EasyChair Preprint

№ 1960

Finding Citation Cartels in Academic Research

Arindrima Koley and Subhankar Mishra

November 16, 2019

Finding Citation Cartels in Academic Research

Arindrima Koley[1][0000−0002−2136−0369] and
Subhankar Mishra[2,3][0000−0002−9910−7291]

[1] Institute of Mathematics and Applications, Bhubaneswar, Odisha 751029, India
smishra@niser.com
[2] School of Computer Sciences, National Institute of Science Education and Research, Bhubaneswar, Odisha 752050, India
[3] Homi Bhabha National Institute. Anushaktinagar, Mumbai 400094, India
smishra@niser.com
https://www.niser.ac.in/users/smishra

Abstract. In recent days plagiarism and invalid citation of papers have become a threat to the research world. Number of times a paper is being cited or the "cited by" number may define the quality of work done in that respective paper. So, to increase that number, researchers are getting involved into cartels of invalid citations to compete with the contemporary research world. In this paper we have tried find citation cartel. Initially we have assigned weights to each reference paper with respect to its use in the main paper. We have checked with different papers and have finally set two conditions depending on which we qualify a cited paper to be a 'valid citation' or 'invalid citation'. Continuing the same process with each reference paper we have tried to find a relation among the authors of papers having invalid citations of a close community. The community of the authors having invalid citations is our generated citation cartel.

Keywords: citation · cartel · academic research

1 Introduction

"We must stop the avalanche of Low-Quality Research." Though there are many factors like Impact Factor to measure the quality of a research journal, there are ways to manipulate such factors. Manipulation is mainly done by coercive journal self-citation or by publishing non-citable papers or by participating in a citation cartels [3]. Group of researches involve into citation network for mutual benefits to improve the standard of their published papers by citing each other's work.

Previously, researchers used valid citations only, ones that were relevant to their publications. But in recent days, at many institutions, a higher number of good publications upgrades a researcher in their profession, thus creating a competitive environment and this is leading to invalid citations and the involvement to citation cartels. "Hey I'm writing an article, your paper X is relevant, I will cite it if you cite one of my articles in one of your future papers (if the topic is appropriate)" [Research Gate]. Acceptance of such proposals for mutual benefits

is diminishing the value of worthy publications. This is a serious threat to the research world. Few researchers have worked on finding an existing citation cartel inside a citation network using graph theory and depending on a threshold value [1].

In this paper, we try to find citation cartels by quantifying the validity of each reference cited. We have used Tf-idf to find the keywords and finding a weight to each reference paper with an algorithm to categorize the invalid citations. With this data, we have tried to set up an author to author correspondence and thus identify an existing cartel. With this, we have also plotted a graph depicting the existence and percentage of false citations in each paper.

Section 2 gives a brief overview of the past research work done in the area of citation cartels. Introduction to the techniques TF-IDF and Community Detection used in the paper are stated in section 3 and section 4. Methodology for finding citation cartels is described in section 5. Results and Conclusion are stated in section 6 and section 7 respectively.

2 Literature Review

Identification of citation cartel has not been seen so prominently, but it is somehow similar to identification of communities in networks. Works are being done on this, which provides systematical approach with theoretical examples for identifying citation cartels using the modern semantic web tools for manipulating the knowledge on the Internet [1].

Again based on citation data, research has been done to study impact factor biased self-citations. It gives a trend of increasingly pervasive journal self-citation malpractices and unwanted consequences such as inflated perceived importance of journals and biased journal ranking [3].

Impact factor for a journal is calculated based on the number of times articles published in the journal are cited. But this is of wide and thus researchers opt not to show specific results because that can impact on academic reputation. Thomas Reuters used the term 'Citation Stacking' instead of 'Citation Cartel' [2]. Works are also done to examine the proximity of authors to those they cite using degrees of separation in a co-author network, eventually using collaboration networks to expand on the notion of self-citation [4].

3 TF-IDF

In order to show how important a particular word is to a document in a collection of documents or corpus [5], a very well known numerical statistic is the term frequency - inverse document frequency also known as TF-IDF. This statistic is also utilized as a measuring factor while retrieving information, mining text and modeling users. The TF considers documents as bag of words, while ignoring the order of words. For example, a document with 5 occurrences of the term is more relevant than a document with 1 occurrence. However, it is not 5 times more relevant. The IDF takes into account the frequency of the term in the entire

corpus as frequent terms are less informative than rare terms. Hence it assigns high weights to rare items and low weights to frequent terms.

4 Community Detection

Community detection algorithms basically help us determine how the groups are structured in a graph or network. The study of communities and their detection is important to the study of networks such as computer and information networks, social networks and biological networks. A community structure in the context of networks is defined as the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network [6].There are several community detection algorithms in the literature, with our focus on Louvain method [7] which is a heuristic method that is based on modularity optimization.

