Classification and Visualization of the Social Science Network by the Minimum Span Clustering Method

Y.F. Chang
Department of Physics, National Taiwan Normal University, Taipei, Taiwan College of Science, China Three Gorges University, YiChang, HuBei 443002, P.R. China. E-mail: yunfengchang82@gmail.com.tw

C.-M. Chen*
Department of Physics, National Taiwan Normal University, Taipei, Taiwan. E-mail: cchen@phy.ntnu.edu.tw

We propose a minimum span clustering (MSC) method for clustering and visualizing complex networks using the interrelationship of network components. To demonstrate this method, it is applied to classify the social science network in terms of aggregated journal-journal citation relations of the Institute of Scientific Information (ISI) Journal Citation Reports. This method of network classification is shown to be efficient, with a processing time that is linear to network size. The classification results provide an in-depth view of the network structure at various scales of resolution. For the social science network, there are 4 resolution scales, including 294 batches of journals at the highest scale, 65 categories of journals at the second, 15 research groups at the third scale, and 3 knowledge domains at the lowest resolution. By comparing the relatedness of journals within clusters, we show that our clustering method gives a better classification of social science journals than ISI's heuristic approach and hierarchical clustering. In combination with the minimum spanning tree approach and multi-dimensional scaling, MSC is also used to investigate the general structure of the network and construct a map of the social science network for visualization.

Introduction

Cluster analysis is the assignment of a set of observations into clusters of components that are similar to each other but different from components in other clusters. It is often used to ascertain whether a complex system comprises a set of distinct clusters, each representing components with substantially different properties. The segmentation of complex systems into clusters might allow us to find specific functions naturally assigned to each cluster, as in the case of human functional brain system (Sporns, Chialvo, Kaiser, & Hilgetag, 2004) and metabolic system (Guimera & Amaral, 2005). Such an approach to statistical data analysis is very useful in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics. A number of clustering methods have been developed as a tool for handling large and heterogeneous collections of systems, e.g., hierarchical clustering (HC; Hastie, Tibshirani, & Friedman, 2008; Manning, Raghavan, & Schütze, 2008; Ward & JR, 1963), k-means clustering (Hastie et al., 2008; MacQueen, 1967; Manning, et al.), and affinity propagation (Frey & Dueck, 2007). However, two crucial drawbacks exist for these clustering methods, including user intrusion and excessive computation time for large data sets. For example, HC recursively merges all components into a cluster and thus requires human intervention to stop the merging process at a specific level of classification. Such a user-defined level of classification, in general, is arbitrary and not related to the characteristic properties of the system under investigation. Moreover, HC involves the calculation of similarity between all pairs of components for each clustering level and is not efficient for handling a large data set.

Several algorithms have been proposed to reduce computation time in cluster analysis by using parallel computation techniques or at the cost of the clustering results (Fernandez & Gomez, 2008; Hastie et al., 2008; Kishida, 2010). However, most clustering algorithms still require pre-given assumptions as inputs for the classification of complex systems, such as the number of clusters, cluster sizes, and boundary conditions. Therefore, it is desirable to develop a convenient and operable clustering algorithm that can efficiently cluster large complex systems.

One interesting application of cluster analysis is the classification of scientific journals. Traditional classification methods are based on subjective analysis, in which case the output could vary from one person to another (Glänzel & Schubert, 2003). Particularly the classification system developed by Institute for Scientific Information (ISI) is worthy

Received May 22, 2011; revised July 16, 2011; accepted July 18, 2011

*Correspondence: C.-M. Chen, Department of Physics, National Taiwan Normal University, Taipei, Taiwan

 DOI: 10.1002/asi.21634

of special attention. The ISI Journal Citation Reports (JCR) presents inter-journal citation frequencies for thousands of journals annually. The aggregated journal-journal citation data in JCR contain extensive information about journal-journal citations, which could enable an in-depth understanding of the interaction among various scientific disciplines. 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. To avoid human intervention, various quantitative methods have been proposed to construct a robust classification system of scientific journals using citation information in JCR. For example, Pudovkin and Garfield (2002) have used the relatedness factor for finding semantically related journals. Leydesdorff and Cozzens (1993) developed an optimization procedure that stabilizes approximated eigenvectors of the scientific network from principal component analysis as representations of clusters. Leydesdorff (2006) further extended the principal component analysis to rotated component analysis, focusing more on specific subsets with internal coherence. Co-citation clustering has also been used to construct a World Atlas of Sciences for ISI (Garfield, Malin, & Small, 1975; L. Leydesdorff, 1987; Small, 1999). Recently, Chen (2008) proposed use of the affinity propagation method to cluster the scientific network. The clustering results of scientific journals are very helpful in understanding mutual interactions among various knowledge domains.

Furthermore, the combination of journal classification and citation mapping can be used to visualize the major features of the scientific network and its time evolution. Previous studies in knowledge domain visualization have unveiled many underlying mechanisms of specific knowledge domains. Price (1965) used citation patterns of documents to depict the topography of current scientific literature. Small and Griffith pioneered the method of mapping the structure of scientific literature by co-citation analysis of the scientific network (Griffith, Small, Stonehill, & Dey, 1974; Small & Griffith, 1974). Narin worked at a more general level by using the citation patterns between journals to define the inter-disciplinary structure of science (Narin, Carpenter, & Berlt, 1972). Journal-journal citations have also been used in scientometric mapping by many authors (Carpenter & Narin, 1973; Doreian & Fararo, 1985; Leydesdorff, 1986; Leydesdorff & Cozzens, 1993; Samoylenko, Chao, Liu, & Chen, 2006; Tijssen, de Leeuw, & van Raan, 1987). Further studies on the visualization of knowledge domains can give a better overall view of human scientific activities and allow the feasibility to predict future trends in the scientific world. A comprehensive review of knowledge domain visualization is available in a recent article by Börner and others (Börner, Chen, & Boyack, 2003).

