Mining Citation Networks to Detect and Analyze Cliques and Cartel-Like Patterns

Joseph G. Davis[0000-0001-9871-8771]

School of Information Technologies, The University of Sydney, J12 1 Cleveland St. SYDNEY NSW 2006, Australia
joseph.davis@sydey.edu.au

Abstract. With the growing emphasis on metrics such as citation count and h-index for research assessment, several reports of gaming and cartel-like formations for boosting citation statistics have emerged. However, such cartels are extremely difficult to detect. This paper presents a systematic approach to visualizing and computing clique and other anomalous patterns through ego-centric citation network analysis by drilling down into the details of individual researcher's citations. After grouping the citations into three categories, namely, self- citations, co-author citations, and distant citations, we focus our analysis on the outliers with relatively very high proportion of self- and co-author citations. By analyzing the complete co-authorship citation networks of these researchers one at a time along with all the co-authors and by merging these networks, we were able to isolate and visualize cliques and anomalous citation patterns that suggest plausible collusion. Our exploratory analysis was carried out using the citation data from Web of Science (Clarivate Analytics) for all the highly cited researchers in Computer Science, Engineering, and Physics. Some of the exciting research opportunities in citation analytics are also outlined.

Keywords: Citation network analysis, citation analytics, cliques.

1 Introduction

Among the different forms of research-related malpractices, citation manipulation is probably the most difficult to track and monitor. Most of the established scientific journals have a clause related to citation manipulation in their publication code of conduct warning authors against including citations with the sole purpose of boosting the citation counts of particular authors or of journals (journal citation stacking to increase the journals' impact factor). However, monitoring and reporting citation manipulation is widely acknowledged to be extremely hard. How can a reviewer or journal editor conclusively argue that certain proportion of references cited by a paper is superfluous or fraudulent? The Web of Science does report on the number of self-citations but this does not address the more sophisticated citation gaming practices involving what are sometimes referred to as citation cartels or rings in which groups of researchers collude to cite a number of papers by the other members of the cartel, independent of the validity of the citations. Like all implicit or explicit collusive behavior, it is never easy to

conclusively demonstrate that cartel-like practices actually exist in academic publishing.

One of the widely known yet poorly researched areas related to research integrity is the practice of excessive self-citations and citations from socially proximate authors such as friends, co-authors, and colleagues. The central role of citations as tokens of esteem and citation practices in the distribution of symbolic capital and its accumulation by researchers is widely acknowledged [1]. The Web of Science, the primary online citation indexing service (previously provided by the Institute of Scientific Information and currently maintained by Clarivate Analytics) views high citation statistics as indicators of the utility and prestige of a given research or researcher as judged by the researchers themselves. It is a signifier of the importance of the researcher's work and status as an expert researcher in the field [2]. The rise in the of use citation indicators and 'highly cited researcher' lists by funding bodies, government agencies, and university ranking systems in particular have contributed to pressure on individual researchers and universities to boost citations. There have been a large number of anecdotal reports of citation stacking and gaming by researchers (and journals) and cartel-like behavior where groups of researchers appear to act with an intent to boost the citation statistics.

In 2011, the editors of the Journal of Parallel and Distributed Computing published by Elsevier took the unprecedented step of retracting a previously published paper for *citation manipulation*. While paper retractions for scientific fraud (data fabrication, falsification of experimental results, among others) and plagiarism are not uncommon, this was one of the first cases in which a paper was retracted because the journal editor-in-chief and the publisher were convinced that many of the citations included the paper were "not of direct relevance to the article and were included to manipulate the citation record" [3]. The paper had 117 citations and it turns out that 50 of these citations were inserted after the final version of the paper had been approved, for no reason other than to boost the citation statistics of the authors and their collaborators.

We carried out exploratory analysis using Web of Science citation data for a sample of highly cited researchers with the goal of detecting plausible citation cliques and anomalous citation patterns.

