2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8357

What should a researcher first read? A bi-relational citation networks model for strategical heuristic reading and scientific discovery

Moreno Pascual, Cesar [a]; Martinez de Ibarreta Zorita, Carlos [b]
[a] EAE Business School, Spain, [b] Universidad Pontificia de Comillas, Spain.

Abstract

Scientists usually try to find relevant and updated documents for their research. Also, they face an abundance of information. Most of the methodologies and algorithms look Backwards, so they suffer an inevitable time delay. We propose a recommendation algorithm combining Forward and Backward citation entire networks and Macro, Meso and Micro metrics that concludes in a strategic map and a heuristic reading path. Underlying it, we found an asymmetric bowtie scientific advance model that informs all, solving the abundance problem with a triple reduction and a heuristic reading path.

Keywords: *citation networks, big data, recommendation system, network metrics, science advance model, bibliometrics*

1. Introduction

On the one hand, Scientists look forward to discovering the emergent knowledge or Research Front (Fujita, Kajikawa, Mori, & Sakata, 2014; Huang & Chang, 2014; Price, 1965; Shibata, Kajikawa, Takeda, & Matsushima, 2009; Small, Boyack, & Klavans, 2014). Usually, they include in that Front those documents cited more frequently (Shibata et al., 2009; Upham & Small, 2010). Additionally, the scientists tend to mention the most recent documents to gather the more updated knowledge. This phenomenon is traditionally called "immediacy factor" (Price, 1965). Therefore, relevance can have several conceptualisations that might be contradictory. On the other hand, the number of scientific documents doubles every 1.8 years (Kleinberg, 1999; Wang, Song, & Barabási, 2013). This abundance makes challenging the choice of which papers to read and how to order their reading. Many methodologies, metrics, recommendations systems and information retrieval algorithms were developed to solve the classic problem of relevance and abundance.

This **Paper proposes** an algorithm to guide that reading. It is based on well-proven networks metrics, Micro (Eigenvector Centrality and Betweenness Centrality), Meso (Modularity maximisation (Blondel, Guillame, Lambiotte, & Lefebvre, 2008; Lancichinetti & Fortunato, 2011; Newman, 2006)) and Macro perspectives combination and it applies them to Backward and Forward entire citation networks. Forward citation has been used until now (Belter, 2016; Couteau, 2014) but locally around a document, and to interpret an existing document when a new one cites it, but saying nothing about the citing new one. Namely, Eigenvector Centrality is an energy diffusion vector in a steady state of energy, representing a ranking of documents where the knowledge (the diffused element) of the network is deposited (Rodriguez & Shinavier, 2010). Therefore, these interpretation offers a relevance criteria about new publications without time delay when it is applied to Forward Networks. The former technical novelty, the combination of the metrics to the two networks over a timeline (taking advantage of the acyclical characteristics) and the usage of three levels of analysis permit two new intimately related outputs:

a)A **Strategic Reading map** that classifies the clusters of documents in 4 areas (emerging mainstream, declining mainstream, emerging new stream, and declining new stream) and ranks them. Inside each cluster, we define a document level ranking based on the three Research Fronts defined in the next output informing a document level ranking.

b)A **Science advance model** of the given topic that supports the previous view, defining three Research Fronts (Forward, Intermediary and Backward). The scheme makes possible the comparison of scientific areas.

Some **metrics have been developed to understand "relevance",** aiming to reduce the number of documents considered by the reader. Most bibliometric indicators, understand that the very relevant are the most cited (Garfield, 1972; Moed, 2010) or the more

prestigious (Bergstrom, West, & Wiseman, 2008) (defining prestige in different ways) in a given period. For comparison, there are many summaries available (Hu, Rousseau, & Chen, 2011; Salvador-Oliván & Agustín-Lacruz, 2015). Notably, the Inmediacy Index considers the average number of times an article is cited in the year it is published, trying to measure the emergence of a document. It is mainly affected by publication patterns and time delays (Salvador-Oliván & Agustín-Lacruz, 2015). Also, some new indicators based on the citation dynamics are developed to substitute the traditional ones using (Hirsch, 2005; Wang et al., 2013).

There are **other network-based techniques**, such as co-citation (Small, 1973) or bibliographic coupling (Kessler, 1963) or a combination of both (Small et al., 2014), but all of them still need a time elapse. As a consequence, a noticeable time delay appears (Fujita et al., 2014). Additionally, reading is usually done by successively incorporating documents in a recursive search (Vazquez, 2000, 2001). In fact, Scientists include referenced materials that attract their interest that there are not among those initially found, expanding the whole system.

