

Loops in publication citation networks

Journal of Information Science
2020, Vol. 46(6) 837–848
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0165551519871826
journals.sagepub.com/home/jis


Yi Bu 

Department of Information Management, Peking University, Beijing, China

Yong Huang

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, China

Wei Lu 

Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, China

Abstract

Traditionally, publication citation networks are regarded as acyclic, that is, no loops in the network as an earlier published article cannot cite a later published article. However, due to the accessibility of pre-print versions of articles, there might be some loops in a publication citation network. This article presents a descriptive statistic on loops in publication citation networks of computer science and physics by employing a network-based indicator, namely, strongly connected component (SCC). By employing computer science and physics disciplines publications from the Web of Science database as examples, this article examines the count of loops, how the count changes over time and how the count relates to the published year difference between publications within the loop in the citation network. Some common structural patterns are also extracted and analysed; we observe that the two disciplines share the most frequent patterns though there exist some minor differences. Moreover, we find that self-citations in terms of authors, authors' institutions and journals contribute to the formation of loops in publication citation networks.

Keywords

Bibliometrics; citation analysis; citation network; informetrics; library and information science; quantitative science studies; scientometrics

1. Introduction

Citation network analysis has long been regarded as a useful strategy to evaluate the scientific literature, scientists, journals and institutions in bibliometrics and scientometrics [1–6]. Citation networks can be constructed in various levels, such as publications [7], authors [8–11] and journals [12]. In a publication citation network, for instance, a node represents a scientific publication, and an edge from a node (A) to another (B) indicates that A has ever cited B (i.e. B occurs in the reference list of A). In a long time, publication citation networks have been regarded as an *acyclic* directed network.¹ That is to say, there are no loops in the network — no nodes connected in a closed chain. This is reasonable from a traditional wisdom: the fact that A cites B hints that B was published prior to A, because a later-published article cannot be cited by an earlier publication.

However, publication citation networks are no longer acyclic in the current era. Instead, there are many loops in citation networks — two publications might cite each other based on the bibliographic record (i.e. A cited B and B cited A), or A cited B, B cited C and C cited A. This, as argued aforementioned, was impossible previously, but in the current publishing system, it makes sense, partly because of the popularity of pre-print platforms [13–15]. For instance, a physics publication (say C) might be under review by a certain journal now, but its authors might have uploaded it to some pre-print platforms, such as *arXiv*. At the same time, the authors of another physics publication (say D) that is typically

Corresponding author:

Wei Lu, Information Retrieval and Knowledge Mining Laboratory, School of Information Management, Wuhan University, Wuhan 430072, China.
Email: weilu@whu.edu.cn

similar to C noticed and cited C. After a short while, D was submitted to another physics journal and was also uploaded to *arXiv*. When C was revised after a round of revision, its authors noticed and cited D as their topical relatedness is great. In this way, C and D cite each other in record, although one of them might be published earlier than the other one, and, therefore, from a retrospective view, a loop occurs in the citation network. As pointed out in Lin and Chalupsky [15], ‘one journal might have a very long revising period and during that period other people can access the previous version’ (p. 176); of course, during this period, research scholars can also cite the previous version. Another possible reason why loops occur in publication citation networks is the time lag between the point of time when an article is available online and that when it is assigned to a certain issue of a journal.

The phenomenon of loops in publication citation networks makes some bibliometric analyses unfeasible. To deal with it, bibliometricians and scientometricians tended to simply delete the loops and all publications in the loop if they find. For instance, in Huang et al. [4], when using the Microsoft Academic Graph (MAG) data set, they found that there are some loops within their built ego-centred citation networks. When this occurs, they simply removed the whole network (including the publications and the citing loop(s)) from their data set. Another potential strategy is that one of the two mutual links is removed (normally the citing relationships from the older to the younger articles are removed based on the official in-record published date) when scientometricians find that the two publications cite each other.

