References of References: How Far is the Knowledge Ancestry

Chao Min\textsuperscript{1}, Yi Bu\textsuperscript{2}, and Tao Han\textsuperscript{3}

\textsuperscript{1}mc@nju.edu.cn, \textsuperscript{2}hantaoer@foxmail.com, \textsuperscript{3}buyi@pku.edu.cn

School of Information Management, Nanjing University, Nanjing (China)

Department of Information Management, Peking University, Beijing (China)

Abstract
Scientometrics studies have extended from direct citations to high-order citations, as simple citation count is found to tell only part of the story regarding scientific impact. This extension is deemed to be beneficial in scenarios like research evaluation, science history modelling, and information retrieval. In contrast to citations of citations (forward citation generations), references of references (backward citation generations) as another side of high-order citations, is relatively less explored. We adopt a series of metrics for measuring the unfolding of backward citations of a focus paper, tracing back to its knowledge ancestors generation by generation. Two sub-fields in Physics are subject to such analysis on a large-scale citation network. Preliminary results show that backward citation generations bear some resemblance to forward ones in the macroscopic aspect, but they do behave differently in several specific details. Citations more than one generation away are found to be still relevant to the focus paper, from either a forward or backward perspective. Yet, backward citation generations are generally smaller in the size of networks but higher in topic relevance to the paper of interest. This is implicational for recommending citations in tasks of searching related literature but further research is needed regarding this question.

Keywords
Reference of reference, backward citation, citation of citation, forward citation, citation generation

Introduction
Simple citation count has long before been recognized its inefficiency in estimating the impact of scientific works, as it may tend to favour applied research but undervalue basic science, bias towards a crowded field in comparison to a new field (Margolis, 1967). Among studies in search of measures on indirect scientific impact, Hu & Rouseau (2011) have proposed the notion of backward citation generations (references of references) and forward citation generations (citations of citations). Forward citation generations were investigated in virtue of a data structure called “citation cascade” in prior studies (Min, Chen & Yan et al., in press; Min, Sun & Ding, 2017). Backward citation generations, however, are relatively underexplored. In this study, we examine the characteristics of backward citation generations on a large real citation network, aiming to find answers to the following questions:

- What characteristics do references of references exhibit along different reference generations? and
- What differences and similarities can we find between references of references and citations of citations?
Related Work

Backward and forward citation generations

Citation is a relatively ambiguous concept: For a specific focus paper, one may not know a “citation” is whether another paper that it cites or another paper that cites it. To differentiate, usually the former is called a “reference” and the latter is called a “citation”. Hu, Rousseau & Chen (2011) adopted a consistent terminology by naming “reference” as backward citation and “citation” as forward citation. This terminology unifies the naming of citation in the timeline and raises the interesting concepts of backward and forward citation generations. Hu et al (2011) defined the focus paper as a zero-generation paper, thus obtaining backward citation generations by moving back in time and forward citation generations by moving forward in time. Fragkiadaki & Evangelidis (2014) examined various definitions of citation generations in detail and discussed a number of indirect bibliometric indicators. Both Hu et al (2011) and Fragkiadaki et al (2014) presented specific mathematical formulas for citation generations. Recently, forward citation generations seem to have received more attention than backward citation generations. Min, Sun & Ding (2017), for instance, observed citation structure from the angle of information cascade and quantified with two structural-wise metrics. What they term as “citation cascades” is in essence a linked structure of forward citation generations. Chen (2018) also investigated this cascading citation expansion process which has potential to extend the coverage of relevant publications by citation links. To model the evolution of follow-up works and their interdependence, Mohapatra et al (2019) define a data structure called Influence Dispersion Tree (IDT). Based on this data structure, they propose a suite of metrics to measure the influence of a paper.