For **example**, if we searched in WOS "*Knowledge diffusion or scientific change*" and "*Knowledge networks*", the system would retrieve 721 and 884 documents, respectively. A Recursive search (Vazquez, 2001) carried out entirely, would involve the revision of 22,986 and 31,925 documents respectively. On the contrary, if this recursive search is not done ultimately, we will most likely be locked in the cluster of the documents in which we started the reading, with little chance of jumping into other groups (Ren *et al.*, 2012).

			Reduction			
Topic	**Seed**	**Expanded network**	**Clustering**	**Eigen($\omega_{1,2}^{2forward}$)**	**Betw($\omega_{1,2}^{2forward}$) (*)**	**Suggested documents to read**
Knowledge diffusion or scientific change	721	22986	11,778	87	20	**107**
Knowledge networks	884	31925	9,184	72	15	**87**

Figure 1 Documents Reduction

This way, when a scientist cites a document, the information is extracted and incorporated into the new one, building a new limit in the research horizon, a new Forward or Destiny Front, that can be quantitatively defined. They also consolidate and reviews the existing literature building an Intermediary Front. Finally, the standard backwards-looking, without controlling the time elapse as some of the mentioned metrics do, makes the Origin or Backward front. All these Fronts can be joint in a whole perspective taking advantage of the citation networks acyclical feature, that is in the basement of the inadequacy of other algorithms. Interestingly, these three Fronts form an asymmetric bowtie connected by the timeline, in analogy with the Internet bow tie (Broder et al., 2000), that conceptualises the

scientific advance with a new perspective. The relatively small number of documents in the Forward Front confirms the traditional Price intuition.

2. The model

There are several precedents, coming from search algorithms such Pagerank (Brin & Page, 1998), Hubs and Authorities (Kleinberg, 1999) and other models that combine several network types. Namely, using co-citation and bibliographic coupling (Boyack & Klavans, 2014), direct networks (Caschili, De Montis, Ganciu, Ledda, & Barra, 2014) and local Backward and Forward networks (Belter, 2016; Couteau, 2014). Also, there are some tools like Sci2, Citeseer, Google Scholar or Researchgate that crawls and apply the mentioned or similar bibliometric metrics or algorithms.

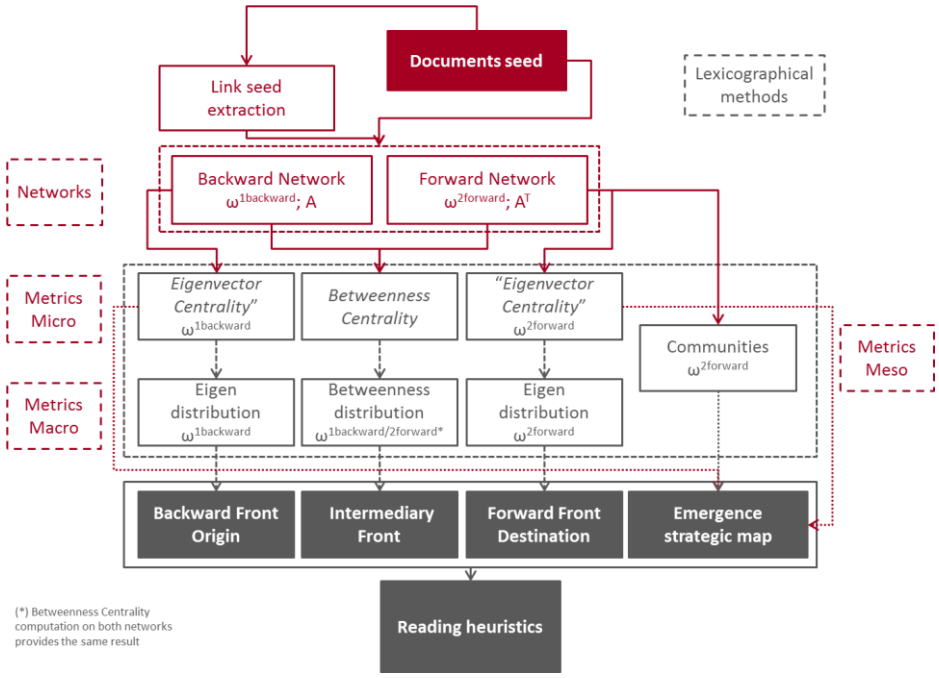

Figure 2 Bi-relational algorithm for reading heuristics

The proposed algorithm applies the following steps:

In the beginning, we apply a lexicographic search using, for example, Web of Science (WoS), retrieving the seed of the system.