Loops have been discussed outside publication citation context for quite a long time. In metabolic networks, for instance, Bilke and Peterson [16] found that there are many loops empirically. In the general social network analysis field, the software Pajek embeds loop analysis as a function in it [17], indicating the universality of loops in networks and graphs. The prevalence of loops in patent citation networks was also mentioned by Madani et al. [18], in which three different models differentiated by retaining or removing loops were proposed to predict future citations of patents. Yet, a systematic investigation of loops in publication citation networks is missing.

Although rare [19], the number of citation loops might increase over years due to the popularity of pre-print platforms. In this article, we aim to understand loops in publication citation networks by addressing the following research questions:

1. How many loops are there in publication citation networks of different disciplines, and how does the number change over years?
2. What are the common structural patterns of loops in publication citation networks?
3. Are these loops caused by self-citations?

This article is outlined as follows. We first introduce our empirical data set as well as the methods employed. We then present our results and discuss the findings. Finally, we summarise the article, demonstrate the implications and illustrate potential future work.

2. Methodology

2.1. Data

We employ the Web of Science (WoS) database to explore our research questions. The WoS database used in this study is hosted by Indiana University Network Science Institute (IUNI) containing the complete set of Clarivate Analytics Web of Science Core Collection in XML format (Web of Knowledge version 5). This data set comprises of 63,590,916 publications indexed by Science Citation Index Expanded (SCI-E), Social Science Citation Index (SSCI), Art and Humanities Citations Index (A&HCI), books and conference proceedings in all disciplines. These publications cover the years 1900 through 2016. Moreover, there are 1,227,494,925 citing relationships among these publications.

In this study, we select computer science and physics as two empirical disciplines to present natural science and engineering, respectively. To this end, we examine the ‘subject’ field of each article – if this field of an article contains ‘computer science’, this article will be categorised as a computer science publication (annotated as ‘WOS-CS’); if it contains ‘physics’, it will be categorised as a physics publication (annotated as ‘WOS-P’). In this way, we select 2,387,985 and 6,267,440 publications in computer science and physics, respectively. We then construct two citation networks for all computer science and for all physics publications. That is to say, each citation network includes all computer science (or physics) publications selected and all citing relationships between these computer science (or physics) publications.

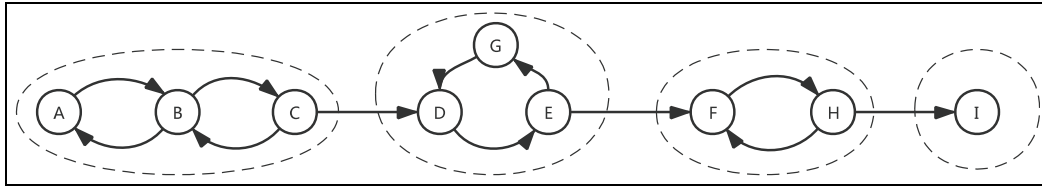


Figure 1. An illustration of strongly connected component (SCC) in a citation network. Nodes represent scientific publications, while edges show citing relationships between two publications.

2.2. Methods

Detecting all loops in citation networks is unfeasible, especially for large loops containing 10+ nodes. Thus, we here employ an alternative way of quantifying loops. In network science, a directed acyclic network is defined as a finite directed network with no direct loops [20]. If there is at least one direct loop in a network, the network is named as directed cyclic network. A directed network is strongly connected if and only if there is a path between all pairs of vertices. A strongly connected component (SCC) of a directed network is a maximal strongly connected sub-network. If a loop exists in a citation network, there must be at least one SCC in citation network. The number of nodes contained in an SCC is named as its size. For instance, in the citation network illustrated in Figure 1, nodes represent publications (e.g. A–I), while edges show citing relationships between two publications (e.g. a link from D to E means that publication D cites E). In this citation network composed of nine publications, there are four SCCs, namely, {A, B, C}, {D, E, G}, {F, H} and {I}, as shown in the four dotted circles in Figure 1, and the sizes of the SCCs are three, three, two and one, respectively. Again, as long as there is at least one loop, there must be an SCC. Yet, note that SCC is not equivalent to loops, because the occurrence of an SCC might indicate more than one loop. For instance, in the SCC {A, B, C}, there are three loops formed by [A and B], [B and C] as well as [A, B and C], respectively. Two real examples of SCCs extracted from WOS-P and WOS-CS are shown in the top and the bottom images of Figure 2, respectively, with their sizes equal to 10.

