1.34	3.30
Engineering, Communication II	9	P IEEE	1.27	4.14
Environmental Sciences	21	Environ Sci Technol	1.14	1.83
Food Science & Technology	18	Trends Food Sci Tech	1.13	2.45
Gastroenterology & Hepatology	26	Gastroenterology	0.63	0.97
Genetics & Heredity	24	Eur J Hum Genet	0.84	1.36
Geography & Remote Sensing	7	Prog Phys Geog	1.28	3.12
Geosciences	54	Earth Sci Rev	1.03	1.95
Hematology	27	Semin Hematol	0.53	0.81
Immunology	52	J Immunol	0.49	0.75
Marine & Freshwater Biology	22	Mar Ecol Prog Ser	0.87	1.35
Materials Science	20	Int Mater Rev	1.22	2.71
Mathematics	11	Ann Math	0.76	1.14
Mathematics, Applied	7	Siam J Sci Comput	0.94	1.32
Mechanics	9	Annu Rev Fluid Mech	0.75	1.16
Medicine	46	JAMA: J Am Med Assoc	0.88	1.37
Meteorology & Atmospheric Sciences	15	Tellus A	0.83	1.30
Microbiology	33	Curr Microbiol	0.74	1.18
Microbiology, Medical	39	Clin Microbiol Rev	0.87	1.55
Mineralogy	21	Contrib Mineral Petr	0.87	1.36
Nephrology	18	Curr Opin Nephrol Hy	0.60	0.99
Neurosciences	94	Neuroreport	0.72	1.14
Nutrition & Dietetics	21	Nutr Rev	0.66	1.02
Oncology	63	J Cancer Res Clin	0.74	1.28
Operations Research & Management Science	5	Math Oper Res	0.73	1.03
Ophthalmology	15	Graef Arch Clin Exp	0.47	0.71
Optics	15	Opt Commun	0.90	1.50
Orthopedics	15	J Orthopaed Res	1.14	2.26
Otorhinolaryngology	6	Arch Otolaryngol	0.62	1.13
Pathology	14	Modern Pathol	0.61	0.87
Pharmacology & Pharmacy	32	Drug Safety	1.00	1.70
Physics	56	Europhys Lett	0.68	1.07
Physics, Applied	41	Jpn J Appl Phys 1	0.78	1.33
Physics, Nuclear & Particle	21	J Phys G Nucl Partic	0.49	0.80
Physiology	18	Acta Physiol Scand	0.73	1.13
Plant Sciences	33	Plant Physiol Bioch	0.64	1.03
Polymer Science	16	Macromol Chem Physic	0.78	1.22
Psychiatry	39	Am J Psychiat	0.67	0.98
Psychology	13	Psychol Rev	1.61	3.78
Public, Environmental, & Occupational Health	25	Int J Epidemiol	1.15	1.95
Radiology, Nuclear Medicine, & Medical Imaging	20	Radiology	0.92	1.50
Respiratory System	15	Eur Respir J	0.69	0.99
Rheumatology	12	Rheumatol Int	0.34	0.47
Statistics & Probability	9	J Roy Stat Soc B	0.94	1.77
Surgery	20	Arch Surg Chicago	0.74	1.20
Toxicology	28	Toxicology	0.82	1.21
Urology	9	Urology	0.34	0.50
Veterinary Sciences	12	Vet J	1.13	1.92
Virology	16	J Gen Virol	0.48	0.69
Water Resources	9	Water Resour Res	0.82	1.34

(out of 170) in SCI Mathematics category are included in the incomplete SCI database (of impact factor greater than 1). Even though most mathematics journals are missing from the incomplete database, a compact and robust Mathematics category is still identified in our classification scheme for both $t = 10^{-4}$ and 10^{-3}, in which all 11 mathematical journals are included as shown in Tables 2 and 3. The robustness of a journal category can also be checked against changing the cutoff parameter. When the value of t varies, the content of a category would be stable if all of its members are relatively far away from other journals. Such robust categories in SCI include Agriculture; Chemistry, Analytical; Computer Science; Dentistry; Oral Surgery and Medicine; Environmental Sciences; Genetics and Heredity; Marine and Freshwater Biology; Materials Science; Mathematics; Mathematics, applied; Operations Research and

Management Science; Ophthalmology; Physics, nuclear and particle; Plant Sciences; Polymer Science; Psychology; Radiology; Nuclear Medicine and Medical Imaging; and Statistics and Probability.