In this article, we propose a minimum span clustering (MSC) method to cluster complex networks based on the interrelationships among network components. Without inputting the number of clusters or their size, this method is able to cluster complex systems efficiently by merging components of the closest proximity into the same cluster. Moreover, during the merging process of MSC, characteristic numbers of clusters are obtained for the complex system, which correspond to various resolution scales for viewing the system. Here, we applied MSC to classify journals in the Social Sciences Citation Index (SSCI) using citation information in JCR. The structure of these journal clusters was further investigated by minimum spanning tree (MST) approach (Cormen, Leiserson, Rivest, & Stein, 2001; Gower & Ross, 1969; Paivinen, 2005). The exact mapping of these clusters on a two-dimensional plane was implemented by classical scaling (CS; Borg & Groenen, 1997) according to the accumulated distance matrix of clusters. The coordinates of clusters on the two-dimensional map were further optimized by Sammon mapping (SM; Borg & Groenen). Such an approach using various multidimensional scaling (MDS) techniques to construct a sensible map of the scientific world has been previously shown to be very successful (Samoylenko et al., 2006).

Visualizing the Scientific Network

The scientific knowledge of human beings is a complex network, which is traditionally divided into many knowledge domains, such as physics, biology, and psychology. Based on JCR, the visualization of this complicated scientific network can therefore be implemented by first clustering the network into domains using MSC, delineating the general structure of these domains using MST, and constructing a map to visualize the relationship among various domains using MDS. In general, the property of a scientific journal can be well described by the citation pattern of its articles. Journals in the same knowledge domains usually have similar citation patterns, while journals in different knowledge domains could have very different citation patterns. To begin with, we derive citation matrix $\{N_{ij}\}$ (number of citations of journal j cited by journal i) directly from extracting the year 2005 SSCI database CD. The citation pattern of a journal (a citation frequency vector f_i, where i runs over all journals in consideration) is calculated by counting its citation frequency of various journals in the dataset. The collection of all citation frequency vectors thus forms a citation matrix $\{N_{ij}\}$. The similarity in citation patterns of two journals i and j is defined as the cosine measure

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

where $c_{ik} \equiv N_{ik}/\left(\sum_{j\in\Omega} N_{ij}\right)$ is the normalized citation matrix element. Although the number of published papers and citations may reflect individual differences of journals, these quantities are not directly related to the research focus of journals. Therefore, effects from number of journal citations are removed from our investigation by the definition of similarity in Equation 1 and only citation patterns are important

DOI: 10.1002/asi

TABLE 1. Shortest distance pairs of network components listed in increasing order for demonstration purposes.

Journal i	3	4	2	1	8	6	7	10	9	5
Journal j	4	3	3	8	1	7	6	3	4	8
d_{ij}	0.078	0.078	0.151	0.491	0.491	0.682	0.682	0.901	0.968	0.980

for our results. Depending on the similarity in citation patterns of journals i and j, the value of s_{ij} ranges from 0 to 1. For mapping or visualization, the similarity in citation patterns of journals is converted into their distance such that closely related journals are short distances apart and remotely related journals are long distances apart. We express the distance matrix as

$$d_{ij} = 1 - s_{ij} \tag{2}$$

To classify the scientific network, we applied MSC to decompose the SSCI journals into several knowledge domains. The procedure of MSC is explained in the following three steps using a simple example comprising 10 components:

Step 1: We identified the closest neighbor of each journal and recorded their distances in a list ascendingly, as shown in Table 1. For a network of N components, MSC needs only to process N elements in this list instead of dealing with a distance matrix of N^2 elements.

Step 2: The knowledge clusters were constructed by starting from the journal pair with the shortest distance and then including more journal pairs from the list in the order of increasing distance. In this example, to begin with, journals 3 and 4 formed the first cluster. This cluster grew by including journal 2 through its connection with journal 3. Then journals 1 and 8 formed the second cluster, while journals 6 and 7 formed the third cluster. Journals 10 and 9 were added to the first cluster through their connection to journals 3 and 4, respectively. Finally, journal 5 was added to the second cluster through its connection to journal 8.

Step 3: For clusters constructed in step 2, they are considered as renormalized components and their accumulated citation matrix is calculated for all journals in each cluster. In other words, we define the renormalized citation frequency vector as $f_i^R = \sum_{k \in \text{cluster } i} f_k$. If the network is decomposed into M clusters, then the collection of these renormalized citation frequency vectors forms an $M \times N$ matrix. The similarity and distance between clusters can be calculated by Equations 1 and 2. The scientific network comprising these renormalized components will be further classified by steps 1 and 2.

The structure of the scientific network can be visualized by the construction of MST using the Kruskal algorithm (Cormen et al., 2001). MST is a spanning tree for which the sum of distances among network components is the smallest. The Kruskal algorithm constructs the MST by connecting components in the order of increasing distance but avoiding those connections that form loops in the network graph.