Apparently, there must be at least one loop in an SCC of a given network. Thus, we can use SCC of a citation network as an indicator to measure the details of loops (e.g. count and network structure) in a publication citation network. In practice, we apply the algorithm in Nuutila and Soisalon-Soininen [21] on our citation networks with Python 2.7 and networkx 1.11 in which a depth-first search strategy is adopted [22].

For a given citing relationship, if the citing publication was published more than 3 years earlier than the cited publication, we delete this citing relationship, as this might result from errors recorded in the database. Also, for the sake of further calculation, we remove all SCCs containing only one publication. For example, in the citation network shown in Figure 1, we will remove {I} and purely keep the remaining three SCCs.

Based on the definition of SCC, we know that a given publication citation network might have multiple SCCs with various structures. We obtain 4011 computer science publications and 58,315 physics publications involved in the corresponding SCCs. The numbers of SCCs in the two disciplines equal 1677 and 25,480, respectively.

3. Results

3.1. Overview

Figures 3 and 4 show some descriptive statistics of loops in publication citation networks of computer science and physics, respectively. In Figures 3(a) and 4(a), the horizontal axis represents the size of SCC, that is, the number of nodes (publications) contained in a certain SCC, while the vertical axis shows the number of SCCs with the corresponding size. From the figures, we can find that most SCCs, regardless of which discipline, feature with a small size while only a few SCCs have a great number of publications inside. Specifically, there are more than 1000 SCCs in computer science featuring a size of two – these are the simplest SCCs. That is to say, these SCCs include two publications mutually citing each other. The number of SCCs with two publications exceeds 20,000 in physics. Nevertheless, there are only 2 and around 20 SCCs whose size is equivalent to 10 in computer science and physics, respectively. Figures 3(b) and 4(b) reveal the relationship between the number of SCCs and the published year difference of the publications contained in SCCs. Here, the published year difference is defined as the difference between the oldest and the youngest publications in an SCC. We observe that in both disciplines, those including articles published in the same year are dominant among all SCCs. Particularly, more than 1000 and 20,000 SCCs contain scientific articles published within the same year (i.e.

