



# Toward the Discovery of Citation Cartels in Citation Networks

Iztok Fister Jr.<sup>1</sup>, Iztok Fister<sup>1</sup> and Matjaž Perc<sup>2,3\*</sup>

<sup>1</sup> Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia, <sup>2</sup> Faculty of Natural Sciences and Mathematics, Institute of Physics, University of Maribor, Maribor, Slovenia, <sup>3</sup> Center for Applied Mathematics and Theoretical Physics, University of Maribor, Maribor, Slovenia

In this perspective, our goal is to present and elucidate a thus far largely overlooked problem that is arising in scientific publishing, namely the identification and discovery of citation cartels in citation networks. Taking from the well-known definition of a community in the realm of network science, namely that people within a community share significantly more links with each other as they do outside of this community, we propose that citation cartels are defined as groups of authors that cite each other disproportionately more than they do other groups of authors that work on the same subject. Evidently, the identification of citation cartels is somewhat different, although similar to the identification of communities in networks. We systematically expose the problem, provide theoretical examples, and outline an algorithmic guide on how to approach the subject.

## OPEN ACCESS

**Keywords:** citation network, citation cartel, network science, community detection, cooperation

### Edited by:

Antonio F. Miguel,  
University of Évora, Portugal

### Reviewed by:

Zoran Levnjacic,  
Institute Jozef Stefan, Slovenia  
Marija Mitrovic Dankulov,  
Institute of Physics Belgrade, Serbia  
Walter Quattrociocchi,  
IMT Lucca, Italy

### \*Correspondence:

Matjaž Perc  
matjaz.perc@uni-mb.si

### Specialty section:

This article was submitted to  
Interdisciplinary Physics,  
a section of the journal  
Frontiers in Physics

**Received:** 02 November 2016

**Accepted:** 29 November 2016

**Published:** 15 December 2016

### Citation:

Fister I Jr., Fister I and Perc M (2016)  
Toward the Discovery of Citation  
Cartels in Citation Networks.  
Front. Phys. 4:49.  
doi: 10.3389/fphy.2016.00049

## 1. INTRODUCTION

Citation network is a kind of Social network that can be represented as a direct graph with nodes representing papers  $\{P_1, \dots, P_n\}$  and edges  $e(P_i, P_j)$  between two nodes  $P_i$  and  $P_j$  denoting a co-citation relationship [1], when the paper  $P_i$  cites a paper  $P_j$ . This relationship is shown in the direct graph as an arrow directed from a node  $P_i$  to  $P_j$ . At the moment, the number of pure citations of scientific articles is becoming one of the most important measures of scientific impact and quality. Hence, the authors are trying to obtain as many citations as possible for their works by creating so-called citation cartels, where members cite each other in order to increase their own number of citations. Nowadays, this phenomenon is also more exposed by additional academic pressure which forces scientists to publish as many papers as possible. In contrast, when they do not publish enough, they suffer from losing their job or go down on the career scale. In line with this, the concept “publish or perish” [2, 3] has been introduced which has added an additional strength to the scientific competitiveness. Moreover, many measures of scientific impact based on citations are stepping into reality [4–8]. By the same token, bibliographical databases [9] as, for example, Web of Science, Scopus, DBLP, are covering more and more data which are later used for evaluations of researchers and their institutions.

In a nutshell, the modern tools for estimating the quality of research work analysis faces scientists with working hard in order to hold their working positions. Additionally, every year, many more researchers join the scientific community (e.g., scientists from China and India). With so many people in science, many new ways for increasing their publicity have been introduced. Especially, lowly ranked researchers are trying to obtain citations synthetically [10]. Thus, new ways, some legitimate others not, for achieving more citations have been developed. Interdisciplinary research collaborations [11] and motivators for work [12] of course help in this endeavor, and rightfully so. But there are also people, sometimes collaborators [13], sometimes friends and colleagues, and

sometimes third party, who are increasingly often citing each other inside citation networks to increase the impact of their papers by establishing citation cartels.

In this paper, we try to expose the problem of detecting citation cartels inside citation networks, visualize the random generated author citation network and show how to discover a citation cartel inside the citation networks.

## 2. THE CITATION RACE

In these days, a lot of people argue that scientific publishing has become a kind of race. Every researcher would like to be reputed and well-known in the scientific world. Some years ago, researchers, especially from the prominent institutions, were the most famous in science but with the arrival of the Internet the knowledge started to spread around the world, while many more people have gained access to data and publications. Therefore, more and more people have decided to work in the scientific community. Consequently, ranking of researchers has become more rigorous. As stated before, many metrics were developed for assessing researchers. On the other hand, the Matthew effect by Perc [14], asserting that people who have huge resources, a good citation pool and connections, are much more awarded than people without these, has been much more visible recently. Additionally, social networks also help some researchers to promote their works [15, 16].

