DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.4, NOV. 2021

Discovering the communication structures of directed weighted citation networks in library and information science through Pathfinder

Ruimin Ma[a] , Rongxu Zhang[b*]

a. School of Economics and Management, Shanxi University, Taiyuan, China

b. School of Government, Beijing Normal University, Beijing, China

ABSTRACT

Communication structures mining is of importance for the understanding of the communities of a domain and knowledge flow among papers and authors. In this paper, we take advantage of Pathfinder, a method for pruning networks, to discover the communities of a directed weighted citation network and its main knowledge flow structure. Meanwhile, in the course of the analysis, necessary data transformations are carried out, and proper parameters for Pathfinder are determined. It is found that Pathfinder plays a multifaceted role in the discovery of communication structures of directed weighted citation networks, which could provide more systematic insights to citation network analytics.

KEYWORDS

Communication structure; Citation network; Knowledge flow; Community; Pathfinder

1 Introduction

Communication forms the foundation of modern science and facilitates knowledge exchange among scholars (Garvey, 1979). Research on the communication structures of knowledge networks has attracted great interests from scholars in a variety of scientific fields such as physics, computer science, sociology, and information science. Broadly perceived, there are two threads in current studies: one is community discovery and the other is knowledge flow identification.

Community discovery using citation networks as a research instrument has benefited from a number of methods, stemming from applied statistics and social network analysis; for instance, clustering analysis, factor analysis, and multidimensional scaling analysis have been employed to reveal the intellectual structure of co-occurrence networks (e.g., Kessler, 1963; McCain, 1990; Small, 1973; White & Griffith, 1981; White & McCain, 1998; Zhao, 2006; Zhao & Strotmann, 2008). In recent years, we have witnessed a growing trend of using advanced methods for community detection, jointly designed in information science, computer science, and sociology, such as self-organizing feature mapping (e.g., Kohonen, 1982; Lin et al., 2003), the Pathfinder algorithm (e.g., Chen, 1998, 2006a, 2006b; White, 2003a), the Gir-

* Corresponding author: 202131260008@mail.bnu.edu.cn

van-Newman algorithm (Girvan & Newman, 2002), and the Louvain algorithm (Blondel et al., 2008). Despite their algorithmic strengths, they are primarily applied to undirected, co-occurrence-based networks such as co-authorship and co-citation networks.

In the meantime, the other thread of research concerning patterns of communication structures is to find knowledge flow trajectories. One pioneering study was conducted by Hummon and Doreian (1989), in which they initiated three algorithms to mine main paths in citation networks, i.e., search path nodes pair (SPNP), node pair projection count (NPPC), and search path link count (SPLC). Batagelj (2003) later proposed another widely-accepted algorithm named search path count (SPC); it has been demonstrated that SPC outperforms other related methods such as SPLC. Liu and Lu (2012) further improved the SPC and proposed four novel algorithms, of which the global key-route, has exhibited its marked performance. Lucio-Arias and Leydesdorff (2008) explored the main paths of a field based on highly cited documents using SPLC by virtue of HistCite developed by Garfield and his colleagues (Garfield et al., 2003). This attempt ensures that every node on the main paths plays a dominant role in the development of a field, which, to a great extent, authentically embodies the essence of being on the main paths. It should be noted that because Lucio-Arias and Leydesdorff's method is based on SPLC which is akin to SPC, some highly cited nodes could be ruled out from networks (De Nooy et al., 2005; Liu & Lu, 2012; Lu & Liu, 2013).

Although there is a body of methods to discover the communication structures of networks from various perspectives, the attribute of networks should be discerned because different methods apply to different types of networks. In general, networks can be classified into four types-directed weighted, directed unweighted, undirected weighted, and undirected unweighted, of which the directed networks can be further subdivided into acyclic and cyclic types. With respect to citation networks, they can be an acyclic unweighted one such as paper citation networks or a cyclic weighted one such as author citation networks. At present, to the best of our knowledge, most path delineation methods are only suitable for undirected networks or acyclic directed networks, and there lacks empirical research on how to use an appropriate method to find knowledge paths in directed weighted networks.

To fill this gap, we aim to employ Pathfinder to address two central issues regarding the use of citation networks to examine knowledge communication structures:

- how to discover the communities of a directed weighted citation network;
- What patterns can be revealed by analyzing the main knowledge flow structure of the specific network.