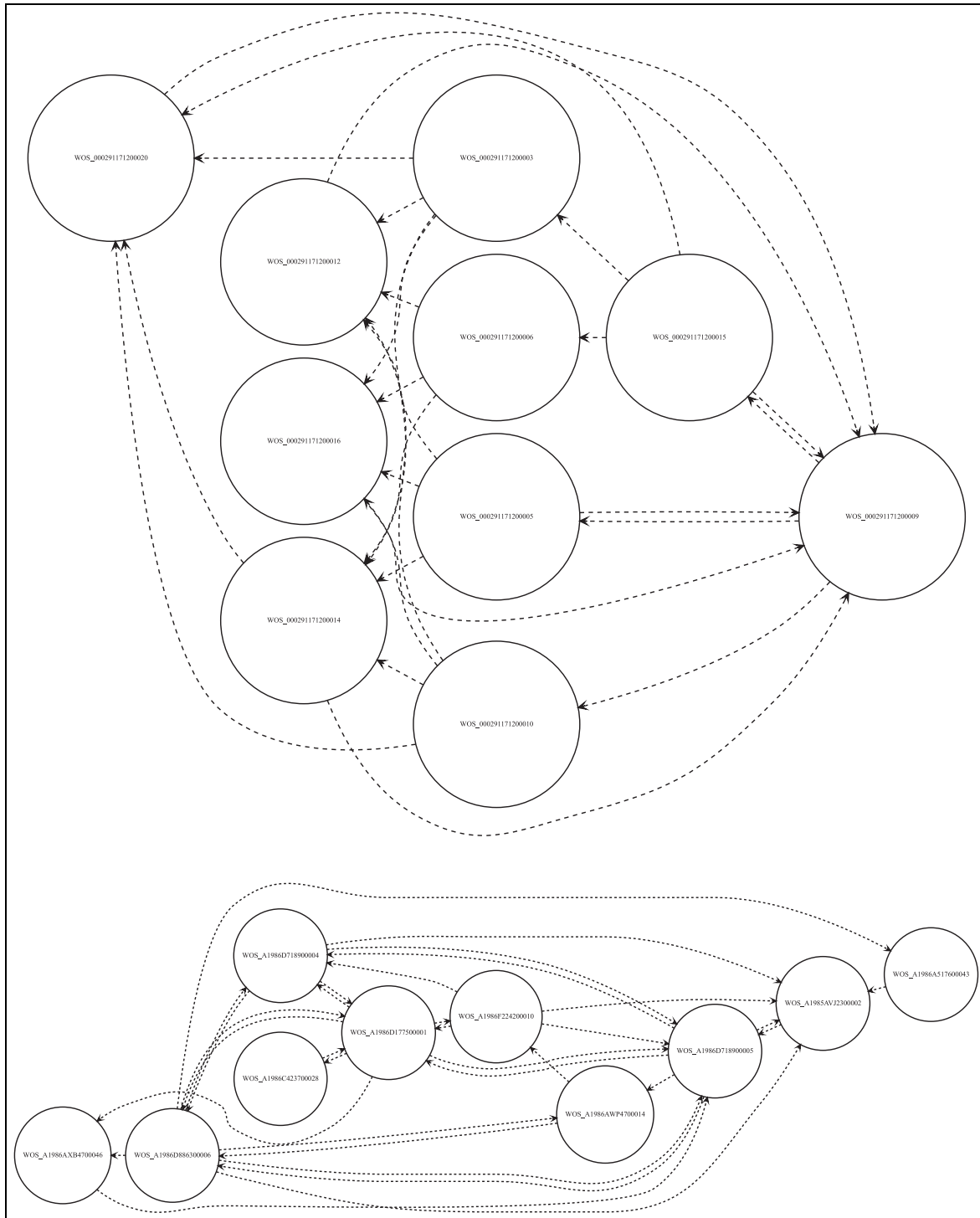


Figure 2. Two real examples extracted from WOS-P (top) and WOS-CS (bottom). Both SCCs have a size of 10. Labels of nodes represent their Web of Science ID.

year difference equals to zero in the figures) in the two disciplines. There are no SCCs in which publications differ in more than 2 years in computer science; that value is equal to five for the physics field.

We also examine how the number of SCCs containing articles published in different years changes over time, as shown in Figures 3(c) and 4(c), where N refers to the published year difference of the oldest and youngest publications

