

Global Maps of Science Based on Paper-level Classifications of Science

Jianghua Ma

School of Social Science, Soochow University, Suzhou Jiangsu 215123, China

Abstract: Science mapping (or mapping knowledge domains) has been a hot topic in scientometrics field since 1990s. Science maps are generally created by analyzing different kinds of scientific literatures, including journal articles, conference proceedings, patents, books, and monographs. Citation analysis (eg. direct citation, co-citation, bibliographic coupling and hybrid approach) with units in different granularity (eg. authors, keywords, articles, journals and categories) is the most popular method to create maps of science. In this study, paper-level classification codes of science based on Chinese Library Classification Scheme (CLC) are used to draw global maps of science. Paper-level classification codes of 3,490,665 academic articles published from 2008 to 2018 in 1,687 core Chinese journals indexed by CSSCI (Chinese Social Sciences Citation Index) and CSCD (Chinese Science Citation Database) are harvested and three kinds of approaches (publishing, co-assignments and citing networks with various of interdisciplinary measures) are applied to construct and analyze global maps of science. The novel CLC-based science maps reveal new relations among different disciplines of science in Chinese knowledge organization system, which offers a new way to comprehensively understand the landscape of scientific domains and how they interact.

Keywords: Science Mapping; Mapping Knowledge Domains; Interdisciplinarity; Classifications of Science; Chinese Library Classification Scheme.

1. Introduction

In scientometrics field, citation analysis (eg. direct citation, co-citation and bibliographic coupling) is the most popular method to create maps of science from academic journal articles and their citations (Shu et al. 2017). Meantime, different maps of science are generated from units in different granularity (eg. authors, keywords, articles, journals and categories). For most category-based maps of science (especially global maps of sciences), they rely on journal-level classifications of science (Leydesdorff & Rafols 2009, 2012; Richardson 2012; Leydesdorff et al. 2013).

Different from international journal-level classification system (eg. Web of Science and Scopus Categories), Chinese bibliographic databases (eg. CNKI and Wanfang Data) classify publications at the paper level using the CLC (National Library of China 2010). Also, most Chinese bibliographic databases (eg. CNKI, Wanfang Data and CSCD) classify journals mainly using the CLC. Besides, CLC is used by all the publishers and libraries in China to classify all publications including journals, books, and monographs (Chen et al. 2019).

Generally speaking, assigning journal articles with paper-level classifications of science is more accurate than journal-level classifications of science. So, science maps created by paper-level classifications of science is probably more accurate than journal-level classifications of science. As the most widespread knowledge organization in China, CLC has never been used to map global science. This study offers a new way to map knowledge and presents a global map of science generated from CLC assigned to articles and journals and investigates differences with previous global maps of science.

2. Data and Method

2.1. Data

Table 1. Sample journal statistics.

Index	CLC Code	# of Journals	# of Papers
CSSCI	Tier-1	447 (78.7%)	720,482 (72.7%)
	Tier-2	121 (21.3%)	270,552 (27.3%)
CSCD	Tier-1	708 (57.6%)	1,571,781 (62.2%)
	Tier-2	411 (33.4%)	955,719 (37.8%)

Table 2. Sample paper statistics.

CLC Code	# of Papers	# of Papers (# of CLC Codes >= 2)
Tier-1	3,490,665 (100%)	297,664 (8.5%)
Tier-2	3,428,575 (98.2%)	411,413 (11.7%)
Tier-3	3,401,177 (97.4%)	472,025 (13.5%)
Tier-4	3,168,248 (90.7%)	440,361 (12.6%)

The sample journals consist of 1,772 core journals (only 1,687 journals can be accessed in this study, detailed in Table 1) indexed by CSSCI (2019-2020 Version) (<http://cssrac.nju.edu.cn/index.html>) and CSCD (2019-2020 Version) (http://sciencechina.cn/cscd_source.jsp). 26 journals are both indexed by CSSCI and CSCD. The sample articles consist of 4,058,005 articles (only 3,490,665 academic articles are kept in this study, detailed in Table 1) published from 2008 to 2018 in 1,687 core Chinese journals indexed by CSSCI and CSCD. The raw data are crawled from March to May, 2020 in CNKI and manipulated from June to November, 2020. Every Chinese journal has a tier-1 or 2 CLC codes and almost each

Chinese article has at least one tier-2, 3, 4, 5 or 6 CLC codes (detailed in Table 2).