The applications of citation generations

Viewing citation data from the perspective of propagating generations has been applied in several studies. The first application one may think of is measuring the total influence in terms of a single paper, taking both direct and indirect citations into account. Rousseau (1987) did so from a backward citation perspective by using Gozinto Theorem and he claimed that this approach provides insight in the stream of ideas that guided an author to the results in a given paper. della Briotta Parolo et al (2020) formalized the question of how much influence a seed publication s has on another publication j and call it persistent influence. They took into account all the citation passages between the two publications and calculated persistent influence in a large real citation network; they concluded that Nobel prize papers have higher ranks of persistent influence than direct citation counts. Another stream of application is the investigation of citation chains. In their work, Frandsen & Nicolaisen (2013) documented a ripple effect of the Nobel Prize delivered through citation chains to mathematician Robert J. Aumann’s scientific works as well as to those works’ references. They showed that the award of the Nobel Prize affected not only the citations to Aumann’s works but also the citations to those works’ references. The effect of citation chain reactions received follow-up attention from researchers (Farys & Wolbring, 2017; Frandsen & Nicolaisen, 2017). In addition, the concept of citation chain has also been applied in such tasks as search path counting (e.g., Jiang & Zhuge, 2019) and main path analysis (e.g., Jiang, Zhu & Chen, 2019).

Methods and Data

Adopting a similar approach to a data structure called “citation cascades” in Min, Chen & Yan et al. (in press) and Min, Sun & Ding (2017), we construct a data structure termed “reference cascades” from the perspective of backward citation. For each paper i, we define its reference cascade RC_i as a directed graph that includes i, all references of i, second-order references (i.e., references of references) of i, third-order references, …, and nth-order references until no
further references (called “pioneer papers”; they are the knowledge ancestry of $i$) can be identified; as for edges, $R_{Ci}$ includes all citing relations within these publications. Thus, we know that the only node whose in-degree equals zero is the root paper $i$, and that nodes whose out-degree equals zero represent pioneer papers.

For each reference cascade, we define two metrics:

1. **Depth**: A certain node’s depth is the shortest path length from the root to that node within the reference cascade. The maximum value of any node’s depth is the depth of the reference cascade, which reflects how deep the reference cascade can reach.
   
2. **Size**: Size is defined as the total number of nodes in a reference cascade. The size may dramatically increase as the depth of a reference cascade increases.

For each generation of a certain reference cascade, we define:

1. **Width**: The number of nodes in a reference generation is called the generation’s width. The maximum width of any generation is taken as the width of the reference cascade, which reflects how wide the reference cascade can extend.

2. **Topic relevance**: Two papers’ topic relevance is defined as the similarity of their topics, measured in the dataset here by the Jaccard similarity of their PACS codes. We then stimulate the topic relevance of a certain generation as the average topic relevance between the root paper and all papers in this generation of the reference cascade.

In this study, we adopt the 2013 version American Physical Society (APS) dataset that covers 450K physics publications and their 6M citing relations. Publications after 1975 are labelled with one or more PACS codes to identify its sub-field. We particularly select publications whose PACS code starting with 2 (Nuclear physics; 19516 papers) or 5 (Physics of gases, plasmas, and electric discharges; 5113 papers) because the topic relevance distributions among different generations in their citation cascades (Min, Chen & Yan et al., in press) are more representative.

**Results**

**Depth distribution.** Nearly 2500 (1/8) and 1000 (1/5) papers respectively in PACS 2 and PACS 5 have no knowledge ancestry within American Physical Society journals (Figure 1). We check the full texts of these papers and find that they are not necessarily very old papers who have few prior knowledge to cite, but quite often they have cited other sources of knowledge outside the APS literature. The remaining papers more or less inherited knowledge from APS. The earliest ancestors in PACS 2 and PACS 5 date back to 60 and 54 generations away, respectively. In PACS 5, as the generation traces backward, a decreasing number of papers can reach to the old time, but there is a rise of papers between 35-45 generations. For PACS 2, the trend is a bit complicated as there are three noteworthy stages on the curve, firstly a decreasing stage between 1-18 generations, next an increasing-and-then-decreasing wave between 19-36 generations, and finally a repeating wave afterwards.