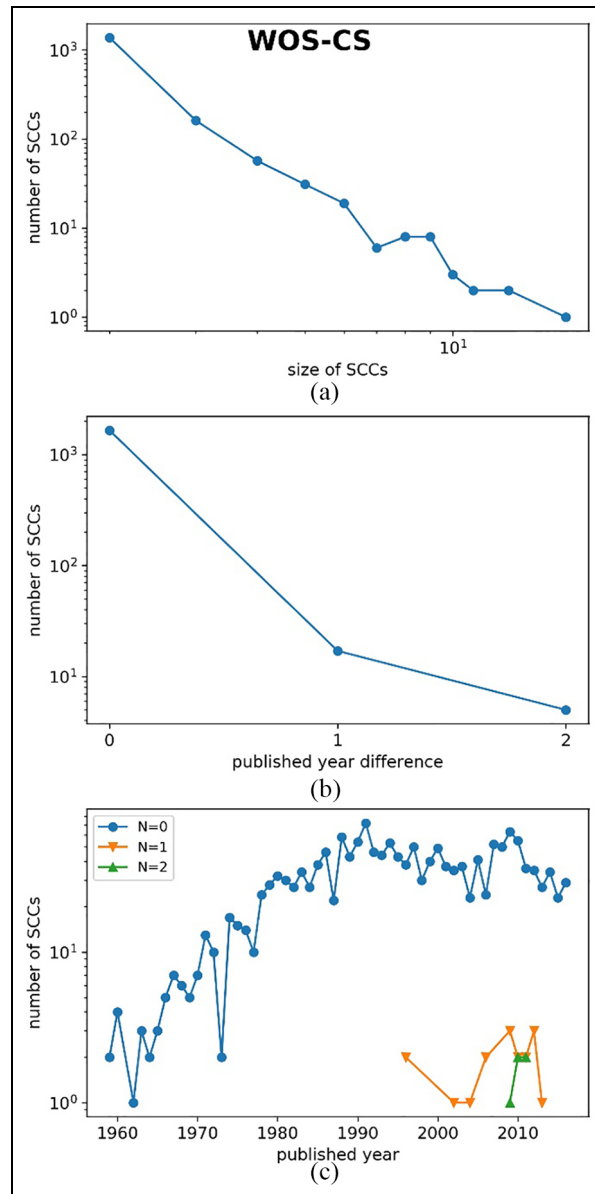









Figure 3. Loops in the citation network of computer science: (a) distribution of the size of strongly connected components (SCCs), (b) time span of the largest year difference between all publications in an SCC and (c) SCC distribution of publications in different years, in which N refers to the published year difference of the oldest and youngest publications in the SCC.

in the SCC – for instance, in both figures, the curve $N = 0$ represents SCCs whose publications were published within the same year based on the WoS records. From Figure 3(c), we observe that publications in SCCs prior to 1990s are almost in the same year and the number of SCCs are very limited in computer science. This attributes to two potential reasons. On one hand, the numbers of publications and citations were limited prior to 1990 in this field; on the other hand, the rare usage of pre-print platforms and limited scholarly communications make it hard for loops to occur in publication citation networks. From Figure 3(c), there are only ~30 SCCs consisting of computer science articles published in 1980. The curves representing $N = 1$ and $N = 2$ does not occur until 1995 and 2008. Even though in recent years, those in the same year ($N = 0$) still dominate all SCCs in terms of counts in the field of computer science. Patterns are similar for those in physics. Specifically, the $N = 1$, $N = 2$, $N = 3$ and $N = 5$ curves first occurred in 1950s, late 1970s, late 1980s and 1990s, respectively, and the numbers of SCCs with $N = 3$ or $N = 5$ are limited, compared with those with $N = 0$ and $N = 1$.

3.2. Patterns of loops

Besides some basic distributions of SCCs as well as time span details, we are also interested in what the common structural patterns of SCCs in the publication citation networks are in two disciplines. In network science, structural patterns are often used to analyse the details of sub-networks and paint a more nuanced picture on the whole network itself, sometimes named as ‘network motifs’. For example, Milo et al. [23] investigated all 13 different three-node connected sub-networks, such as  and .

Tables 1 and 2 present top patterns of SCCs, their frequencies, proportions, published year difference and the details of loops in SCCs. Take computer science as an example: we find that SCCs containing two publications mutually citing each other (i.e., ) occur most frequently (1379 out of 1677 SCCs). Among these SCCs, almost ~1000 feature no published year difference. The second most frequent pattern consists of three publications citing each other (i.e., ) but its frequency is only 54. Obviously, in this pattern, there are three loops with two publications and one loop with three publications. This pattern shows quite close relationships among the three publications. As for physics, the most frequent pattern is also , but the second most frequent pattern is  with two loops with two publications and one loop with three publications. This pattern shows a slightly less close relationship than . Yet, according to Tables 1 and 2, although there are slight differences between frequent patterns, as well as their rankings, in the two disciplines, most patterns are the same with similar ranking.