2.2. Method

Small summarized that science mapping is difficult to grasp because three conceptual steps are required before it makes sense, two of which involve some degree of mathematical manipulation. First a unit of analysis must be chosen which should comprise the elementary particles of our science universe. Secondly a measure of association between the units must be defined. Thirdly, a means must be found for depicting the units and their relations in a low dimensional space (usually two dimensions) that can be perceived by humans. Once these intellectual leaps are made science mapping seems natural and even inevitable (Chen 2013).

Based on the above theories, three kinds of relations between Chinese journal-level and article-level CLC codes are depicted: publishing, citing and CLC codes assignment or co-assignment (shown in Figure 1). Then, three methods are used to construct networks of CLC codes and interdisciplinary measures (diversity & network coherence) are applied to measure nodes and their relations.

Method 1 (Publishing network): Article-level CLC codes point to journal-level CLC code (perspective of publishing). When any article-level CLC code k in any article A points to journal-level CLC code l of article A 's publishing journal X , their value is number one. The total relation intensity ϕ between CLC code k and l is the sum of CLC code k points to l between any article A and its publishing journal X . The total relation intensity is not normalized.

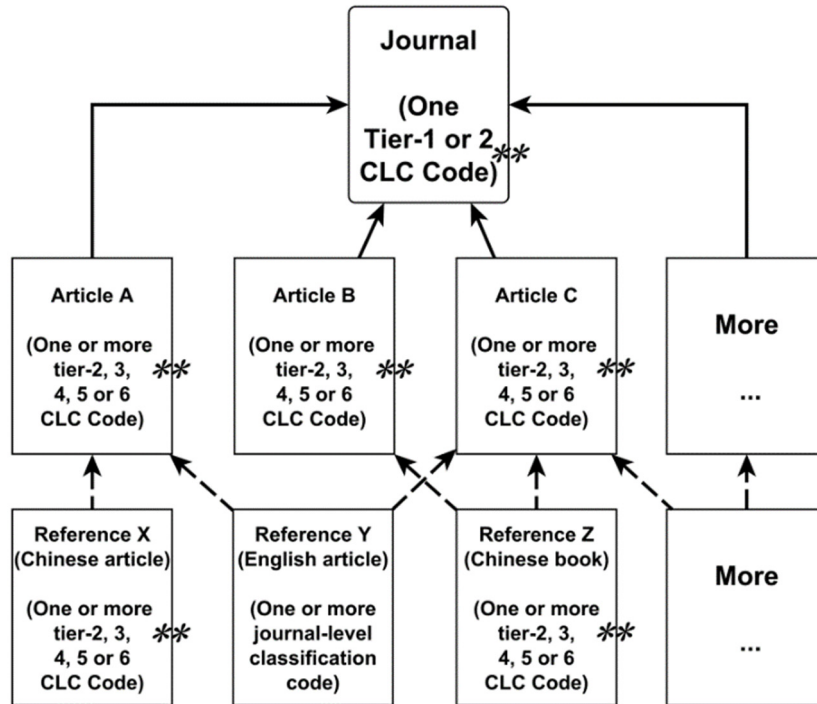


Fig 1. Relations of Chinese journal-level and article-level CLC codes (solid lines mean publishing, stars mean CLC codes assignment or co-assignment, dashed lines mean citing).