It is worth mentioning that the citation race is much more visible nowadays than it was in the past. However, researchers have found many ways to promote their papers and raise the number of citations of their articles. Former professional road bike rider Hamilton [17] exposed in his book “The Secret Race: Inside the Hidden World of the Tour de France” tricks which riders used to improve their results in cycling. The most words were given to blood doping and the use of many medications. A very similar story was also depicted by Millar [18]. The question is if scientific publishing is not a kind of doping as well? After analyzing a bunch of papers, it seems that there are many patterns and complex connections in citation networks. In the next sections, we will outline our concept of citation cartels.

## 3. CITATION CARTELS

Citation networks describe relationships between researchers and papers connected with reasonable citation relationships. These networks are a useful way for analyzing the hidden relationships, e.g., information about the citation behavior of researchers. In line with this, many questions can be asked such as, for instance, how authors select which publications to cite.

In the past, the citation process was fair. An intention to cite some paper follows a principle saying “the right cites are only those obtained from unknown readers.” In this case, readers recognize the author’s work at first, instead of recognizing the author earlier. However, each cite means an acknowledgment for an author’s good work while, at the same time, it increases an author’s reputation. This fact has caused that the fair

citation game has been becoming more and more unfair. Finally, almost all rules have been ignored in the struggle for citation.

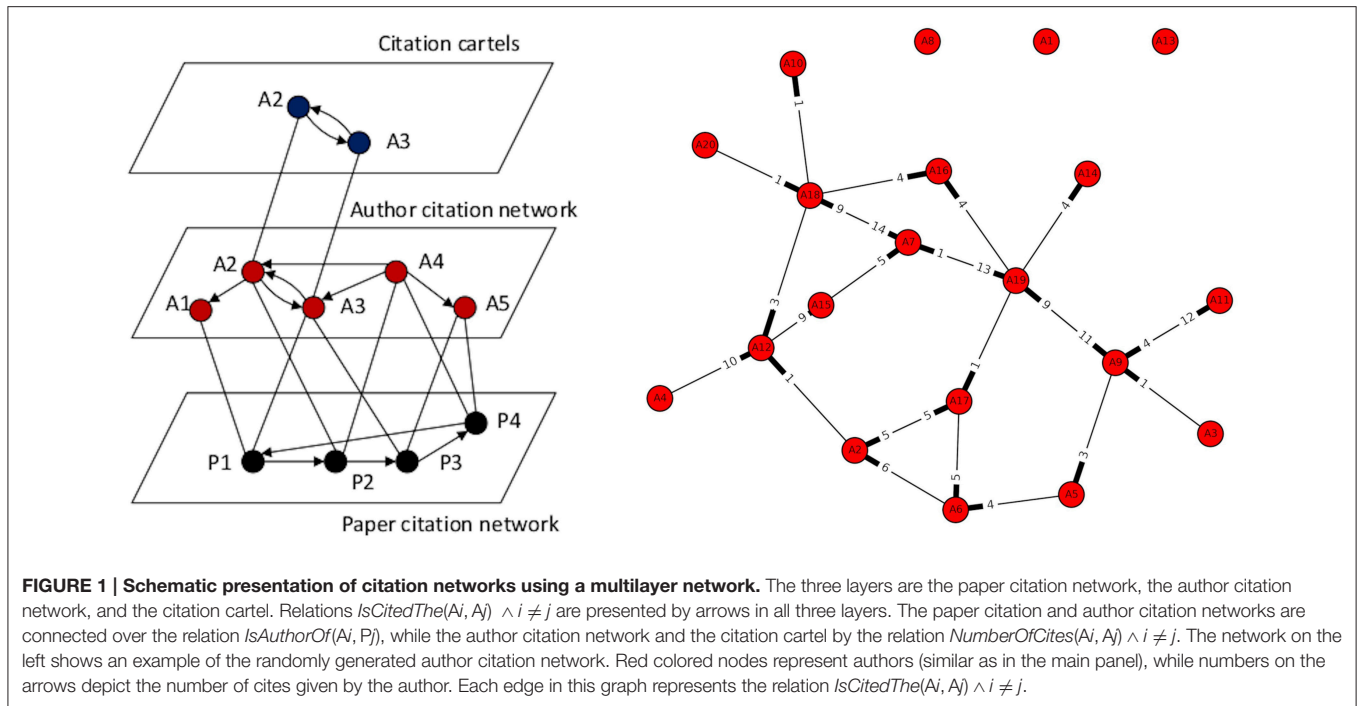
Consequently, the citation cartels have been established in order to make the difference between the quality of the scientists, measured by the number of cites, higher. The concept of citation cartels was firstly exposed in 1999 in an essay by Franck [19] that defined this phenomenon as groups of Editors and Journals working together for mutual benefit. Actually, this definition refers to Editors that were using the inter-journal cites to increase the Impact Factors of their Journals. Recently, the citation cartels have also addressed other relationships, like Editor to authors or authors to authors.