2 Methods

In this section, the principle of Pathfinder is first introduced; then, how to make Pathfinder applicable to the directed weighted network is described.

Pathfinder is an effective algorithm for pruning directed and undirected networks (Schvaneveldt, 1990). This algorithm is regulated by two parameters: r is to determine the distance between two nodes that are not directly linked; q is a constraining factor that limits the number of links in a path (Schvaneveldt, 1990). The result of a Pathfinder pruning is named PFNET(r, q). Pathfinder produces graphs in which all nodes are connected and retain only salient relations among nodes (White, 2003a; Chen, 2006a). Distances between two nodes are calculated $W(ij)=(\sum^{k}_{m=1} w^{r}_{m})^{1/r}$, in which for r=1, it becomes the sum of the link weights of a path; for r=∞, it is the same as the maximum weight associated with any link of a path (Schvaneveldt, 1990). It has been proved that for a symmetric network (i.e., an undirected weighted network), when r=∞ and q=n-1(n is the total number of nodes), we get a minimum-cost network (MCN), which is essentially the union of all minimal spanning trees (MSTs) of the network (Schvaneveldt, 1990; Chen, 2006b). Additionally, Pathfinder is predicated upon dissimilarity measures, which requires a transformation of similarity-based networks such as citation networks.

The procedures to obtain PFNETs are summarized as follows: (1) calculate the weights of a path w1, w2, ..., wm between nodes i and j with the Minkowski r-metric $W(ij)=(\sum^{k}_{m=1} w^{r}_{m})^{1/r}$,; (2) similarly, calculate the weight of all paths between nodes i and j; (3) the new distance between nodes i and j is calculated using D_{ij} = MIN (W ($P_{i/1}$), (W ($P_{i/2}$), ... , (W ($P_{i/p}$)), where p is the number of paths between i and j; (4) compare wij (the weight of the direct link between i and j) and Dij. If wij >Dij, it violates the triangle inequality and should be omitted; and (5) carry out the iteration for pairwise nodes through steps (1)-(4) until all links that violate triangle inequalities are omitted, and eventually obtain PFNET(r, q). Generally, q is set as n-1, while r is contingent on different settings.

2.1 Community detection of directed weighted citation networks with pathfinder

As far as community detection is concerned, we focus more on the similarity rather than the knowledge flow among nodes. Thus, directed weighted citation networks can be transformed into undirected weighted ones. The number of citations between two nodes indicates the extent to which they are similar: if two nodes mutually frequently cite each other, they have more affinities. We can adopt the sum of the mutual citations between two nodes as the similarity metric (De Nooy et al., 2005; Leicht & Newman, 2008; Yu et al., 2010) so that the raw network is transformed into a symmetric (i.e., undirected weighted) one.

To meet the requirement of Pathfinder, the inverse of the sum is set as the dissimilarity metric between two nodes. According to PFNET(∞, n-1), the transformed network can be pruned to be a more concise one that only keeps salient relations, which is also the union of all possible MSTs (Schvaneveldt, 1990; Chen, 2006b). Additionally, White (2003a) pointed that PFNETs should be generated from raw counts, and therefore we need not carry out any further transformations such as Pearson or Cosine correlation coefficients for networks (Ahlgren et al., 2003; White, 2003b).

2.2 Knowledge flow display of directed weighted citation networks with pathfinder

For directed weighted citation networks, there could be countless paths between two nodes when there is a cycle in the paths so that it is difficult to distinguish which one is the root node. Therefore, we mainly use Pathfinder to reveal the dominant knowledge flow among nodes.

We use citation counts to measure the volume of knowledge flow between two authors (Yan, 2014). If node A cites B five times, there are five units of knowledge flowing from B to A. To satisfy the premise of Pathfinder, the inverse of knowledge flow is adopted to calculate the distance between two authors (another transformation approach, negative values, cannot be used because there are cycles in the networks, which would lead to many isolated nodes). Network trimmings can be accomplished by setting the parameters in Pathfinder r=∞.

In order to make the visualization result more readable, we reorganize the layout of nodes. Here, we classify nodes into three types: nodes with no input degrees are conceived as "sources" and positioned to the left; nodes with no output degrees are considered as "sinks" and placed to the right; and rest nodes with both input and output degrees are the brokers and scattered in the middle.