3. Results

Due to the limited pages, only publishing network and co-assignment network of tier-1 CLC codes are shown in Figure 2 and 3 (global maps of science based on tier-2, 3 and 4 paper-level classifications of science can also be mapped, which can convey more details than tier-1 paper-level classifications of

$$\phi_{kl} = \sum_{A \text{ and } X} \text{One}(\text{CLC}_k, \text{CLC}_l) \quad (1)$$

Method 2 (Co-assignment network): Co-occurrence of article-level CLC codes (≥ 2) in any article (perspective of CLC codes co-assignment for articles). When any article-level CLC code i and j co-occur in any article B , their co-occurring value is number one. The total relation intensity ϕ between CLC code i and j is the sum of CLC code i and j co-occurring in any article B . The total relation intensity is normalized by Jaccard Index.

$$\phi_{ij} = \sum_B \text{One}(\text{CLC}_i, \text{CLC}_j) \quad (2)$$

Method 3 (Citing network): CLC codes in references point to CLC codes in citing articles (perspective of citing). When any article-level CLC code m in reference Y points to article-level CLC code n in any citing article C , their value is number one. The total relation intensity ϕ between CLC code m and n is the sum of CLC code m points to n in reference Y and its citing article C . The total relation intensity is not normalized.

$$\phi_{mn} = \sum_{C \text{ and } Y} \text{One}(\text{CLC}_m, \text{CLC}_n) \quad (3)$$

Interdisciplinary measures (Diversity & Network coherence): Various of diversity (eg. Brillouin index and Gini index) and network indicators (eg. network density, mean path length and node centrality) are calculated to measure global maps of science created by paper-level and journal-level classifications of science. Besides, tier-4 categories can be upgraded to tier-3, 2 and 1 categories. So, all upgraded categories are also included in each upper-level network.

science). Also, only citing networks of tier-1 and 2 CLC codes in Library, Information and Archival Science (49,554 academic articles with 32,6264 citations) are drawn (shown in Figure 4) because massive citation data are difficult to preprocess and it is still in process.

From Figure 2 to 4, networks are placed by Circular or Force Atlas layout algorithm, each node indicates a category

(labeled by CLC codes, detailed in Table 3), node size indicates its weighted degree, weighted in-degree or weighted out-degree, node label size is proportional to node size, node color indicates its category type (blue nodes indicate Arts &

Humanities, green nodes indicate Social Sciences, red nodes indicate Natural Science), links between nodes indicate their relation intensity (original value or normalized value).

Table 3. Tier-1 categories of CLC codes (normal fonts: Arts & Humanities; bold fonts: Social Sciences; italic fonts: Natural Sciences).

Code	Category	Code	Category
A	Marxism-Leninism, Mao Zedong Thought & Deng Xiaoping Theory	N	General Natural Science
B	Philosophy & Religion	O	Mathematics, Physics & Chemistry
C	General Social Science	P	Astronomy & Earth Science
D	Politics & Law	Q	Biology
E	<i>Military</i>	R	Medical Science
F	Economics	S	Agriculture
G	Culture, Science, Education & Sports	T	Industrial Technology
H	Language & Character	U	Transportation
I	Literature	V	Aeronautics & Astronautics
J	Arts	X	Environmental & Safety Science
K	History & Geography	Z	Comprehensive books

3.1. Publishing Network

In Figure 2, it shows that there are more knowledge interactions within the macro-disciplines (eg. social science, natural science) than outside the macro-discipline. From Figure (b), *T* (Industrial Technology), *R* (Medical Science), *C* (General Social Science), *S* (Agriculture), *O* (Mathematics,

Physics & Chemistry), *P* (Astronomy & Earth Science) and *N* (General Natural Science) absorb the most knowledge from other disciplines. From Figure (c), *T*, *R*, *F* (Economics), *S*, *O*, *P* and *X* (Environmental & Safety Science) contribute the most knowledge to other disciplines. Besides, *E* (Military) rank last in Figure 2(b) because journals in *E* (Military) category are not indexed by CSSCI and CSCD.

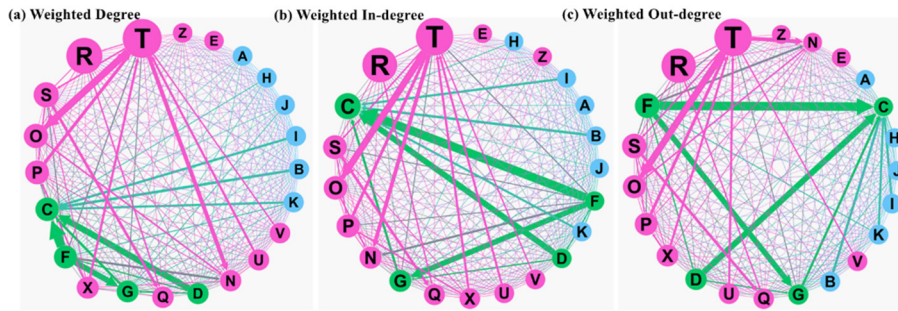


Fig 2. Publishing network of tier-1 CLC codes (node size in Figure (a), (b) and (c) indicates its weighted degree, weighted in-degree and weighted out-degree respectively.)