Although MST delineates the general structure of the scientific network, specific details for those unconnected components in a MST are not preserved. To construct a two-dimensional map of the scientific network, we apply CS to transform the high-dimensional structure of the scientific network to its low-dimensional representation. CS uses the principal component analysis to calculate eigenvalues and eigenvectors of the distance matrix, from which a two-dimensional projection of the scientific network onto the two principal axes is obtained. In other words, the CS method is used to find a set of two-dimensional vectors $\{\mathbf{x}^m\}$, such that the squared distance matrix between the $\{\mathbf{x}^m\}$ points matches $\{d_{ij}^2\}$ in Equation 2 as closely as possible. The quality of this two-dimensional representation of the scientific network in general depends on the distribution of eigenvalues, particularly the largest two eigenvalues. To minimize possible distortion of network structure, the coordinates of network components in the two-dimensional map are further optimized by Sammon mapping, which tries to preserve the structure of inter-component distances in high-dimensional space on the lower dimension projection. Specifically, the optimization is implemented by minimizing the Sammon stress

$$E = \frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{[d_{ij} - d\left(x^i, x^j\right)]^2}{d_{ij}}, \tag{3}$$

where the summation runs over the dataset under investigation, and $d(a, b)$ is the distance between points a and b.

Results

To demonstrate the applicability of the MSC method in clustering complex systems, we consider the classification of 1,575 journals in the SSCI at various resolution scales of MSC. At each scale, we then use MST to view the general structure of the network, and use MDS to construct its two-dimensional map for visualization. The citation data analyzed in this work are extracted from the CD version of the 2005 SSCI dataset. There are 1,583 journals in this dataset, of which 1,578 journals have nonzero contents. In total, the 2005 SSCI dataset contains 66,051 articles and 2,437,389 citations. The SSCI dataset is decomposed into 55 categories in the ISI classification scheme, which is available at our web page (http://phy.ntnu.edu.tw/∼cchen/paper/msc.htm). Our analysis of the SSCI network is mainly based on 1,575 journals, and then we add another three journals to check the robustness of our method when the number of journals increases.

The implementation of MSC does not require number of clusters or cluster size as pre-given inputs. By repeatedly going through the three steps described in the previous section, classification of the network can be obtained at various resolution scales. The grouping of clusters in the second to fourth runs of MSC is given in Table 2, while a complete

DOI: 10.1002/asi

TABLE 2. Clustering of the SSCI network from 2nd to 4th runs of MSC.

2nd run (65 categories)	3rd run (15 groups)	4th run (3 domains)
6 Psychiatry (25) 10 Psychopathology (40)	1 Psychiatry (65)	1 (984)
17 Drug, alcohol & health (7) 18 Substance abuse (16)	3 Substance abuse (23)	
3 Psychology, applied (23) 7 Psychology, multidisciplinary (113) 19 Psychophysiology (22) 21 Group management (7) 25 Management (40) 32 Technology management (7) 45 Psychology, mathematical (11) 59 Information science & library science (32)	4 Psychology & Management (255)	
24 Psychology, clinical (39) 26 Psychology, behavior assessment (12) 37 Psychology, psychoanalysis (11)	5 Psychology, clinical (62)	
27 Nursing (27) 29 Health (53) 31 Gerontology (34) 34 Public health (29) 41 Health policy & services (14) 49 Ergonomics (13)	6 Health & nursing (170)	
12 Business (30) 23 Information management (10) 42 Operations management (9)	7 Business (49)	
2 Psychology, developmental (67) 28 Linguistic rehabilitation (10) 46 Social work (10) 47 Family studies (22) 57 Developmental disabilities (10)	10 Psychology, developmental (119)	
9 Psychology, experimental (54) 39 Neuropsychology (14) 58 Psychology, biological (9) 62 Linguistics (19)	13 Psychology, biological (96)	
40 Psychology, educational (54) 44 Literacy education (8) 53 Education, general (18) 54 Women's studies & education (17)	14 Psychology, educational (97)	
20 Anthropological archaeology (13) 55 Anthropology (29) 61 Culture & society (6)	15 Anthropology (48)	
4 Law (66) 5 Economics (119) 8 Financial economics (10) 11 Econometrics (22) 15 Theoretical economics (11) 16 Industrial economics (8) 33 Urban, regional, & transportation studies (22) 52 Medical history (5) 56 History of social science (8) 63 History (10) 65 History & philosophy of science (16)	2 Economics, law, & history (297)	2 (321)
1 Finance (19) 22 Accounting (5)	11 Finance & accounting (24)	
14 Demography (9) 36 Sociology (49) 43 Religion (6) 48 Criminology & penology (27) 60 Social administration (10) 64 Public administration (9)	8 Sociology (110)	3 (270)

(*Continued*)

DOI: 10.1002/asi

TABLE 2. (*Continued*)

2nd run (65 categories)	3rd run (15 groups)	4th run (3 domains)
35 Geography (29) 50 Urban studies (18)	9 Geography (47)	
13 Political science (58) 30 Area studies (11) 38 International relations (19) 51 International policy (25)	12 Politics (113)	

Note. SSCI = Social Sciences Citation Index; MSC = minimum span clustering.
The number in parenthesis shows the number of journals of each category/group/domain.

TABLE 3. Research groups and their associated properties in the SSCI network.