In a **second step**, we create two citation networks. The Backward network $\omega^{1backward}$ contains directed edges that come from the document that cites the cited document. Conversely, the Forward network $\omega^{2forward}$ uses edges that comes from the cited documents to the one that cites. We understand both networks as a representation of information flow.

In a **third step**, we compute the Eigenvector Centrality to both entire networks and betweenness centrality to any of them, capturing how information circulates through the documents and across its edges in both directions, ranking them. Both, Eigenvector Centrality (Newman, 2012) and Betweenness Centrality (Freeman, 1977; Newman, 2012), are widely used.

Eigenvector Centrality computed in both networks allows us to interpret the relevance of the origin of the information and the destination. The resulting vector is a ranking of documents where the knowledge of the network is deposited, according to the interpretation of the steady state of energy (Rodriguez & Shinavier, 2010). Defining it more clearly, the Eigenvector Centrality measure, applied to the Forward network $\omega^{2forward}$ offers us the documents that, in the researchers' perspective, gather the most relevant information at the current time. The Backward vision has a symmetrical interpretation.

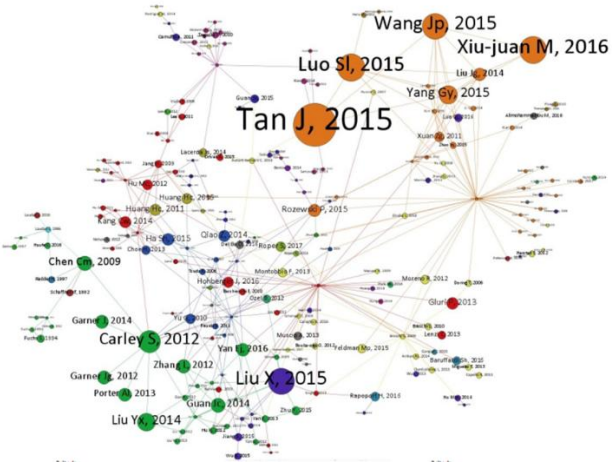

Figure 3 Forward Network example retrieved end 2016. Topic "Knowledge diffusion" or "Scientific change."

However, a complete panorama forces us to analyse what happens between these two states, origin and destination. To do this, we propose the usage of the "Betweenness Centrality" metric. The measure offers a ranking of documents through which the most considerable amount of information passes, mediating the flow of the network. In this case, its calculation also involves the consideration of the entire network structure (Shibata, Kajikawa, & Matsushima, 2007, p. 881). Therefore, all these measures of relevance asses prestige from different and combined perspectives.

The **fourth step** consists of the Community detection. We considered for it **Modularity Maximisation** (Newman, 2006) using the Blondel resolution methodology (Blondel et al., 2008) but applied to the Forward network $\omega^{2forward}$ that detects the clusters formed currently. The clustering step also makes possible to unravel different citation knowledge areas patterns at the same time, as Business and Economics.

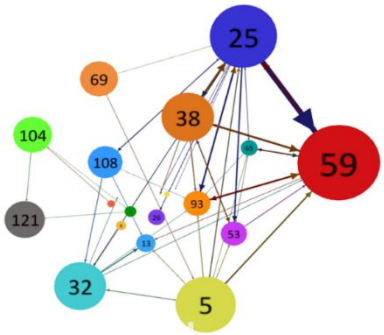

Figure 4 Topic "Knowledge diffusion" or "Scientific change" community detection

The **fifth step** compares the Backward and the Forward Eigenvector relevance in each community to depict the Strategic Map. The vertical axis expresses the emerging factor. For that, we add the Forward Eigenvector Centrality, on the one hand, and in the other, the Backward Eigenvector Centrality score and we measure the distance to the regression line relating the two criteria. The horizontal axis is the Forward Eigenvector Score. Then we can interpret that the more emergency, the higher vertical position, the more current relevance, the more on the right horizontal position, concluding in a four-quadrant interpretation.

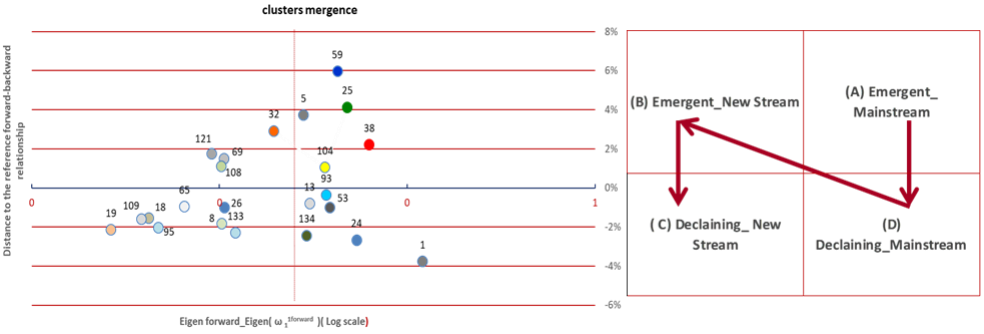

Figure 5 Strategic Map and Reading order example

As a result, in the example depicted in Figure 5, the reading order would be: A) 59,25,5,38,104 B)93,13,53,134,24,1 C) 32,121,69,108. Given that order, inside each

Community, we can logically begin reading the intermediary ranking and then the forward one, concluding with a specific reading suggestion.