5 Methodology

To start with we have taken any published paper (say P) and the papers mentioned in the reference (say $P1, P2, \ldots$.). We have done our research mainly in two phases:

1. To classify as invalid citation
2. To find an existing cartel centering a certain paper

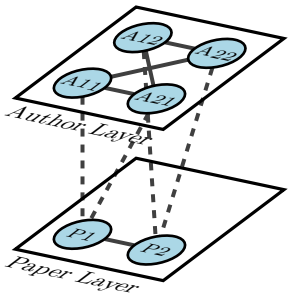

Fig. 1. Graphical representation of authors (A11, A12, A21, A22) and papers (P1 and P2) in the citation network.

5.1 Finding invalid citation

There may be numerous papers mentioned in the reference, let us consider the first one (say P1). We first extract the portion from the paper P where concepts or extract from paper P1 is used. Let us take this extract as sample S. So we have three texts to deal with:

1. Main paper: P
2. Reference paper : P_1
3. Extracted text from P : S

To proceed with we have found the key words (depending on importance level) from the sample S using tf-idf with R programming language [8].

So we get a list (we are working with 20 words) of words from sample S and also the corresponding frequencies in P, P_1, S. Let $a_1, a_2, \ldots, a_{20}$ be the 20 key words obtained from sample $S, (x_1, x_2, \ldots, x_20), (y_1, y_2, \ldots, y_{20}), (z_1, z_2, \ldots, z_{20})$ be corresponding frequencies of $(a_1, a_2, \ldots, a_{20})$ in $P, P1, S$ respectively. Let the weight of these words be defined by the ratio (x_i/z_i) with respect to the paper P. Next we have categorize the word list into two categories depending on the ratio (zi/yi):

1. Category 1 (C1): { z_i s.t $(z_i/y_i) > 1$ i.e. $z_i > y_i$ }
2. Category 2 (C2): { z_i s.t $(z_i/y_i) \leq 1$ i.e. $z_i \leq y_i$ }

Let n be the number of words in $C1$ and m be the number of words in $C2$. Next we have considered two summations:

$$X = \sum_{i=1}^{n} x_i \quad \text{and} \quad z_i \in C1 \tag{1}$$

$$Y = \sum_{i=1}^{m} x_i \quad \text{and} \quad z_i \in C2 \tag{2}$$

After experimenting with various papers with these weights, we have come to conclusion that if

1. $m > 5$ and
2. $X \leq Y$

Then the cited paper P1 can be called as a valid citation or may be not at least an invalid one.

5.2 Finding Citation Cartel

We have approach this theoretically. After finding the papers (from the papers mentioned in the reference) used as invalid citation, we save those papers into an array say Papers [50]. Next we can check the authors of each paper and can create a matrix with the 1st column as the name of the papers with the main paper at the top and each row having the mane of the corresponding paper. Let $A, B, C, ...$ represent the authors of the main paper and $(A_1, A_2, ...)$, $(P_1, P_2, ..., P_m)$ represent the respective authors column and paper rows.

If we take an example of matrix with maximum 4 authors of any paper then:

If any of $A, B, ...$ matches with any of A_{ij}, then there is an invalid self-citation since we are already working with papers being cited wrongly (referring to section 5.1) Next we create similar matrices with each papers $P_1, P_2, \ldots, P_m$

Table 1. Table captions should be placed above the tables.

	A1	A2	A3	A4
Main Paper (P)	A	B	C	D
P1	A_{11}	A_{12}	A_{13}	D_{14}
P2	A_{21}	A_{22}	A_{23}	D_{24}
P3	A_{31}	A_{32}	A_{33}	D_{34}
P4	A_{41}	A_{42}	A_{43}	D_{44}
P5	A_{51}	A_{52}	A_{53}	D_{54}
P6	A_{61}	A_{62}	A_{63}	D_{64}

as the' main paper' and name the matrices as $I, II, III, ...,$ ($m + 1$ matrices). Then we check if any author is appearing in more than 2 matrices, we can say that the author is involved into some cartel. Let us same those authors into some array called Authors [50] s.t $[a_1, a_2, \ldots., a_n]$. So we get an array of authors involved into some kind of citation cartel. Next we need to check whether any of a_i is citing any of a_j's paper and again if aj is citing any other a_i,s paper from that array. If this search result goes around into the list then we can say that the authors are forming a cartel.

The goal is to find cartels in this matrix which can also be viewed a network of invalid citations building the by the previous processing of the data. We propose the use of community detection algorithms to find clusters of the authors involved in invalid citations regularly and constantly among each other. Specifically we propose the Louvain method [7] as it is a very simple method to find the clusters (cartels here) in the invalid citation graph by optimizing modularity.

6 Results

The table 2 shows our experimental data showing valid and invalid citations for $X \leq Y$ and $X > Y$. The first two data show valid citation and the last two for invalid citations.

Table 2. Calculation of X and Y for a sample set.

Paper Name	X	Y
P1	86	219
P2	19	200
P3	250	55
P4	197	108

We have plot graphs with the frequencies of the keywords of paper P and paper P1, and the difference is well understood in between a valid citation in Figure 2 and invalid citation in Figure 3.