	Title	N_j	Representative Journal	$\bar{S}_{j-j}$
1	Psychiatry	65	American Journal of Psychiatry	0.63648
2	Economics, Law & History	297	American Economic Review	0.336661
3	Substance abuse	23	Addiction	0.619084
4	Psychology & Management	255	Journal of Personality and Social Psychology	0.288527
5	Psychology, clinical	62	Psychological Bulletin	0.240879
6	Health & Nursing	170	Social Science & Medicine	0.269675
7	Business	49	Journal of Marketing	0.422846
8	Sociology	110	American Sociological Review	0.380592
9	Geography	47	Urban Studies	0.301904
10	Psychology, developmental	119	Child Development	0.358018
11	Finance & Accounting	24	Journal of Finance	0.702856
12	Politics	113	American Political Science Review	0.333083
13	Psychology, biological	96	Psychological Review	0.331181
14	Education	97	Journal of Educational Psychology	0.232264
15	Anthropology	48	International Journal of Osteoarchaeology	0.174875

Note. SSCI = Social Sciences Citation Index.

list at various resolution scales is available on our web page (http://phy.ntnu.edu.tw/∼cchen/paper/msc.htm). For the network of SSCI journals, 294 batches of journals emerge from the first run of MSC. These batches are mostly specific subfields of social sciences. For example, identified subfields related to gerontology include: "Nursing, gerontology," "Psychology & aging," "Aging & health," "Aging & society," and "Death studies." Identified subfields related to finance include "Finance," "Corporate finance," "Financial markets," "Real estate finance & economics," and "Mathematical finance." Each batch of journals is named by the most popular words that appear in the title and scope of its members.

In the second run of MSC, these 294 batches are further grouped into 65 categories, which closely resemble the 55 subject categories of SSCI journals in ISI's classification scheme. For example, the above five finance related subfields merge to form the "Finance" category, and 18 out of 19 journals in this category also belong to the category "Business, finance" in the ISI's scheme. The third run of MSC merges 65 categories to form 15 research groups. An example is the mergence of categories "Finance" and "Accounting" to form a group of "Finance & Accounting." Table 3 lists their statistical properties, including number of journals (N_j), representative journal, and average intra-cluster journal-journal similarity ($\bar{S}_{j-j}$). The value of $\bar{S}_{j-j}$ shows the relatedness of journals within a cluster, which is greater than 0.5 for "Psychiatry," "Substance abuse," and "Finance & Accounting." These three groups are relatively small in their size and more specialized in their research scope. The fourth run of MSC further clusters the 15 groups into three knowledge domains. Domain I includes Psychology and related research groups, such as "Psychiatry," "Health & nursing," and "Business." Domain II comprises "Economics, Law & History" and "Finance & Accounting." Domain III is composed of "Sociology," "Politics," and "Geography."

In Figure 1, we show the cumulative probability distribution (CPD) of the intra-cluster journal-journal similarity (S_{j-j}) at various resolution scales (A) and for various clustering methods (B). Here the curve labeled "unclustered" is the CPD for the unclustered SSCI network, the curves labeled "MSCX" are the CPD for the MSC clustered network with X clusters, and the curve labeled "closest pairs" is the CPD for the 1575 closest journal pairs. From Figure 1(A), it is found that 90% of journal pairs have a similarity less than 0.1 for the unclustered network, while the CPD is 0.5 for journal pairs

DOI: 10.1002/asi

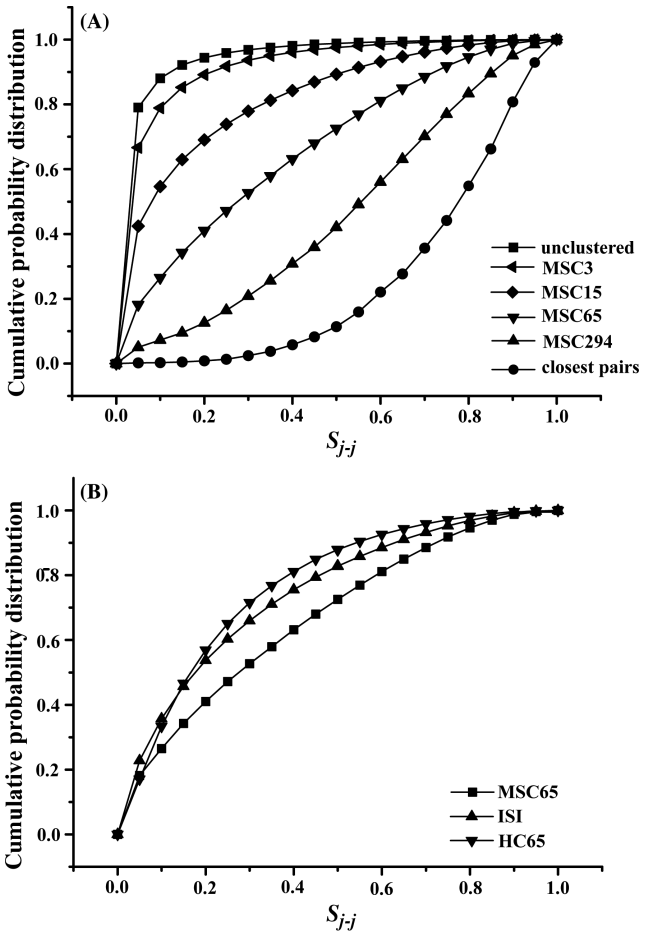

FIG. 1. Cumulative probability distribution (CPD) of the intra-cluster journal-journal similarity at various resolution scales (A) and for various clustering methods (B).