Of course, the methodology faces several limitations: "grey literature" (Cooper, Hedges, & Valentine, 2009, pp. 92–93), Ortega hypothesis, the strategic or social citation (Stremersch, Camacho, Vanneste, & Verniers, 2015), obliterated citations (Cole & Cole, 1972; MacRoberts & MacRoberts, 2010) and citation interpretation limits (Amsterdamska & Leydesdorff, 1989; Bellis, 2009), may affect the result. The usage of the whole network as a system, using enough data, might vanish this effects.

3. Reductions and Asymmetric Bowtie

All the mentioned Fronts follows, in all the cases, high skewed distributions that provoke steep decreases in the number of relevant documents. However, the relevant communities detected are only a few attending to the appropriate content and its emergency. Then, a triple reduction its taking place solving the abundance problem.

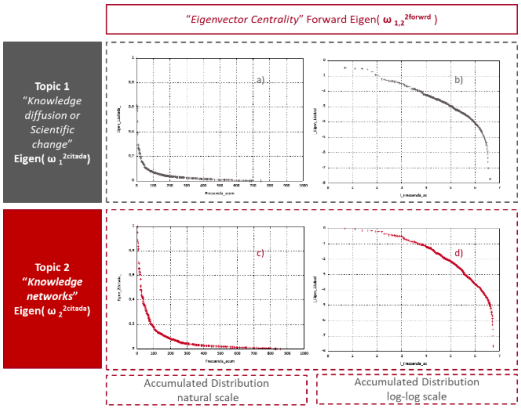

Figure 6 Skewed distribution in several topics. Forward Eigenvector

Finally, a **scientific advance model** is underlying all the reasoning. Interestingly, this model would suggest an asymmetric relation between several fronts, meaning that the Forward is relatively short compared with the Backward front, and the intermediary appears very concentrated, as Price realised many years ago.

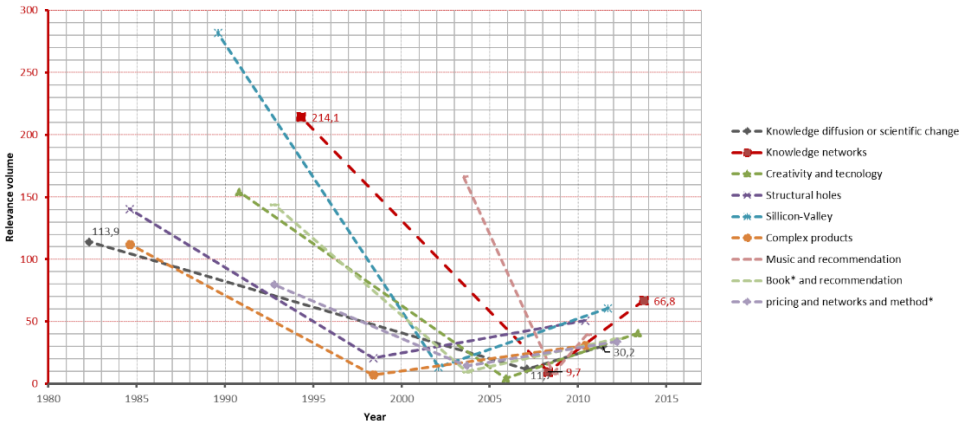

Figure 7 Asymmetric bowtie for several topics

4. Conclusion

A combination of direct Forward and Backward, complete, Networks and Micro, Meso and Macro perspectives not only drives a triple reduction that solves the abundance problem but also gives several "relevance" criteria even for at the very early moments of a document publication. That combination can reconcile relevance and time, understanding the physical interpretation of the Eigenvector Centrality and its application to the forward network as a whole system. The Research fronts conceptualisation, its asymmetric bowtie shape and its relationships, confirm Price intuitions and give light to the reading process. Future research may include weighted edges, an automatic tool and large application for its verification using several databases.