3 Data

The data were harvested from the Chinese Social Sciences Citation Index (CSSCI), in which there is a category named Library, Information and Archives Science (CLIS). Between 2009 and 2013, in this category, 15,401 authors have published 20,436 papers and cited 167,649 references containing 78,814 cited authors. Note that only the first authors of citing papers and references were taken into account because CSSCI indexes first authors only for publications. Ninety-three authors with citations over 60 within this time frame were selected. Four isolated authors and five authors whose maximum similarity with others lower than 2 were then precluded, thus resulting in 84 authors. Then an 84*84 author citation matrix was built. At last, the data of two subfields of CLIS-informetrics and competitive intelligence-were extracted from the 84*84 matrix.

4 Results

4.1 Discovering communities of directed weighted citation networks

Author citation networks are a representative of directed weighted citation networks. Figure 1 shows the communities of CLIS based on an author citation network with Pathfinder. In Figure 1, the size of a node indicates an author's citation count and the thickness of a line represents the similarity between two authors.

There are two distinct components within CLIS: library science to the right of the middle dashed line and information science located to the left. The library science component is densely connected and cannot be further separated into smaller groups. With respect to information science, there are three subgroups: knowledge organization retrieval, informetrics, and competitive intelligence. In addition, we observe that some individuals play pivotal roles in their respective groups. For example, J.P. Qiu is a renowned information scientist who specializes in informetrics, theories and methods of information science, knowledge management, and citation indices. Yet, within this subfield, there are also scholars whose research interests go beyond informetrics, for instance, X.P. Sheng and I. Nonaka who are interested in knowledge management, C.P. Hu, M.S. Lai, and F.C. Ma focusing on basic theories and methods of information science, and Q. Y. Zhang dedicated to the research of information indices. In this case, it is hard to distinguish some groups with fewer members who could be mixed into other groups. Therefore, a proper interpretation of Pathfinder results needs one necessary premise: every anticipated group should have a certain number of members.

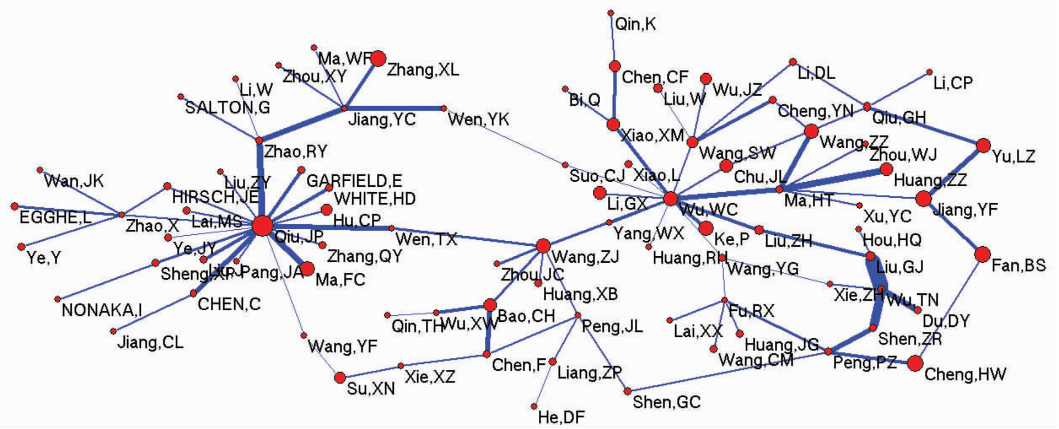

Figure 1 The intellectual structure of CLIS based on an author citation network.

4.2 Revealing main knowledge flow structures of directed weighted citation networks

Figure 2 shows the main knowledge flow structure of informetrics discovered in Figure 1. In informetrics, there are several foreign scholars such as Garfield, White, Chen, Egghe, and Hirsch, and the majority of whom did not publish any articles in Chinese journals, and therefore they act as sources who only delivered knowledge to Chinese scholars. There is only one sink "X. Zhao" who absorbed knowledge from others. Qiu occupied a pivotal position in the structure, who not only gained a great deal of knowledge from others but transferred knowledge to other scholars, most of whom were his doctoral students such as R. Y. Zhao and T. X. Wen.