3.2. Co-assignment Network

In Figure 3(a) and 3(b), co-occurrence value between any two nodes is original or normalized respectively. From Figure 3 and Table 4, the relationships between *T* (Industrial Technology), *O* (Mathematics, Physics & Chemistry), *X*

(Environmental & Safety Science), *R* (Medical Science), *F* (Economics), *D* (Politics & Law) and *G* (Culture, Science, Education & Sports) are some of the strongest as indicated by the thick lines between any two nodes. Also, in Figure 2(b), the relationships between *S* (Agriculture) and *Q* (Biology), *H* (Language & Character) and *I* (Literature) are the strongest.

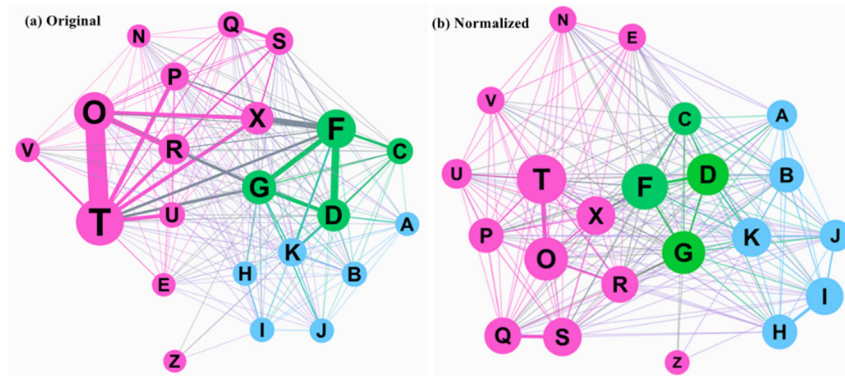


Fig 3. Co-assignment Network of Tier-1 CLC codes.

Table 4. Top10 original or normalized co-occurrence value of co-assignments (reds fonts denote the same node pairs).

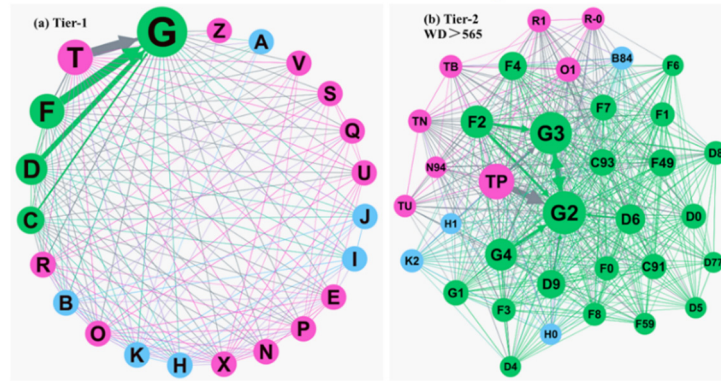
Node 1	Node 2	Original	Node 1	Node 2	Normalized
T	O	24890	T	O	0.315
F	D	9628	S	Q	0.268
F	X	7643	H	I	0.221
O	R	7345	F	D	0.176
F	G	6540	F	X	0.135
T	P	5499	O	R	0.133
O	X	5259	F	G	0.107
G	D	4983	G	D	0.105
T	U	4960	G	R	0.089
T	X	4899	D	K	0.087

3.3. Citing Network of Library, Information and Archival Science

From Figure 4(a) and Table 5, *T* (Industrial Technology), *F* (Economics), *D* (Politics & Law) and *C* (General Social Science) contribute the most knowledge to *G* (Culture, Science, Education & Sports), then *R* (Medical Science), *B* (Philosophy & Religion), *O* (Mathematics, Physics & Chemistry), *K* (History & Geography) and *H* (Language & Character).