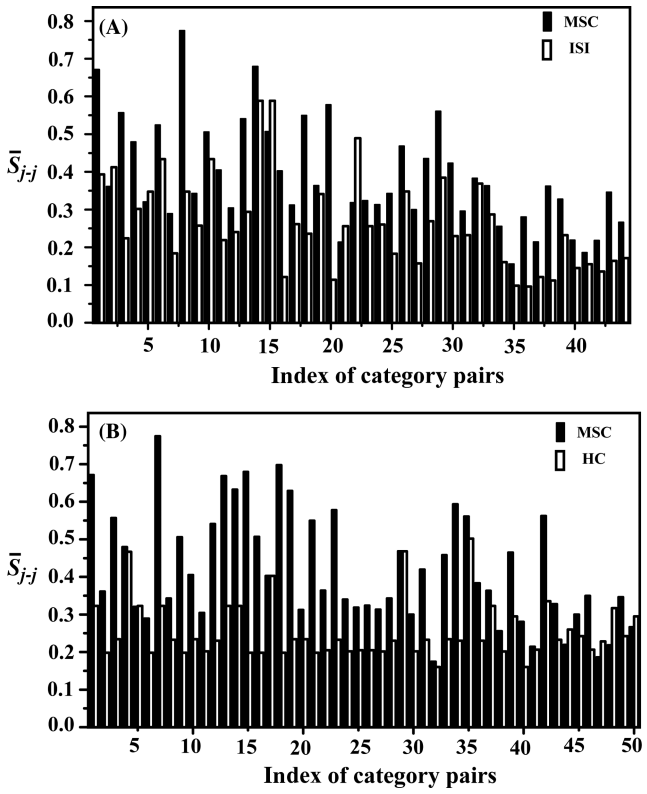

FIG. 2. Comparison of relatedness of journals within a category for (A) MSC and ISI and (B) MSC and HC.

with $S_{j-j} > 0.8$ in the 1,575 closest journal pairs. For MSC clustered network with X clusters, average cluster similarity drops significantly as X decreases because less related journals are grouped in the same cluster. The CPD has a value of 0.5 for journal pairs with $S_{j-j} < 0.55$ in the case of $X = 294$, for journal pairs with $S_{j-j} < 0.3$ in the case of $X = 65$, and for journal pairs with $S_{j-j} < 0.1$ in the case of X=15. In Figure 1(B), we compare the calculated CPDs using four different clustering methods, including MSC, HC, and ISI's heuristic approach. There are 55 clusters in ISI's classification scheme, but 65 clusters in the other two methods. The comparison in Figure 1(B) concludes that, for the relatedness of journals within the same cluster, MSC > ISI $\geq$ HC.

Further comparison was made to the classifications of MSC and ISI of SSCI journals. Among the 65 categories identified in the second run of MSC, 44 clusters have sufficient resemblance ($R > 0.5$) to SSCI subject categories of ISI's classification. The resemblance of two categories is defined as $R = \frac{r_o + r_s}{2}$, where r_o and r_s are the percentage of overlapping journals and the relative size of the two categories. For two identical categories, the value of resemblance is $R = 1$. In Fig. 2(A), we show the comparison of relatedness of journals ($\bar{S}_{j-j}$) between MSC and ISI for the 44 pairs of categories with $R > 0.5$. Detailed information about the title and contents of these 44 pairs of categories can be found at our website (http://phy.ntnu.edu.tw/∼cchen/paper/msc.htm). In 39 out of 44 pairs of categories, the relatedness of journals of MSC categories is significantly larger than that of ISI categories. However, as shown in the category pair 22 of Fig. 2(A), the value of $\bar{S}_{j-j}$ in the category "Gerontology" of MSC classification is considerably smaller than that of its counterpart in ISI classification. The members of this category include "Nursing, gerontology," "Psychology & aging," "Aging & health," "Aging & society," "Rehabilitation & assistive technology," and "Occupational rehabilitation." Among these members, the last two are less related to gerontology. Their connection to other members in this category is through the journal "Assistive technology," which accidentally cited "Journal of rehabilitation research and development" 14 times (out of its 64 citations) in year 2005. From our analysis, journals with very few citations tend to change the local structure, but not the general structure of our classification. We note that multiassigned journals (belonging to more than one category) are allowed in ISI's classification scheme, but not in our classification scheme. Including multiassigned journals in our scheme will slightly decrease the relatedness of journals in the categories of our classifications, because these journals are also included in slightly less favored categories. In Figure 2(B), we show the comparison of relatedness

DOI: 10.1002/asi

of journals ($\bar{S}_{j-j}$) between MSC and HC for the 50 pairs of categories with $R > 0.5$. In 43 out of 50 pairs of categories, the relatedness of journals of MSC categories is significantly larger than that of HC categories. In only four pairs of categories, the relatedness of journals of MSC categories is smaller than that of HC categories.

We further examine the effects of journal insertion on the classification results of MSC. In the 2005 SSCI dataset, 1,578 journals have nonzero contents. The average number of citations for each journal is 1,544. In the above analysis, we have removed three journals from the dataset because they have very few citations ("Feminist Review" 40 citations; "Futures" 322 citations; "New Left Review" seven citations). By inspecting the clustering results of MSC and the MST structure of the network at various resolution scales, we found that the insertion of these three journals does not change the general structure of the SSCI network. However, local structures of the network are affected. For example, due to the insertion of "New Left Review," the closest journal of "Third World Quarterly" (a member of batch 288) becomes "Review of International Political Economy" (a member of batch 137), which was "Journal of Palestine Studies" (a member of batch 288) before insertion. This change provides a link to merge batches 137 and 288 to form a new batch (137) of International Studies in the first run of MSC. The relatedness of journals is 0.497 for batch 137 and 0.510 for batch 288 in the dataset of 1,575 journals, while it is 0.393 for batch 137 in the dataset of 1,578 journals. Similar local changes have also been observed in the second to fourth runs of MSC. In general, the classification scheme of MSC is slightly improved by removing journals with very few citations from the dataset.