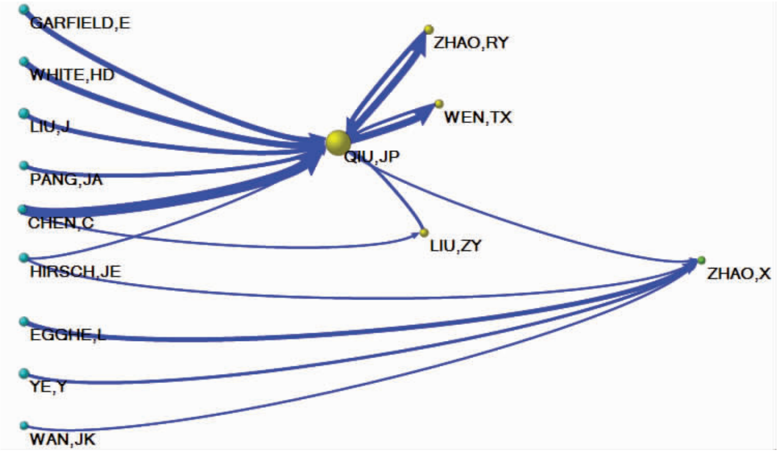

Figure 2 The knowledge flow structure among salient authors in informetrics of CLIS, ruling out the lines with values lower than 2 and isolated authors.

Figure 3 shows the knowledge flow structure of competitive intelligence that is also discovered in Figure 1. Different from Figure 2, the scholars in this field communicated with others more frequently, and the number of scholars in each role-sources, brokers, and sinks appeared almost even. Unlike Figure 2, this structure is more stable: if we remove one scholar, it would impose little impact on the communication among most of the scholars, which may benefit the sustainable development of this field in the long run.

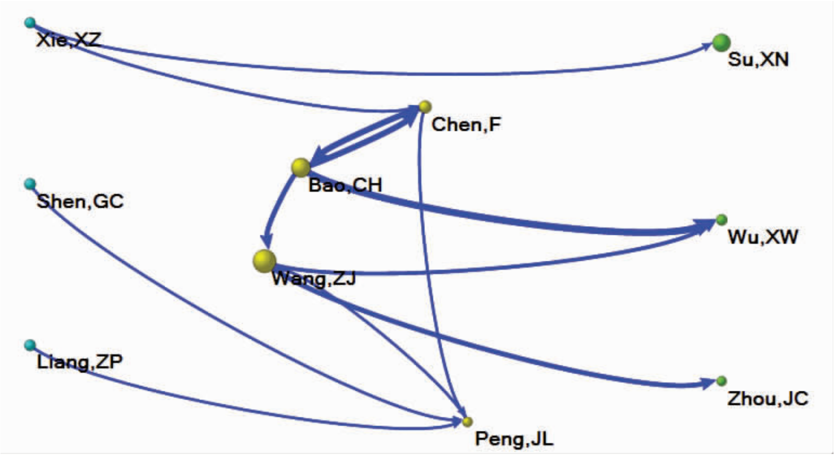

Figure 3 The knowledge flow structure among salient authors in competitive intelligence of CLIS, ruling out the lines with values lower than 2 and isolated authors.

Additionally, we should note that because in citation networks, the transfer of knowledge for non-contacting nodes is implied but not factual, we would better focus on the relation between two authors rather than a chain of knowledge flow; for example, we can find in Figure 2 that Hirsch mainly delivered knowledge to Qiu and X. Zhao but we are unaware of whether the knowledge RY Zhao and Wen received from Qiu is the same one from Hirsch to Qiu.

5 Concluding remarks

This paper constructed a framework for discovering communication structures of directed weighted citation networks through Pathfinder. In this context, communication structures mainly comprise two types: communication communities and main knowledge flow structures.

We probed into the intellectual structure of a directed weighted citation network. We converted it into an undirected weighted one through the use of the sum of mutual citations and set the parameters in Pathfinder as r=∞ and q=n-1, which severed to pruning the network and captured main subgroups of the network. We found that each expectant community should contain a certain number of members when employing Pathfinder to tap into the communities of a network, or it could lead to inaccurate interpretations of communities' profiles.

