How are academic articles cited over time?

Chun Guo, Indiana University Bloomington
Staša Milojević, Indiana University Bloomington
Xiaozhong Liu, Indiana University Bloomington

Abstract
Few studies have focused on understanding the changing relationship between the cited work and the works that cite it. In this study we use publications from the ACM Digital Library published between 1980 and 1989 to follow their citation patterns over the thirty-year period (till 2010). We focus on how these trends differ for articles that are highly cited with respect to those that are not. The analyses are based on pairwise title similarity and similarity measures computed from the heterogeneous paper graph, including papers, authors, venues, and topics as nodes. We find that in general as the time passes the citing papers become more dissimilar from the cited paper. Furthermore, highly cited papers get cited by topically more distant papers in all time periods. However, they tend to share venues more than the other groups. In addition, they are less cited by collaborators than less cited papers.
Keywords: citations; informetrics; scientometrics
Citation: Guo, C., Milojevic, S., Liu, X. (2015). How are academic articles cited over time?. In *iConference 2015 Proceedings*.

Contact: chunguo@indiana.edu, smilojev@indiana.edu, liu237@indiana.edu

1 Introduction

In metrics-driven world of science evaluation, citations have played a major, and occasionally controversial role. Citations have been extensively used to evaluate individual scientists, journals, departments, universities, disciplines, and scientific output of entire countries (Moed, 2010). Numerous researchers have worked on developing theories of citation (e.g., Cronin, 1981; Gilbert, 1977; Kaplan, 1965; Small, 1978). Using a wide variety of methods and approaches researchers have worked on understanding citation behavior, what citations measure, and factors that influence it. Studies of citation have entered a new phase with the increased access to *full text* in electronic format. Researchers started utilizing all elements of journal articles to examine citation content and context (e.g., Liu, Zhang, & Guo, 2013; Zhang, Ding, & Milojević, 2013). For example, some studies utilized the techniques of text mining and natural language processing to automatically characterize the contributions of a cited paper to the citing one (e.g., Jochim & Schütze, 2012; Teufel, Siddharthan, & Tidhar, 2006). However, few studies have explored the changes in the character of citations made to scholarly publications over time.

In this poster, we present initial results of a study that explores the relationship between citing and cited works over a period of several decades. Furthermore, we focus on how these trends differ for articles that are highly cited with respect to those that are not. For example, we address a question of whether the articles that cite some highly popular article become increasingly dissimilar from the cited work. In the study we consider different characteristics of citing and cited documents, such as their authors, venues, and topics. We characterize aggregated relationships between citing and cited articles in terms of pairwise textual similarity and global relatedness through heterogeneous network.

2 Methods

Data on both the cited and the citing articles comes from The Association for Computing Machinery (ACM) Digital Library. We select as *cited* papers those published over a ten-year period between 1980 and 1989. The publication window needs to be several times shorter than the intervening time period in order to ensure similar citation window for all cited articles. These 11,817 papers were cited by 100,459 papers published through year 2010. We split cited papers into three groups: highly cited (cited more than 41 times), cited (cited 13 to 41 times), and rarely cited (cited less than 13 times) (Table 1). Thresholds were chosen such that each group had roughly the same number of citations, thus ensuring that the

	Citation counts	Number of papers in the group	Total number of citations
Highly cited group	42-443	403	33675
Cited group	13-41	1491	33916
Rarely cited group	1-12	9923	32868

Table 1. ACM dataset used and the division into highly cited, cited, and rarely cited group.

accuracy of the results in each group will be comparable. For the analysis we aggregated citations for each group of articles by the citing year and for each citation feature we derive the average value.

Once we have divided papers into groups, for each group we analyze citation patterns over time and the patterns as related of topical similarity, citation by collaborators, usage of similar knowledge base, and usage of similar publication venues. We analyze topical similarity using two approaches. The first is based on the pairwise title similarity between the citing and the cited document measured using cosine similarity (Salton & McGill, 1983). The second uses global relatedness between the cited and citing documents based on the keywords on the paths that connect these two papers in a global heterogeneous graph.

We have created a global heterogeneous graph for the whole ACM database with four kinds of nodes: Paper (P), Author (A), Keyword labeled Topic (K), and Venue (V). Thus, citing and cited papers are positioned (as paper vertices) in the graph, and the relationship between them is characterized by a number of paths on the graph between the nodes pair. The graph contains seven types of edges that depict the relations between pairs of nodes: P→A (written by), A→A (co-author), P→K (relevant for), P←K (contribute to), P→V (published at), P→P (cite), V←K (contribute to).