3.3. Source analysis of loops: a multi-level self-citation investigation

As aforementioned, Chalupsky [13], as well as Lin and Chalupsky [15], pointed out the phenomenon of loops in citation networks mostly results from self-citations, but they did not provide quantitative proofs. In this section, we are investigating this hypothesis based on our empirical data. We know that there are different levels and perspectives of self-citations, such as authors, institutions and journals. For instance, an author-level self-citation indicates that there is at least one shared co-author in the bylines of citing and cited publications. An institution-level self-citation reveals that the sets of institutions of citing and cited authors have a non-empty intersection. There is at least one shared co-author having the same institution in the bylines of citing and cited publications. A journal-level self-citation illustrates that both citing and cited articles are published in the same journal.

To this end, we analyse the relation between self-citations and the formation of loops in publication citation networks. Due to the data quality issue, not all publications contain author-, authors’ institution- and journal-related metadata. The numbers of publications available with different levels of self-citations are shown in Table 3, where we can find that most publications contain author and journal details, but only some have author institution metadata (~72% and ~66% for computer science and physics publications, respectively) in the WoS database.


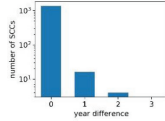
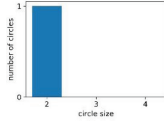

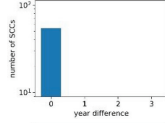
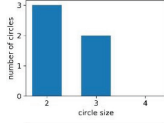

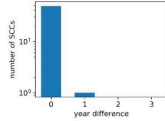
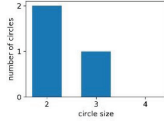

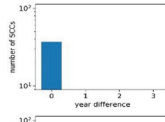
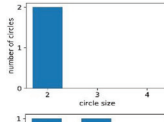

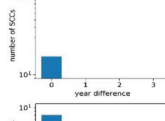
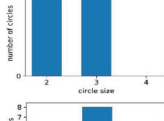

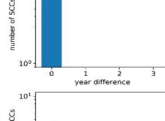
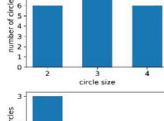

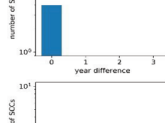
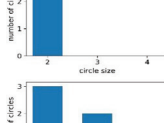

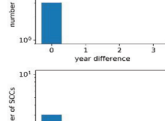
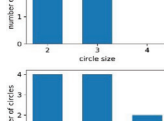

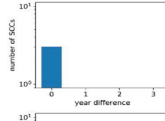
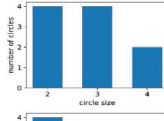

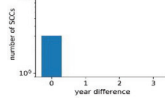
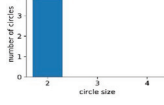
Note that we only compare the institutions on the university level. For instance, if two authors have their institutions in the same university but in different schools, departments, colleges and/or centres, we still regard them as the same institutions. Furthermore, due to the limited number of loops in records, there is no need to disambiguate authors’ names. Instead, for the sake of simplicity in the empirical study, we stipulate that if two authors share the same last name and initial, they are the same person.

Table 4 shows the result of author-level self-citation analysis, where one can find that in both disciplines, nearly half of the loops result from the fact that the citing and cited publications share the first authors. If we loosen the constraint to all authors instead of only first authors, we observe that 70.9% and 80.5% loops come from shared at least one co-author in computer science and physics. Table 5 shows how authors’ institution self-citations influence loops in the citation networks. In both disciplines, more than 4/5 loops were triggered by co-authors in the same institution. Similarly, based on Table 6, we also find that self-citations are probably the reason of loops in the citation networks on the journal level. Our empirical results support the arguments from Chalupsky [13] and Lin and Chalupsky [15] in a quantitative way.