We also delved into the main knowledge flow structure of the directed weighted network and set the parameters of Pathfinder as r=∞ and q=n-1. By virtue of redistributing the nodes-arraying nodes with only output degrees to the left, nodes with only input degrees to the right, and the rest in the middle, we observed more clearly the roles of sources, sinks, and brokers of a network as if they were distributed chronologically.

This paper mainly focused on applying Pathfinder algorithm to mine the communities of directed weighted network and display communication flow within a specific community. Nonetheless, there were two limitations. First of all, we did not compare Pathfinder with other algorithms to demonstrate the difference. Second, we only applied Pathfinder to discover the communication structure of library and information science which is a relatively small subject considering the citation networks of other subjects are more complicated. Thus, in future research, the comparisons should be carried out systematically and the stability and robustness of Pathfinder also should be tested in different subjects.

References

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson´s correlation coefficient. *Journal of the American Society for Information Science and Technology, 54* (6) , 550–560.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023.*

Blondel, V.D., Guillaume, J.L., Lambiotte, R., & Lefebvre, E.(2008).Fast unfolding of communities in large net–works. *Journal of Statistical Mechanics: Theory and Experiment, 2008* (10), P10008.

Chen, C. (1998). Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers, 10* (2), 107–128.

Chen, C. (2006a). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific lit–erature. *Journal of the American Society for Information Science and Technology, 57* (3), 359–377.

Chen, C. (2006b). *Information visualization: beyond the horizon*. London: Springer Science & Business Media.

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek.* New York: Cam–bridge University Press.

Garfield, E., Pudovkin, A. I., & Istomin, V. S.(2003). Why do we need algorithmic historiography?. *Journal of the American Society for Information Science and Technology, 54* (5), 400–412.

Garvey, W. D. (1979). *Communication: the essence of science*. New York: Pergamon Press Inc.

Girvan, M., & Newman, M. E.(2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99* (12), 7821–7826.

Hummon, N. P., & Doreian, P.(1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11* (1), 39–63.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.

Kohonen, T. (1982). Self–organized formation of topologically correct feature maps. *Biological Cybernetics, 43* (1), 59–69.

Leicht, E. A., & Newman, M. E. (2008). Community structure in directed networks. *Physical Review Letters, 100* (11), 118703.Lin, X., White, H. D., & Buzydlowski, J. (2003). Real–time author co–citation mapping for online searching. Information Processing & Management, 39 (5), 689–706.

Liu, J. S., & Lu, L. Y. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch in–dex as an example. *Journal of the American Society for Information Science and Technology, 63* (3), 528–542.

Lu, L. Y. Y., & Liu, J. S. (2013). An innovative approach to identify the knowledge diffusion path: The case of resource–based theory. *Scientometrics, 94* (1), 225–246.

Lucio–Arias, D., & Leydesdorff, L. (2008). Main–path analysis and path–dependent transitions in HistCite(TM)–based historiograms. *Journal of the American Society for Information Science and Technology, 59* (12), 1948–1962.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American So–ciety for Information Science, 41* (6), 433.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: studies in knowledge organization.* New York: Ablex Publishing.

Small, H. (1973). Co–citation in the scientific literature: A new measure of the relationship between two docu–ments. *Journal of the American Society for Information Science, 24* (4), 265–269.

White, H. D. (2003a). Pathfinder networks and author cocitation analysis: A remapping of paradigmatic informa–tion scientists. *Journal of the American Society for Information Science and Technology, 54* (5), 423–434.

White, H. D. (2003b). Author cocitation analysis and Pearson´s r. *Journal of the American Society for Informa–*

tion Science and Technology, 54 (13), 1250–1259.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science, 32* (3), 163–171.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co–citation analysis of information sci–ence, 1972–1995. *Journal of the American Society for Information Science, 49* (4), 327–355.

Yan, E. (2014). Finding knowledge paths among scientific disciplines. *Journal of the Association for Information Science and Technology, 65* (11), 2331–2347.

Yu, W., Chen, G., Cao, M., & Kurths, J. (2010). Second–order consensus for multiagent systems with directed topologies and nonlinear dynamics. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 40* (3), 881–891.

Zhao, D. (2006). Towards all–author co–citation analysis. *Information Processing & Management, 42* (6), 1578–1591.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information sci–ence 1996–2005: Introducing author bibliographic–coupling analysis. *Journal of the American Society for In–formation Science and Technology, 59* (13), 2070–2086.