In line with this, new levels of citation networks have emerged. These can be represented in the sense of multi-layer graphs [20] (**Figure 1**, left). For instance, an original paper citation network is represented as a direct graph with nodes denoting papers and edges describing a relation  $IsCitedThe(P_i, P_j)$  meaning the paper  $P_i$  is citing a paper  $P_j$ . Citation networks, the authors’ collaboration, as well as author citation networks can be extracted from the original paper. Interestingly, each edge in the author citation networks (**Figure 1**, right) includes also a weight denoting the numbers of cites that author  $A_i$  gives to author  $A_j$  and vice versa. Finally, the citation cartels are derived from the author citation networks, where the number of inter-citations between two nodes needs to exceed some threshold value in a clique of order 2. In this sense, the discovery of citation cartels in networks is much the same as the discovery of communities in networks [21, 22], although with a twist in that members of citation cartels might do their best to stay undetected.

## 4. DISCOVERING THE CARTELS

Discovering the citation cartels in multi-level graphs can be implemented easily by using the modern semantic web tools for manipulating the knowledge on the Internet, i.e., Resource Description Framework (RDF) and RDF query language (SPARQL). Knowledge stored in RDF format in the form of structured meta-data is stored as graphs, where nodes represent resources consisting of triplets “subject-predicate-object.” RDF uses graphs as a formal basis [23], where subjects are connected with their corresponding objects using edges. Edges denote property relations. The SPARQL query language provides semantic query language for retrieving and handling data stored in RDF format. This language was used for discovering the citation cartels in this study. The results of simulations on an artificially generated citation network showed that it was easy to discover citation cartels in this network using the mentioned semantic web tools. Let us suppose two researchers, i.e., researcher A1 and researcher A2, who published two of their own papers P1 and P2, respectively. In the first paper P1 published by author A1, the reference is on the second paper P2 published by the author A2. Here, two scenarios can be taken into account, as follows:

1. when the paper P1 is not coauthored by author A2 and vice versa (**Figure 2**, left), and



2. when authors A1 and A2 are coauthors of the other mutual papers (Figure 2, right).

In both cases there is a suspicion that both authors belong to a citation cartel, especially when the number of their mutual citations is higher. However, the first scenario is more restrictive and, therefore, a more sophisticated method of discovering the citation networks.

The illustrated scenarios are described easily using a first-order predicate logic, where concepts like authors and papers are connected by corresponding relations. Let us assume that a relation connects an author (also subject) to his/her paper (object) using the property (predicate). Then, the relation can be defined as a predicate connecting the subject with the object in general. The relation can be described as a triplet *subject-predicate-object*. In predicate logic, this triplet can also be written as *predicate(subject, object)*.

The corresponding citation network is constructed by the following relations:

- *IsCitedThe*: The paper  $P_i$  is cited by the paper  $P_j$  in the paper citation networks.
- *IsAuthorOf*: The paper  $P_i$  is co-authored by the author  $A_j$  and connects paper citation networks with author citation networks.
- *NumberOfCites*: Determines the number of cites that the author  $A_i$  gives to the author  $A_j$  and connects the author citation networks with the citation cartels.

Let us notice that the relation *NumberOfCites* represents a filter allowing that authors  $A_i$  and  $A_j$  belong to a citation cartel, when the