In this study, in order to characterize the relationship between citing and cited papers we examined the global relatedness derived from four paths: PAAP, PVKVP, PPP, and PKP_relevant. The list of all the factors we have used to characterize the relationship between citing and cited papers and their descriptions are provided in Table 2. For each path we calculate the random walk probability between P_{citing} and P_{cited} nodes (citing and cited papers). If citing and cited papers are closely related (given a path), the random walk probability between them is high.

3 Results and Conclusion

The results of analyses are shown in Figure 1, as different citation trends for each of the three groups of articles (highly cited, cited, and rarely cited) over the period of three decades. Upper left panel shows that publications have similar overall trends in the number of received citations. Namely, all groups experience the peak in the total number of citations by the end of the first decade, to be followed by a decline in the following decade, and the plateau in the third decade. The decline experienced in the second decade is the steepest for the rarely cited papers and the least steep for the highly cited ones. The constant rate in the third decade is the higher for the highly cited papers than for the other two groups, even though there are many times fewer of them.

To determine if the citing papers cover the same topics as the cited papers and if and how that changes over time, we utilized two approaches. The first one, based on title similarity (middle left panel), shows that all three groups follow similar trends, with similarity between citing and cited work decreasing over time. This may be the result of the overall shift in focus as the time passes. However, the highly cited papers are the least similar (in terms of terminology used) to the papers that cite them (at all periods) and the rarely cited papers exhibit the highest level of similarity. Possible reasons for this behavior to be examined further are: (a) rarely cited papers are more often self-cited, i.e. written by the same author(s) whose topics stay similar, and (b) highly cited papers are more appealing for the more varied audiences, producing lower levels of similarity. The second measure of topical similarity we used comes from the

Factor	Abbreviation	Description
Title similarity	Title similarity	Cosine similarity between titles of the citing and cited paper.
$P_{citing} \rightarrow A \rightarrow A \leftarrow P_{cited}$	path PAAP	The authors of both citing paper and cited paper have collaborated with each other.
$P_{citing} \rightarrow V \leftarrow K \rightarrow V \leftarrow P_{cited}$	path PVKVP	The venues, where the citing paper and cited paper published, contribute to the same topics.
$P_{citing} \rightarrow P \leftarrow P_{cited}$	path PPP	Citing paper and cited paper cite the same papers.
$P_{citing} \rightarrow K \leftarrow P_{cited}$ (relevant edge)	path PKP_relevant	Citing paper and cited paper are relevant to the same topics.

Table 2. List of factors used to examine the relatedness between the citing and cited papers.

general heterogeneous graph in which we examine the extent to which cited and citing papers share topics (identified as either author-provided keywords or inferred from topic modeling). We observe similar decreasing trends (upper right panel) over time as with the title similarities in all three groups. However, the highly cited papers exhibit different overall trend than the other two groups. Namely, in the first decade their relevance increases to be followed by a decrease in the following two decades, while relevance in other two groups decreases throughout the thirty-year period.

Another measure of topical similarity comes from the comparison of venues in which the cited and citing paper have been published (lower right panel). We see that three groups have different trends. Highly cited papers have the highest similarity and they are initially cited in more similar venues than later, while rarely cited papers have the lowest, but for them the venue similarity increases, i.e., they are becoming more isolated. To further understand these trends we need to examine other factors possibly at play, such as the impact factor of the venue, its breadth and/or specificity.

When it comes to sharing knowledge bases, namely citing the same body of literature (middle right panel), all three groups show decrease over time. It is interesting that highly cited papers share the least in all time periods and the rarely cited ones share the most. The general decreasing trend would be expected as the result of time, because more and more papers become available to cite.

Finally, when it comes to papers being cited by collaborators, (lower left panel) all groups exhibit the same trends of being more cited by the collaborators in the beginning with gradual decrease over time. Rarely cited and moderately cited papers are more cited by their author(s)' collaborators in the very beginning than the highly-cited ones, perhaps owing to self-citation.