From Figure 4(b) and Table 6, *G2* (Information and Knowledge Dissemination) and *G3* (Science & Science Studies), *G2* and *G4* (Education) mutually contribute the most

knowledge to each other. *TP* (Automation & Computer Technology), *F2* (Economic Management), *D6* (Chinese Politics), *D9* (Law), *F49* (Information Industry Economy), *F7* (Trade Economy), *C93* (Management Science) and *F4* (Industrial Economy) contribute the most knowledge to *G2* or *G3*. Also, some sub-disciplines of arts and humanities, *B84* (Psychology), *H1* (Chinese Language), *H0* (Linguistics) and *K2* (Chinese History), contribute the most knowledge to *G2* or *G3*. Apart from *TP*, other sub-disciplines of natural science, *TN* (Radio Electronics & Telecommunications), *O1* (Mathematics), *R1* (Preventive Medicine & Hygienics), *TU* (Architecture Science), *N94* (System Science) and *TB* (General Industrial Technology), contribute the most knowledge to *G2* or *G3*.



4. Discussions and Conclusion

In this study, global maps of science created by paper-level and journal-level classifications of science are constructed in the perspective of publishing, co-assignments and citing. In publishing and co-assignments networks, they reveal in common that *T (Industrial Technology)*, *R (Medical Science)*, *S (Agriculture)*, *O (Mathematics, Physics & Chemistry)*, *P (Astronomy & Earth Science)*, *X (Environmental & Safety Science)* and *Q (Biology)* are strongly related, *F (Economics)*, *D (Politics & Law)* and *G (Culture, Science, Education & Sports)* are also strongly related. Shu et.al (2017) explored that *Physics*, *Engineering*, *Medical Sciences*, *Mathematics*, *Environmental Sciences*, *Biology*, *Chemistry* and *Agriculture* are strongly related, which is very similar to the results in this study.

There are some limitations in this study. Only global maps of science based on tier-1 CLC codes (only one global map of science based on tier-2 CLC codes) are depicted, global maps of tier-2, 3 and 4 CLC codes are needed to be drawn to reveal more details. Also, only about 32% journals are assigned with tier-2 CLC codes in this study, so science map based on tier-2 journal-level CLC codes is not global. Also, only about 12.6% papers are assigned with more than two tier-2 CLC codes in this study, so the reasonableness and accuracy of co-assignments networks is needed to be verified. Although there are some deficiencies in sample data, publishing, co-assignments and citing networks provides us a comprehensive way to explore and understand global maps of science. Besides, as a unified standard in China, paper-level CLC

codes will provide us a unique perspective in science mapping and interdisciplinary research, even promising practices or experiences.

References

- [1] Chen B, Cheng M, Li P, et al.: Interdisciplinary research based on paper-level classifications of science- A preliminary case study of Chinese journals. In: Proceedings of the 17th International Conference on Scientometrics and Informetrics, pp. 2718-2719. ISSI, Rome (2019).
- [2] Chen C.: Mapping scientific frontiers: the quest for knowledge visualization. 2nd edn. Springer, Heidelberg (2013).
- [3] Leydesdorff, L., Carley, S., & Rafols, I.: Global maps of science based on the new Web-of-Science categories. *Scientometrics* 94(2), 589–593 (2013).
- [4] Leydesdorff, L., & Rafols, I.: A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science & Technology* 60(2), 348-362 (2009).
- [5] Leydesdorff, L., & Rafols, I.: Interactive overlays: A new method for generating global journal maps from web-of-science data. *Journal of Informetrics* 6(2), 318–332 (2012).
- [6] National Library of China.: Chinese Library Classification. 5th edn. National Library of China Publishing House, Beijing (2010).
- [7] Richardson M.: Citography: the visualization of nineteen thousand journals through their recent citations[J]. *Research Trends* 26(3), (2012).
- [8] Shu, F., Dinneen, J. D., Asadi, B., & Julien, C.: A. Mapping science using library of congress subject headings. *Journal of Informetrics* 11(4), 1080-1094 (2017).