$$NumberOfCites(A1, A2) \geq Threshold \wedge \quad (1)$$

$$NumberOfCites(A2, A1) \geq Threshold,$$

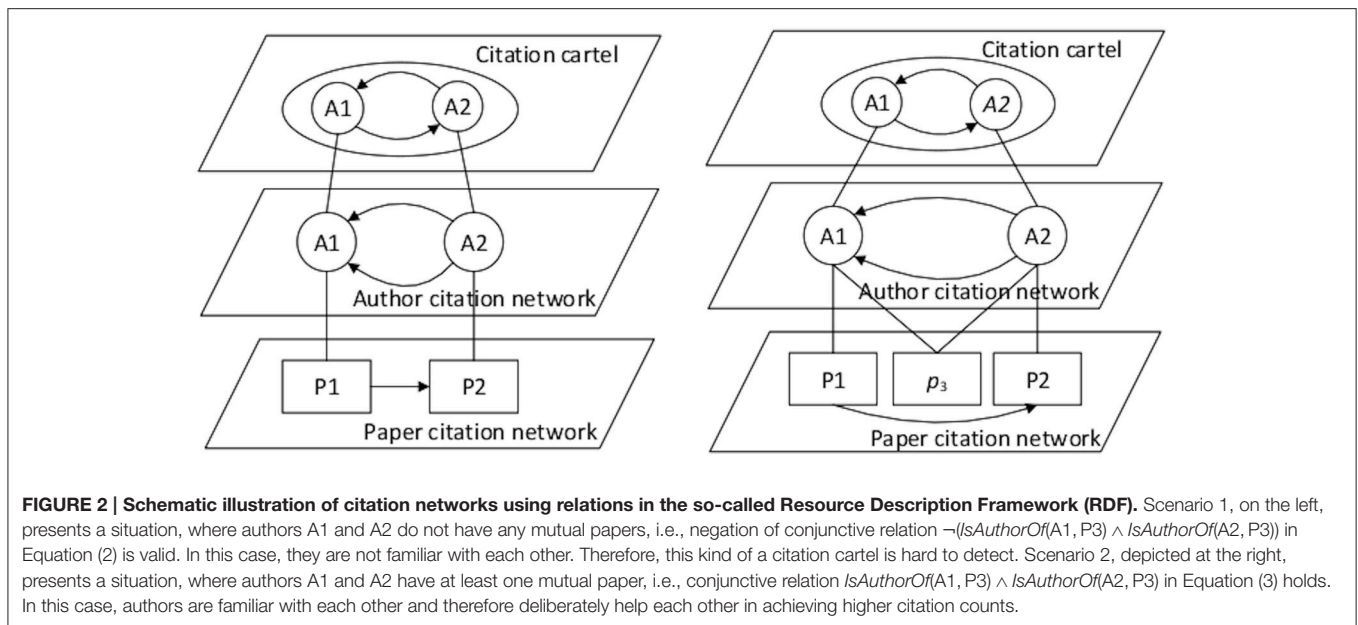
where *Threshold* determines the level needed to declare the existence of a citation cartel. In place of the threshold value, other measures can also be used for filtering, e.g., the ratio between the number of cites that the author  $A_i$  gives the author  $A_j$  and the number of all cites by the author  $A_j$ , and vice versa. The citation cartel for Scenario 1 can be detected by defining the following SPARQL query formally expressed in first-order predicate logic as

$$\begin{aligned} IsInCartelWithV1(A1, A2) = & \\ & IsAuthorOf(A1, P1) \wedge IsCitedThe(A1, A2) \wedge \\ & IsAuthorOf(A2, P2) \wedge IsCitedThe(A2, A1) \wedge \\ & \neg(IsAuthorOf(A1, P3) \wedge IsAuthorOf(A2, P3)) \wedge \\ & NumberOfCites(A1, A2) \geq 10 \wedge NumberOfCites(A2, A1) \geq 10), \end{aligned} \quad (2)$$

while for Scenario 2 as

$$\begin{aligned} isInCartelWithV2(A1, A2) = & \\ & IsAuthorOf(A1, P1) \wedge IsCitedThe(A1, A2) \wedge \\ & IsAuthorOf(A2, P2) \wedge IsCitedThe(A2, A1) \wedge \\ & IsAuthorOf(A1, P3) \wedge IsAuthorOf(A2, P3) \wedge \\ & NumberOfCites(A1, A2) \geq 10 \wedge NumberOfCites(A2, A1) \geq 10). \end{aligned} \quad (3)$$

Note that the last conjunction of predicate *isAuthorOf* denotes all papers written or not by either authors A or B. In order to declare the existence of a citation cartel, the number of inter-citations given by authors A1 and A2 must be higher than the threshold value of 10.



## 5. CONCLUDING REMARKS

Today, citation cartels, where members of these mutually cited papers of authors with which they are known or not known, have become reality in the research domain. In the everyday harder citation race, the cartels imply an easy way to obtain scientific excellence by increasing the number of one's own citations.