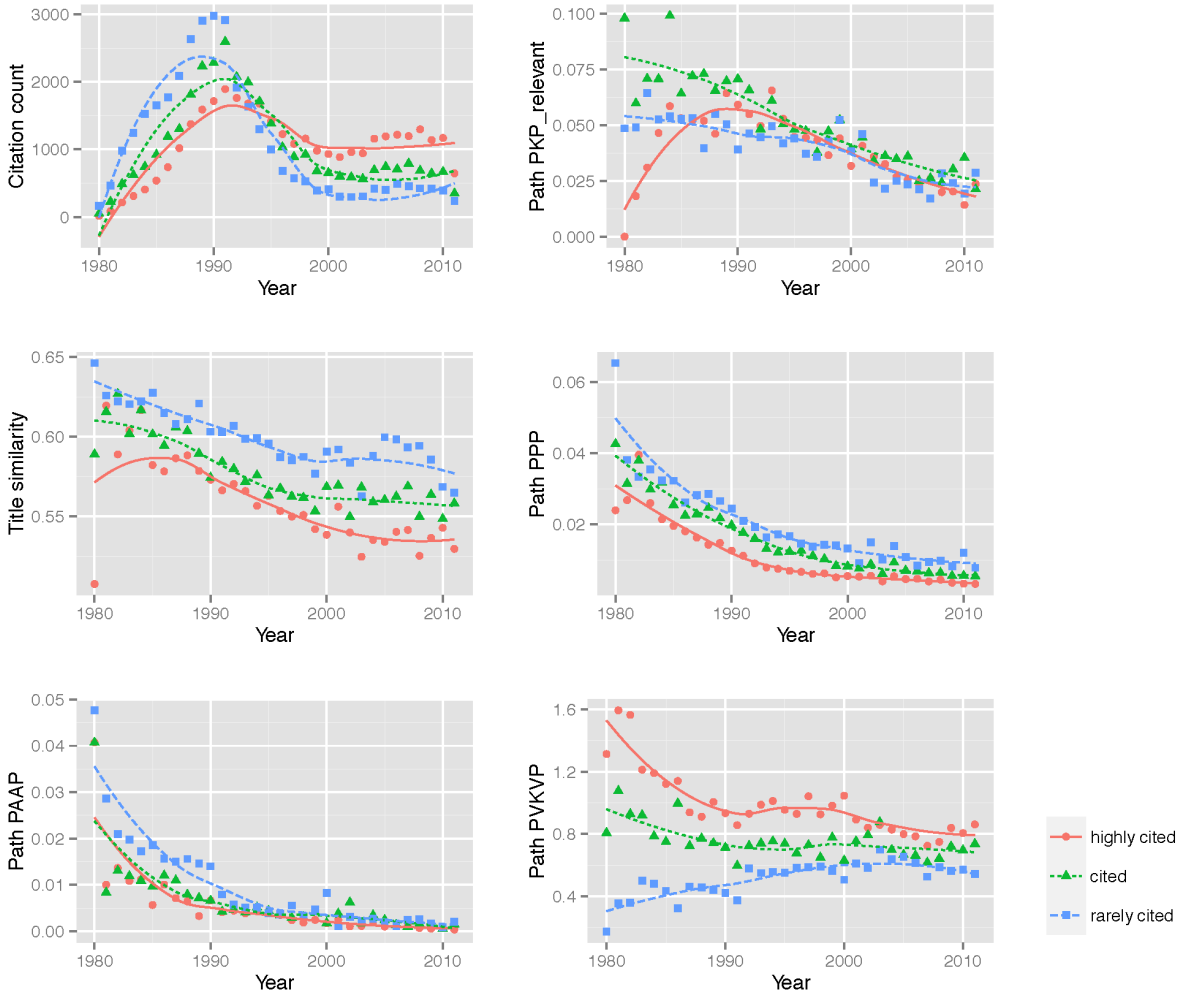

Figure 1. Factors affecting the citation patterns for highly cited, cited, and rarely cited papers in the ACM Digital Library published 1980-1989.

The initial results presented here are very promising. In further research we plan to use other time periods, other disciplines, and additional analysis to identify possible driving mechanisms for the observed behaviors. We also plan to expand the initial group of factors we were observing.

References

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation, 37*(1), 16-24.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science, 7*(1), 113-122.

Jochim, C., & Schütze, H. (2012). *Towards a generic and flexible citation classifier based on a faceted classification scheme*. Paper presented at the Proceedings of the 2012 International Conference on Computational Linguistics, Mumbai, India.

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation, 16*(3), 179-184.

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology, 64*(9), 1852-1863.

Moed, H. F. (2010). *Citation analysis in research evaluation*. Dordrecht: Springer.

Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book Company.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327-340.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). *Automatic classification of citation function*. Paper presented at the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of ctiation content. *Journal of the American Society for Information Science and Technology, 64*(7), 1490-1503.

Table of Figures

Table of Tables