So far, we have demonstrated the efficiency and superiority of MSC in clustering complex networks, such as the SSCI network. We further delineate the general structure of the SSCI network using MST. As shown in Figure 3, we construct the MST of the 65 subject categories identified in the second run of MSC (A) and the MST of the 15 research groups in the third run of MSC (B). The mergence of these 65 categories to form 15 groups in the third run of MSC and the mergence of 15 groups to form three knowledge domains are shown by solid circles in Figure 3. In general, it is rather satisfactory to see the mergence of neighboring categories (groups) to form a group (domain). It is observed from the MST that the group "Economics, Law & History" is connected to "Psychology & Management" through the pair of categories "Urban, regional & transportation studies" and "Technology management," while the group "Sociology" is connected to "Psychology, developmental" through categories "Sociology" and "Family studies." Moreover, we have also observed that, on the lower right of MST, there is a medical zone including: "Psychiatry," "Substance abuse," "Health & Nursing" and "Psychology, clinical." However, the group of "Geography" is found to be adjacent to "Economics, Law & History" in MST through categories 50 (Urban studies) and 33 (Urban, regional & transportation studies). This is inconsistent with our results from MSC, in which groups "Geography," "Sociology," and "Politics" form a knowledge domain. This inconsistency is related to the instability of MST against node deletion or insertion (Samoylenko et al., 2006) since MST only conserves the closest pairs in the network. A more robust and sensible MST is the MST of research groups as shown in Figure 3(B), which can be constructed by calculating the accumulated distance matrix among groups. In Figure 3(B), the formation of 3 knowledge domains is consistent with the results of MSC.

To visualize the world of social sciences, we further apply MDS to construct a two-dimensional map of the SSCI network comprising 15 knowledge groups. CS implemented the exact mapping of these knowledge groups on a two-dimensional plane (X_1, X_2) according to their accumulated distance matrix, followed by optimization using SM. As shown in Figure 4(A), the location of each group (represented by an open circle) in the SSCI network is projected on the two-dimensional map. The radius of a circle is proportional to the number of journals (N_j) in a group. The MST structure of these groups is shown by solid lines. Clearly the map of SSCI network comprises three knowledge domains, including a Psychology related domain on the left (I), a domain of Economics & Law on the lower right (II), and a domain of Politics & Sociology on the upper right (III). This map is consistent with the classification of MSC (groups enclosed by dashed lines) and the general structure of MST. In this map, we find that linkages between domains are through large groups, because of the fact that small groups are usually more specialized and less likely to be located at domain boundaries. For example, the linkage between domains I and III is "Psychology & Management" (255 journals) and "Sociology" (110 journals), and the linkage between domains II and III is "Economics, Law & History" (297 journals) and "Politics" (113 journals). The relationship between "Economics, Law & History" and "Politics" is not obvious. Analysis shows that the connection between these two groups is mainly through the categories "Economics" and "Political science," which is because of strong links among batches "Economics & sociology," "Economics & law," "Tax & public economics," and "Political science I." Similar analysis reveals that the connection between groups "Psychology & Management" and "Sociology" is mainly through categories "Psychology, multidisciplinary," "Management," and "Sociology." For the medical zone discussed earlier, we find that the group "Psychology, clinical" is at the central position of the zone. Furthermore, there seems to be a strong connection between "Psychiatry" and "Substance abuse," which is invisible from MSTs in Figure 3. Note that the map in Figure 4(A) presents only a panorama for the SSCI network to avoid confusing readers with a lot of detailed information.

Mapping of 1,575 journals or 264 batches on a two-dimensional plane by MDS is not realistic because of the large distortion of the relative distance between journals or batches. Specific details of the SSCI network can be investigated by choosing a knowledge area on the map and exploring it with a desired resolution. For example, as shown in Figure 4(B), we explore the domain III in Figure 4(A) with a higher