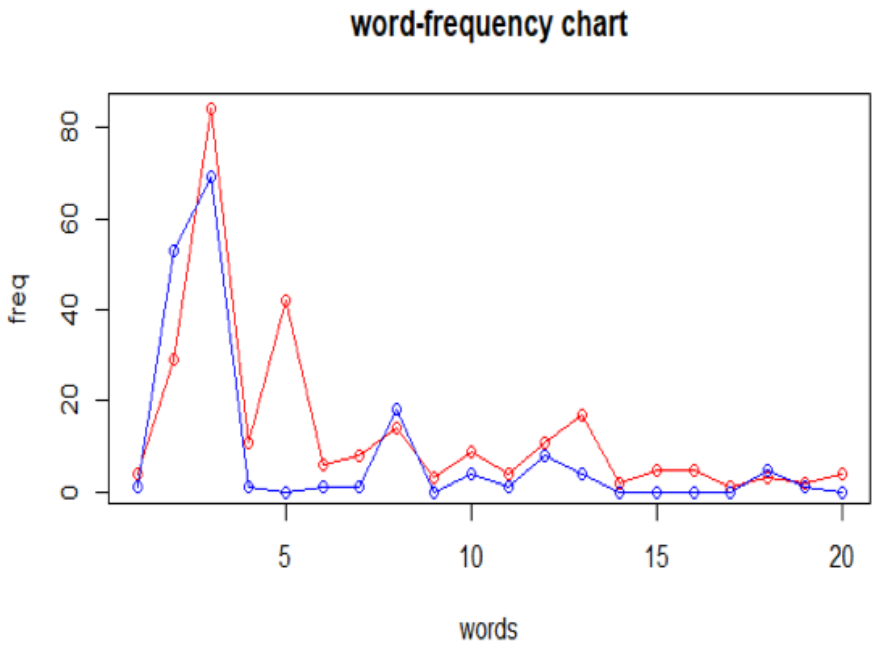

Fig. 2. Good Citation

7 Conclusion

Academia has set the number of citations that an article receives as one of the most important measures of scientific research impact and quality. This follows the acceptance in grant proposals as well as promotions and other job related success. This has resulted in citations cartel. Citation cartel is spreading across the world rapidly and authors are applying various means to increase one's own citation for research excellence. Invalid citation of each other's papers has really come into existence to a great extent. In this paper we have found the irrelevant citations and then have established a way to find citation cartels of authors involved in such cartels. Although we are aware that there might be such instances where authors could have been part of such a cartel accidentally, our future work would involve tuning our approach to reduce false positives and strengthen the validation of existence of cartels in citations.

References

1. Fister Jr, I., Fister, I., & Perc, M. (2016). Toward the discovery of citation cartels in citation networks. Frontiers in Physics, 4, 49.

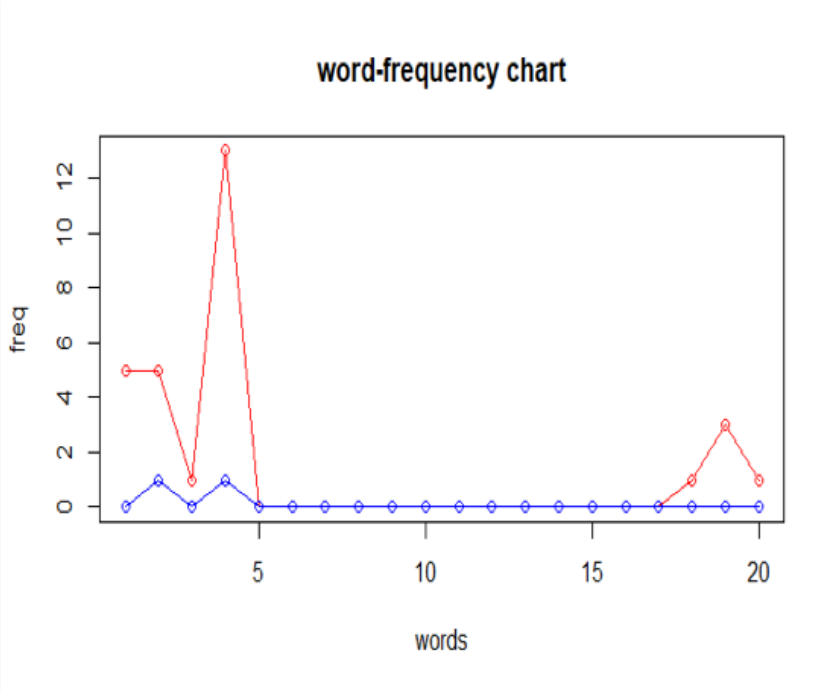

Fig. 3. Invalid Citation

2. "Citation Cartels: The Mafia of Scientific Publishing - Enago." https://www.enago.com/academy/citation-cartels-the-mafia-of-scientific-publishing/. Accessed 14 Nov. 2019.
3. Chorus, C., & Waltman, L. (2016). A large-scale analysis of impact factor biased journal self-citations. PLoS One, 11(8), e0161021.
4. Wallace, M. L., Larivière, V., & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. PloS one, 7(3), e33339.
5. Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning (Vol. 242, pp. 133-142).
6. Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), 7821–7826. doi:10.1073/pnas.122653799
7. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.
8. Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. Journal of computational and graphical statistics, 5(3), 299-314.