2. Approach and Methods

Our approach was data driven using the Web of Science data on researchers, their citations, and their co-authors. We carried out the exploratory analysis for all the highly cited researchers in three fields, Computer Science, Physics, and Engineering. The focus on highly cited researchers was motivated by the large number of citations most of them have garnered and their prominent citation profiles. Complete citation data was downloaded for all the researchers in the three lists for the period 2005 to 2016. Given

our interest in self and affinity network citations, percentage distributions for the researchers classified as self, co-author, and distant citations were computed. The sum of self and co-author citation percentages values ranged from 4.44 to 74 in CS, 8.25 to 79 in Physics, and 3.46 to 76 in Engineering. As an illustration, the frequency distribution of the self and highly cited citation percentage sums for the highly cited Physics researchers is shown in Fig. 1.

We performed directed, ego-centric, co-author citation network analyses for each of 5-6 outliers with the highest self and co-author citation totals from each of the three lists. The primary ego was the highly cited researcher and all the co-authors are the alters. The area of the node represents self-citations and thickness of the links represents number of citations. To avoid duplication in citation counts, if a citation with 4 authors of whom only 1 happened to be a co-author of the ego, then this was counted as 0.25. We found that several nodes in the co-author citation networks were themselves highly cited researchers which made it possible to merge the egocentric co-author citation networks of 4 or 5 egos. Clique analysis was carried out on the merged network to compute the fully connected, non-contained cliques.

3. Key Findings

Our analysis detected preliminary evidence of significant clique activity in the merged graphs. With a merged co-author citation network of 197 nodes, we found four 11-cliques, five 10 cliques, and twenty-eight, 9 cliques, forty-four 8-cliques and much higher numbers of non-contained, lower order cliques were detected. Of the 197 nodes, 23 of them turned to be highly cited researchers themselves. We also found evidence of citation clique activity among these 23 highly cited researchers. We believe that we have sufficient preliminary evidence for further investigation of plausible collusion. The data will also enable us to explore additional ties such as institutional affiliation, country, and publication outlets. More importantly, our work will allow us to develop detailed citation profiles for researchers beyond simple metrics such as citation counts and h-indices.

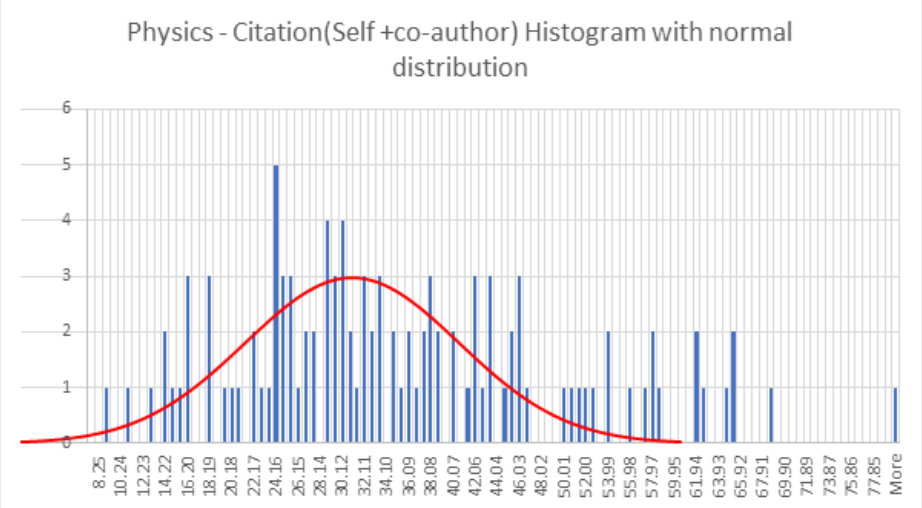

Fig. 1. Frequency Distribution of Self and co-author citation percentage total for 108 highly cited Physics researchers for the year 2016

References

1. Gilbert G.N.: Referencing as Persuasion. Notes and Letters, Social Studies of Science, vol. 7, 113-122 (1977).
2. Web of Science: Look up to the Brightest Stars: Introducing 2017's Highly Cited Researchers. Clarivate Analytics (2016).
3. ScienceDirect: RETRACTED: Impacts of sensor node distributions on coverage in sensor networks, http://www.sciencedirect.com/science/article/pii/S0743731511000864.