DOI: 10.1002/asi

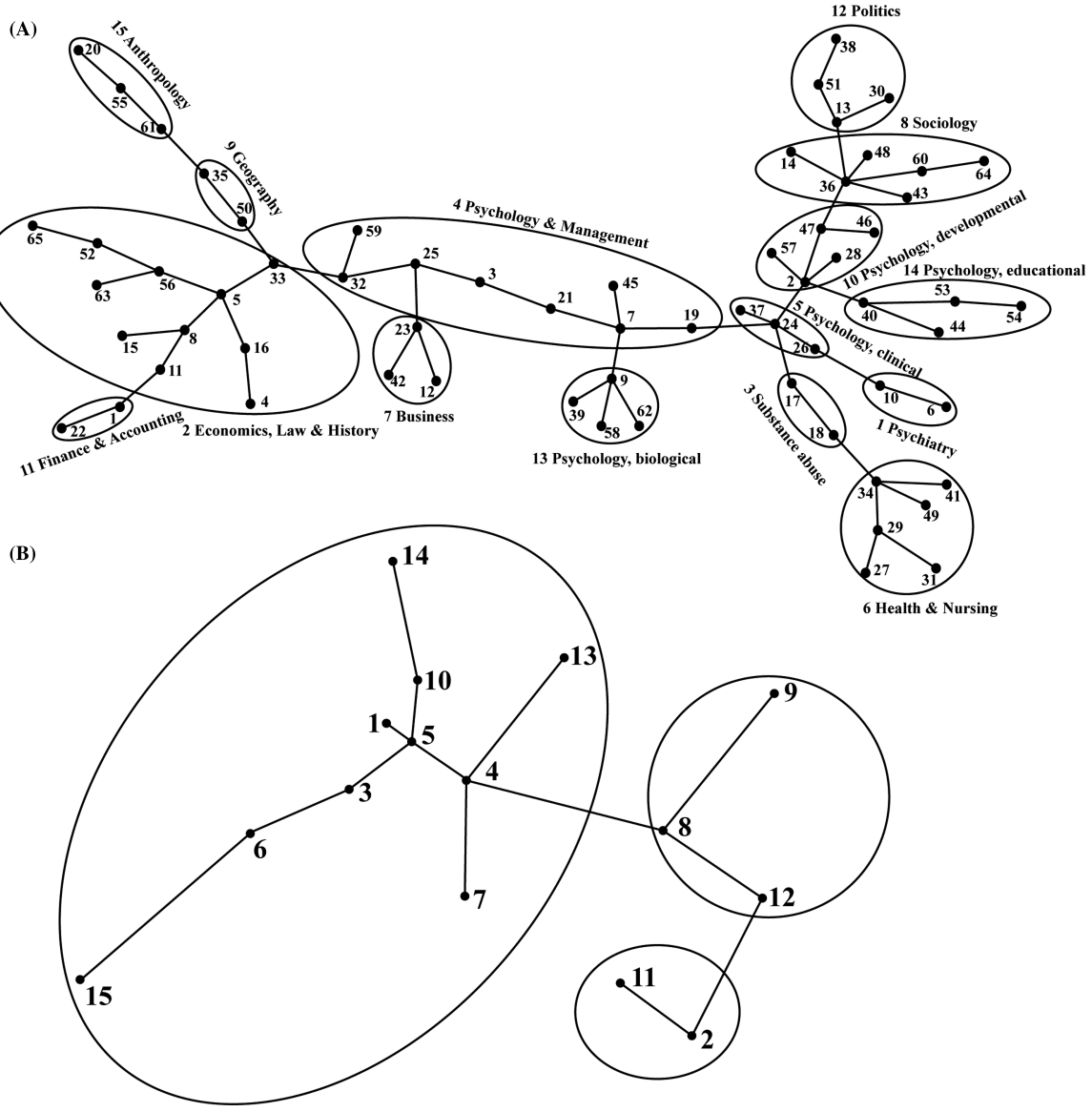

FIG. 3. An MST of 65 categories (A) and an MST of 15 groups (B) for the SSCI network. Each category (group) is represented by a circle and labeled by a number as in Table 2. The line length between two categories (groups) is proportional to their distance. Categories in the same group are enclosed by a solid line.

resolution. By doing so, readers soon discover that the group "Sociology" contains six subject categories, group "Geography" contains two categories, and group "Politics" contains four categories. Readers will also find that the contents of groups 9 and 12 are more localized, while the contents of group 8 are spread out. For group 8, categories 14, 36, 43, and 48 form a head group, while the other two categories 60 ("Social administration") and 54 ("Public administration") form a tail region. The relative positions of groups 8, 9, and 12 in Figure 4(B) are consistent with their positions in Figure 4(A).

Conclusion

In summary, we have proposed a MSC method to cluster large complex systems. This clustering method is efficient since its processing time is linear to the size of the dataset (N), and its results are satisfactory. For the SSCI network, there exist four scales of resolution in processing MSC, and each of them provides a meaningful perspective of the network. At the highest resolution (first run), the network can be clustered into 294 batches of journals, which delineate various subfields in social science. At a slightly lower resolution (second run), the network comprises 65 categories. Because these 65 categories are comparable to the 55 categories of SSCI journals in ISI's classification, we have found that the value of ($\bar{S}_{j-j}$) (relatedness of journals within a category) of our categories is significantly larger than their counterpart in the ISI scheme. At an even lower resolution (third run), the network contains 15 research groups. These 15 groups coalesce to form three knowledge domains at the lowest resolution (fourth run). A two-dimensional map of these research