**Size distribution.** The size of reference cascades reaches a maximum of 115,384 in PACS 2, and 92,661 in PACS 5. The distribution of size (Figure 2 Left) is apparently a long-tail distribution for both two Physics sub-fields, with most papers having small sizes (usually within several hundred) and others scattering over a relatively extensive range (up to hundreds of thousands). When the scale is zoomed in within 200 (Figure 2 Right), the long-tail shape is still prominent. An interesting finding in the left panel of Figure 2 is that there seems to exist a vacuum zone between size of 5-20K, where we observe no reference cascades in this range, but
there do exist papers with either lower or higher sizes. This is in line with the observation on “citation cascades” in a prior study (Min, Chen & Yan et al., in press).

![Distribution of cascade depth](image1.png)

**Figure 1** Paper distribution of cascade depth

![Distribution of cascade size](image2.png)

![Distribution of cascade size](image3.png)

**Figure 2** Paper distribution of cascade size. Left: log scale. Right: normal scale between 0 and 200.

**Width distribution.** To explore how generation width (the number of papers in a generation) grows as references reach into prior knowledge space, we select the median value of generation width among all cascades (Figure 3 Left). This gives a macro perspective on the general pattern of generation width growth. Both PACS 2 and PACS 5 show a similar trend on which the width increases from the first generation and then starts to decrease after reaching the peak. The peak appears at the 12th generation for PACS 2 and 4 generations later for PACS 5. Interestingly, both width curves dip at the 27th generation and then turn up for a bit afterwards.

**Topic relevance.** The overall topic relevance between a focus paper and a generation of prior papers is shown in Figure 3 (Right). The first generation (that is, direct references), without doubt, presents a rather high relevance with the focus paper, nearly 0.7 for PACS 5 and 0.6 for PACS 2. Further generations, however, are not entirely irrelevant. Actually, the relevance is as high as 0.56 for PACS 5 and 0.46 for PACS 2 in the second generation. The third generation still has a relevance of 0.47 and 0.39 for PACS 5 and PACS 2, respectively. Furthermore, in both Physics sub-fields, up until the sixth generation, the “knowledge ancestors” (i.e., the pioneering papers) still keep relevant with the focus paper above the level of 0.2. The relevance decreases sharply and stabilizes around the sixteenth generation, after which the “ancestry” relationship seems rather weak.
Discussion

Here we compare the structural characteristics of reference cascades with those of citation cascades, based on the results in this study and from a prior study (Min, Chen & Yan et al., in press). Our main observation is that reference cascades and citation cascades are similar in the macroscopic aspect, but they do have different features in specific details.

On the one hand, the differences are subtle, yet they do exist. First, most (actually almost all) papers have knowledge ancestors (references) that can be traced back, but only a proportion of papers will have knowledge descendants (forward citations). In addition, there are still many papers who can reach rather deep reference generations (e.g., PACS 2), but for most papers (80%) the citation depth is within only 20. Second, the size of reference cascade is much smaller than that of citation cascade based on the data sample in this study. Also, the latter’s long-tail distribution is fatter than the former’s, indicating a small head but fat tail in the backward-forward citation generation structure. Third, compared with citation generations at the same distance, reference generations seem to be a bit more relevant to the focus paper. This is particularly evident in the first several generations, but further research is needed to compare the relevance difference in the level of the whole APS dataset.

On the other hand, there are also similarities between reference cascades and citation cascades in terms of depth, size, width, and topic relevance. Firstly, both can extend (regardless of backward or forward) citations to generations as far as 60. Secondly, both size distributions exhibit a long-tail shape. And on both distributions, there is a vacuum zone without any papers but there are papers either before or after the zone. Thirdly, both width distributions show a similar increase-and-then-decrease trend. Yet more specifics of width distribution in individual paper level need further examination. Lastly, both topic relevance curves present a decrease-and-then-stabilize trend, resembling a negative exponential or pow law distribution.