The aim of this paper is to discover the citation cartels using the modern semantic web tools for manipulating the knowledge on the Internet, i.e., RDF and SPARQL. However, the proclamation that two authors create a citation cartel is very dangerous, because we cannot ever be sure that this indictment really holds in the real-world. We can only indicate that there is a high probability of citation cartel existence, but this fact needs to be confirmed using a detailed analysis.

## REFERENCES

- Egghe L, Rousseau R. *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Amsterdam: Elsevier Science Publishers (1990).
- Linton JD, Tierney R, Walsh ST. Publish or perish: how are research and reputation related? *Ser Rev.* (2011) 37:244–57. doi: 10.1080/00987913.2011.10765398
- Clapham P. Publish or perish. *Bioscience* (2005) 55:390–1. doi: 10.1641/0006-3568(2005)055[0390:POP]2.0.CO;2
- Hirsch JE. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci USA* (2005) 102:16569–72. doi: 10.1073/pnas.0507655102
- Hirsch JE. Does the h index have predictive power? *Proc Natl Acad Sci USA* (2007) 104:19193–8. doi: 10.1073/pnas.0707962104
- Newman ME. Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci USA* (2004) 101:5200–5. doi: 10.1073/pnas.0307545100
- Petersen AM, Wang F, Stanley HE. Methods for measuring the citations and productivity of scientists across time and discipline. *Phys Rev E* (2010) 81:036114. doi: 10.1103/PhysRevE.81.036114
- Igoumenou A, Ebmeier K, Roberts N, Fazel S. Geographic trends of scientific output and citation practices in psychiatry. *BMC Psychiatry* (2014) 14:332. doi: 10.1186/s12888-014-0332-6
- Šubelj L, Bajec M, Boshkoska BM, Kastrin A, Levnjajić Z. Quantifying the consistency of scientific databases. *PLoS ONE* (2015) 10:e0127390. doi: 10.1371/journal.pone.0127390
- Fister I Jr, Mlakar U, Brest J, Fister I. A new population-based nature-inspired algorithm every month: is the current era coming to the end? In: *StuCoSRcC: Proceedings of the 2016 3rd Student Computer Science Research Conference*. Koper: University of Primorska (2016). p. 33–37.
- Lužar B, Levnjajić Z, Povh J, Perc M. Community structure and the evolution of interdisciplinarity in Slovenia's scientific collaboration network. *PLoS ONE* (2014) 9:e94429. doi: 10.1371/journal.pone.0094429
- Damij N, Levnjajić Z, Rejec Skrt V, Suklan J. What motivates us for work? intricate web of factors beyond money and prestige. *PLoS ONE* (2015) 10:e0132641. doi: 10.1371/journal.pone.0132641
- Chen C. Visualising semantic spaces and author co-citation networks in digital libraries. *Inform Process Manage.* (1999) 35:401–20. doi: 10.1016/S0306-4573(98)00068-5
- Perc M. The Matthew effect in empirical data. *J R Soc Interf.* (2014) 11:20140378. doi: 10.1098/rsif.2014.0378

15. Hall N. The Kardashian index: a measure of discrepant social media profile for scientists. *Genome Biol.* (2014) **15**:424. doi: 10.1186/s13059-014-0424-0
16. Kapp JM, Hensel B, Schnoring KT. Is twitter a forum for disseminating research to health policy makers? *Ann Epidemiol.* (2015) **25**:883–87. doi: 10.1016/j.annepidem.2015.09.002
17. Hamilton T, Coyle D. *The Secret Race: Inside the Hidden World of the Tour de France*. New York, NY: Bantam Dell Publishing Group (2013).
18. Millar D. *Racing Through the Dark: The Fall and Rise of David Millar*. London: Orion (2012).
19. Franck G. Scientific communication—a vanity fair? *Science* (1999) **286**:53–5. doi: 10.1126/science.286.5437.53
20. Cui J, Wang F, Zhai J. *Citation Networks as a Multi-layer Graph: Link Prediction and Importance Ranking*. CS224W Project Report (2010).
21. Fortunato S, Hric D. Community detection in networks: a user guide. *Phys Rep.* (2016) **659**:1–44. doi: 10.1016/j.physrep.2016.09.002
22. Fortunato S. Community detection in graphs. *Phys Rep.* (2010) **486**:75–174. doi: 10.1016/j.physrep.2009.11.002
23. Braatz B, Brandt C. Graph transformations for the resource description framework. In: *Proceedings of the Seventh International Workshop on Graph Transformation and Visual Modeling Techniques*. New York, NY (2008). p. 16.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Fister, Fister and Perc. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.