In the clustered SCI database using $t = 10^{-3}$, the two categories of smallest $\overline{D}_{\mathrm{RJ}}$ and $\overline{D}_{\mathrm{J\text{-}J}}$ are Chemistry, Catalysis and Anesthesiology, whose values of $\overline{D}_{\mathrm{RJ}}$ and $\overline{D}_{\mathrm{J\text{-}J}}$ are 0.24 and 0.32, and 0.28 and 0.33, respectively. The category of Chemistry, Catalysis contains 8 journals, and Anesthesiology contains 6 journals. Both categories are small clusters, which are separated from larger journal categories. All journals in Anesthesiology are also included in the ISI Anesthesiology category, and all journals in Chemistry; catalysis are under the ISI category of Chemistry, Physical. The calculated value of $\overline{D}_{\mathrm{J\text{-}J}}$ in ISI Anesthesiology is 0.78, which is larger than its corresponding value of 0.33 in our classification. On the other hand, the two categories of largest $\overline{D}_{\mathrm{R\text{-}J}}$ and $\overline{D}_{\mathrm{J\text{-}J}}$ are Psychology and Engineering, Communication I, whose values of $\overline{D}_{\mathrm{RJ}}$ and $\overline{D}_{\mathrm{J\text{-}J}}$ are 1.61 and 3.78, and 1.34 and 3.30, respectively. The probability distributions of $D_{\mathrm{J\text{-}J}}$ for the above-mentioned four categories are depicted in the supplementary figure S1 (Chen, 2008). The probability distributions of $D_{\mathrm{J\text{-}J}}$ of Chemistry, Catalysis and Anesthesiology are all between 0 and 1, while the distribution of $D_{\mathrm{J\text{-}J}}$ in Psychology has a nonzero value for $D_{\mathrm{J\text{-}J}} > 20$.

Conclusion

In this paper, we have presented an automatic algorithm, the affinity propagation method, for clustering scientific journals. This clustering method has been applied to the classification of SCI and SSCI databases. Our results demonstrate that the affinity propagation method can provide a reasonable classification scheme for either a complete database or an incomplete database. This method does not need the number of categories or their size as an input. Distance between journals is calculated from the similarity of their annual citation patterns with a cutoff parameter to restrain the maximal distance. Different values of the cutoff parameter lead to different levels of resolution in the classification of journal network. A more coarse-grained classification is obtained when a smaller value of the cutoff parameter (or a larger maximal J-J distance) is used. The property of each category is determined by its core journals, which include RJ and other journals that are closely related to RJ. The level of specificity of a category and relatedness between category members can be investigated by looking at the value of $\overline{D}_{\mathrm{RJ}}$ and $\overline{D}_{\mathrm{J\text{-}J}}$. The stability of a journal category against varying the cutoff parameter implies that its members are remotely related to journals in other categories. Twenty journal categories in SCI are found to be particularly stable despite a difference of an order of magnitude in t. We note that, unlike the ISI classification scheme, which allows overlap in the content of journal categories by subjective decisions, each journal uniquely belongs to a category in our classification scheme. Also note that our classification scheme is not necessary a perfect one, although it tries to maximize the responsibility and availability between RJ and other members for each category of the entire database. Some minor adjustments might still be needed depending on the value of $\overline{D}_{\mathrm{RJ}}$ and $\overline{D}_{\mathrm{J\text{-}J}}$ of each category. In general, our classification results are consistent with the ISI classification scheme. The average J-J distance within a category in our classification is significantly smaller than that in a similar ISI category (if it is available), which implies a higher level of relatedness among category members for those predicted categories in our classification scheme.

Acknowledgments

I thank W.-C. Liu and C.-C. Chen for preparing the preliminary data as well as M. Zuckermann for stimulating discussions. I am grateful to the Science and Technology Center of the National Science Council of Taiwan for providing the ISI database, and the hospitality at the Department of Physics, Simon Fraser University. This work is supported, in part, by the National Science Council of Taiwan under grant no. NSC 96-2112-M-003-014.

References

Chen, C.-M. Classification of the scientific network using aggregated journal-journal citation relations in the Journal Citation Reports: Supporting information. Retrieved July 4, 2008, from http://www.phy.ntnu.edu.tw/~cchen/paper/cluster.htm

Doreian, P., & Fararo, T.J. (1985). Structural equivalence in a journal network. Journal of the American Society for Information Science, 36, 28–37.

Frey, B.J., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 315, 972–976.

Garfield, E., Malin, M.V., & Small, H. (1975). A system for automatic classification of scientific literature. Journal of the Indian Institute of Science, 57, 61–74.

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

Leydesdorff, L. (1986). The development of frames of references. Scientometrics, 9, 103–125.

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

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, 57, 601–613.

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.

Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity (pp. 190–203). Cherry Hill, NJ: Computer Horizons.

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

Samoylenko, I., Chao, T.-C., Liu, W.-C., & Chen, C.-M. (2006). Visualizing the scientific world and its evolution. Journal of the American Society for Information Science, 57, 1461–1469.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50, 799–813.

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