Conclusion

This paper reports the results of a research-in-progress study on citation generations. Different from many existing studies regarding forward citation generations, we trace backward into knowledge ancestry, or what we call references of references, of an individual paper. Two sets of papers from two sub-fields in Physics, as well as their backward citation generations are extracted for analysis. Preliminary results show that backward citation generations bear some resemblance to forward ones in the macroscopic aspect, but they do behave differently in several specific details. There seems to exist an asymmetric structure between an individual paper’s backward and forward citation generations. This is reflected in the observation that
backward citation generations are generally smaller in cascade size but higher in relevance to the paper of interest, compared with forward citation generations. Yet citations more than one generation away are found to be still relevant to the focus paper, from either a forward or backward perspective. The topic relevance is not low in quite some cases, indicating that relevant citations are not merely within the first generation. This is implicational for recommending citations in tasks of searching related literature either manually or with automatic algorithms.

Since within six generations, backward citations are relevant to the paper of interest in the level of no lower than 0.2, a higher-generation approach would have potential over a direct-citation approach in finding more citations that may be of one’s interest. However, further research is needed to more rigorously validate and extend the findings in this study. Also, easy-to-use scientometrics indicators might be needed to measure the high-generation citation structures.

Acknowledgments
Financial support from the National Science Foundation of China (NSFC No. 71904081, No. 71874077) and Humanities and Social Sciences Program of the Ministry of Education (No. 19YJC870017) and is gratefully acknowledged.

References
Chen, C. (2018). Cascading citation expansion. Journal of Information Science Theory and Practice, 6(2), 6-23.
della Briotta Parolo, P., Kujala, R., Kaski, K., & Kivelä, M. (2020). Tracking the cumulative knowledge spreading in a comprehensive citation network. Physical Review Research, 2(1), 013181.
Farys, R., & Wolbring, T. (2017). Matched control groups for modeling events in citation data: An illustration of Nobel Prize effects in citation networks. Journal of the Association for Information Science and Technology, 68(9), 2201-2210.
Fragkiadaki, E., & Evangelidis, G. (2014). Review of the indirect citations paradigm: theory and practice of the assessment of papers, authors and journals. Scientometrics, 99(2), 261-288.
Frandsen, T. F., & Nicolaisen, J. (2013). The ripple effect: Citation chain reactions of a Nobel prize. Journal of the American Society for Information Science and Technology, 64(3), 437-447.
Frandsen, T. F., & Nicolaisen, J. (2017). Rejoinder: Noble prize effects in citation networks. Journal of the Association for Information Science and Technology, 68(12), 2844-2845.
Hu, X., Rousseau, R., & Chen, J. (2011). On the definition of forward and backward citation generations. Journal of Informetrics, 5(1), 27-36.
Jiang, X., Zhu, X., & Chen, J. (2019). Main path analysis on cyclic citation networks. Journal of the Association for Information Science and Technology.
Jiang, X., & Zhuge, H. (2019). Forward search path count as an alternative indirect citation impact indicator. Journal of Informetrics, 13(4), 100977.
Margolis, J. (1967). Citation indexing and evaluation of scientific papers. Science, 155(3767), 1213-1219.
Min, C., Chen, Q., Yan, E., Bu, Y., & Sun, J. (In press). Citation Cascade and the Evolution of Topic Relevance. Journal of the Association for Information Science and Technology. https://arxiv.org/abs/2004.12275
Min, C., Sun, J., & Ding, Y. (2017). Quantifying the evolution of citation cascades. Proceedings of the Association for Information Science and Technology, 54(1), 761-763.
Mohapatra, D., Maiti, A., Bhatia, S., & Chakraborty, T. (2019, June). Go wide, go deep: quantifying the impact of scientific papers through influence dispersion trees. In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 305-314). IEEE.
Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influences on a scientific publication. Scientometrics, 11(3-4), 217-229.