References

Amsterdamska, O., & Leydesdorff, L. (1989). Citations: Indicators of significance? *Scientometrics*, *15*(5–6), 449–471. https://doi.org/10.1007/BF02017065

Bellis, N. De. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to Cybermetrics*. Plymouth United Kingdom: The Scarecrow Press, Inc.

Belter, C. W. (2016). Citation Analysis as a Literature Search Method for Systematic Reviews. *Journal of the Association for Information Science and Technology*, *67*(11), 2766–2777. https://doi.org/10.1002/asi.23605

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(45), 11433–4. https://doi.org/10.1523/JNEUROSCI.0003-08.2008

Blondel, V. D., Guillame, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685. https://doi.org/10.1002/asi

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine BT - Computer Networks and ISDN Systems. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., … Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, *33*(1), 309–320. https://doi.org/10.1016/S1389-1286(00)00083-9

Caschili, S., De Montis, A., Ganciu, A., Ledda, A., & Barra, M. (2014). The Strategic Environment Assessment bibliographic network: A quantitative literature review analysis. *Environmental Impact Assessment Review*, *47*, 14–28. https://doi.org/10.1016/j.eiar.2014.03.003

Cobo, M. J. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630. https://doi.org/10.1002/asi.22688

Cole, J. R., & Cole, S. (1972). The Ortega Hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress. *Science (New York, N.Y.)*, *178*(4059), 368–75. https://doi.org/10.1126/science.178.4059.368

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of Research synthesis and meta-analysis*. (R. S. Foundation, Ed.) (2nd Editio). New York.

Couteau, O. (2014). Forward searching - A complement to keyword- and class-based patentability searches. *World Patent Information*, *37*, 33–38. https://doi.org/10.1016/j.wpi.2014.01.007

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*. https://doi.org/10.2307/3033543

Fujita, K., Kajikawa, Y., Mori, J., & Sakata, I. (2014). Detecting research fronts using different types of weighted citation networks. *Journal of Engineering and Technology Management - JET-M*, *32*, 129–146. https://doi.org/10.1016/j.jengtecman.2013.07.002

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science (New York, N.Y.)*, *178*(178), 471–479. https://doi.org/10.1300/J123v20n02_05

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hu, X., Rousseau, R., & Chen, J. (2011). Structural indicators in citation networks. *Scientometrics*, *91*(2), 451–460. https://doi.org/10.1007/s11192-011-0587-3

Huang, M. H., & Chang, C. P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics*, *98*(3), 1721–1744. https://doi.org/10.1007/s11192-013-1126-1

Indiana University and SciTech Strategies. (2009). Science of Science (Sci2) Tool. Retrieved from https://sci2.cns.iu.edu

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. https://doi.org/10.1002/asi.5090140103

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, *46*(May 1997), 668–677. https://doi.org/10.1.1.120.3875

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *84*(6), 1–8. https://doi.org/10.1103/PhysRevE.84.066122

Lawrence, S., & Bollacker, K. (2018). CiteSeerX. Retrieved May 28, 2018, from http://citeseerx.ist.psu.edu/index

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, *61*(1), 1–12. https://doi.org/10.1002/asi.21228

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Newman, M. E. J. (2012). *Networks: An introduction*. New York: Oxford University Press.

Price, D. S. (1965). Networks of Scientific Papers. *SCIENCE*, *149*. Retrieved from http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=17&SID=Z15uohRqct1ddYYr4hv&page=1&doc=4

Rodriguez, M. A., & Shinavier, J. (2010). Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics*, *4*(1), 29–41. https://doi.org/10.1016/j.joi.2009.06.004

Salvador-Oliván, J. A., & Agustín-Lacruz, C. (2015). Correlación entre indicadores bibliométricos en revistas de Web of Science y Scopus. *Revista General de Información y Documentación*, *25*(2). https://doi.org/10.5209/rev_RGID.2015.v25.n2.51241

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology*, *58*(6), 872–882. https://doi.org/10.1002/asi.20529

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation. *Journal of the American Society for Information Science and Technology*, *60*(1971), 571–580. https://doi.org/10.1002/asi

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*(8), 1450–1467. https://doi.org/10.1016/j.respol.2014.02.005

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, *32*(1), 64–77. https://doi.org/10.1016/j.ijresmar.2014.09.004

Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics*, *83*(1), 15–38. https://doi.org/10.1007/s11192-009-0051-9

Van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823. https://doi.org/10.1016/j.joi.2014.07.006

Vazquez, A. (2001). Statistics of citation networks. *Science*, 12. Retrieved from http://arxiv.org/abs/cond-mat/0105031

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*(6154), 127–133. https://doi.org/10.1126/science.1237825