DOI: 10.1002/asi

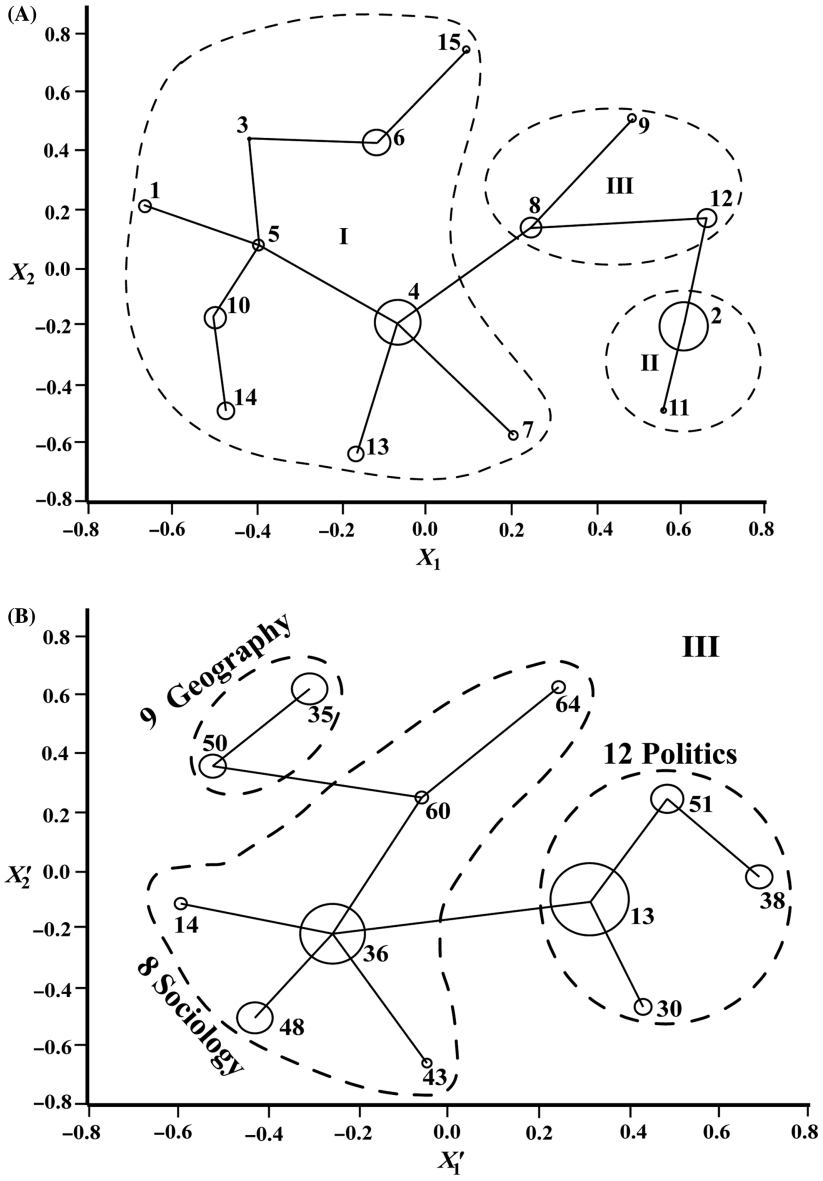

FIG. 4. A two-dimensional map of the SSCI network at the lowest resolution scale (A) and an enlargement of domain III at the third scale (B). Solid lines show the MST structure of 15 groups in (A) and the MST of 12 categories in (B). Solid circles indicate the number of journals in each group (A) or category (B). Clusters enclosed by dashed lines are results of MSC.

groups and knowledge domains has been constructed using MDS and MST to visualize the general structure of the SSCI network and the interaction among various groups. Although the two-dimensional map shows only a panorama of the network, specific detailed information of the network can be investigated by exploring various knowledge areas with a desired resolution. Because of the efficiency and superiority of MSC in clustering complex systems, this method can also be applied to investigate various other complex networks, such as the genome network, proteome network, corporate networks, and merchandising networks.

Acknowledgment

The authors are grateful to the Science and Technology Center of the National Science Council of Taiwan for providing the ISI database. This work is supported, in part, by the National Science Council of Taiwan under grant no. NSC 99-2112-M-003-011-MY3.

References

Borg, I., & Groenen, P. (1997). Modern multidimensional scaling: Theory and applications. New York: Springer.

DOI: 10.1002/asi

Börner, K., Chen, C.M., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37, 179–255.

Carpenter, M.P., & Narin, F. (1973). Clustering of scientific journals. Journal of the American Society for Information Science and Technology, 24(6), 425–436.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., & Stein, C. (2001). Introduction to Algorithms (2nd ed.). Cambridge, MA: MIT Press.

Doreian, P., & Fararo, T.J. (1985). Structural equivalence in a journal network. Journal of the American Society for Information Science and Technology, 36(1), 28–37.

Fernandez, A., & Gomez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. Journal of Classification, 25(1), 43–65.

Frey, B.J., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 315(5814), 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(3), 357–367.

Gower, J., & Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. Applied Statistics, 18, 54–64.

Griffith, B.C., Small, H.G., Stonehill, J.A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro- and microstructure of science. Science Studies, 4, 339–365.

Guimera, R., & Amaral, L.A.N. (2005). Functional cartography of complex metabolic networks. Nature, 433, 895–900.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). The Elements of Statistical Learning (2nd ed.). New York: Springer-Verlag.

Kishida, K. (2010). High-speed rough clustering for very large document collections. Journal of the American Society for Information Science and Technology, 61(6), 1092–1104.

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., & 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.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (pp. 281–297). Berkeley. California: University of California Press.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.

Narin, F., Carpenter, M., & Berlt, N C. (1972). Interrelationships of scientific journals. Journal of the American Society for Information Science and Technology, 23, 323–331.

Paivinen, N. (2005). Clustering with a minimum spanning tree of scale-free-like structure. Pattern Recognition Letters, 26(7), 921–930.

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 and Technology, 57(11), 1461–1469.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science, 50(9), 799–813.

Small, H.G., & Griffith, B.C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. Science Studies, 4, 17–40.

Sporns, O., Chialvo, D.R., Kaiser, M., & Hilgetag, C.C. (2004). Organization, development and function of complex brain networks. Trends in Cognitive Sciences, 8(9), 418–425.

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

Ward, J.H., Jr. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58(301), 236–244.

DOI: 10.1002/asi