4. Discussion and conclusion

By employing computer science and physics disciplines’ publications from the WoS database as examples, this article investigates loops in publication citation networks. Specifically, this study examines the count of loop, how the count changes over time and how the count relates to the published year difference between publications within the loop. Some common structural patterns are also extracted and analysed; we observe that the two disciplines share the most frequent patterns though there exist some minor differences. Moreover, we find that self-citations in terms of authors, authors’ institutions and journals contribute to the formation of loops in citation networks.

Table 1 Common structural patterns of SCCs in the citation network of computer science.


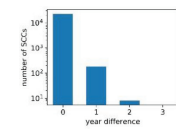
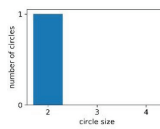

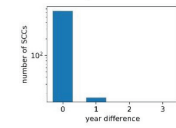
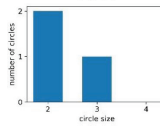

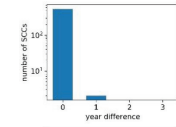
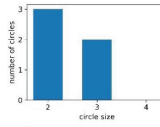

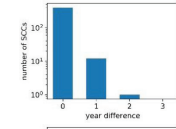
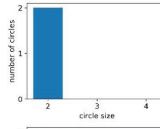

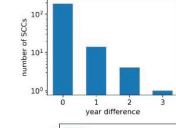
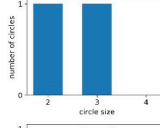

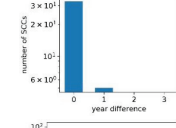
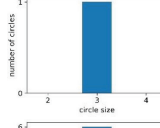

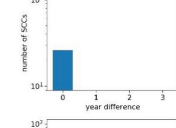
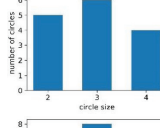

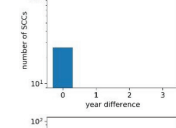
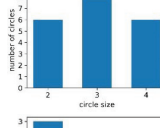

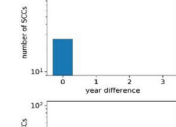
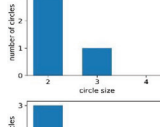

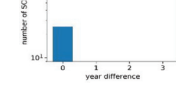
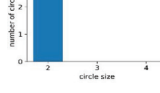
Rank	Pattern	Frequency	Proportion (%)	Size	Publication year difference	Loops contained
1		1379	82.2	2		
2		54	3.2	3		
3		50	3.0	3		
4		37	2.2	3		
5		17	1.0	3		
6		8	0.5	4		
7		4	0.2	4		
8		3	0.2	4		
8		3	0.2	4		
8		3	0.2	5		

SCCs: strongly connected components.

Retaining loops in publication citation networks without removing them might have some influences in bibliometric analyses. For instance, in our previous work [3,4], we considered an ego-centred publication citation network containing direct citations of a publication (say X) and all citing relationships between X's citing publications, named a citing cascade, for understanding direct citations between citing publications by distinguishing three different types of citing publications of X, namely, connectors (defined as citing publications of X that are also cited by other X's citing publications), late endorsers (defined as citing publications of X that also cite connectors of X) and isolate endorsers (defined as citing publications of X that are not cited by any of X's citing publications or cite X's other citing publications). If loops are retained in a citing cascade, it would be difficult to calculate the length of longest paths of the cascade. Yet, for studies not taking into consideration citation network typologies such as counting number of citations, loops do not need to be removed from the networks given the rareness of their occurrence.

'Super-early citing' is the essential reason why loops occur in publication citation networks. Here, 'super-early citing' means that authors cite a publication that has not been officially published. Specifically, there are four types of 'super-

Table 2. Common structural patterns of SCCs in the citation network of physics.

Rank	Pattern	Frequency	Proportion (%)	Size	Published year difference	Loops contained
1		22,584	88.6	2		
2		659	2.6	3		
3		544	2.1	3		
4		409	1.6	3		
5		207	0.8	3		
6		38	0.2	3		
7		26	0.1	4		
7		26	0.1	4		
7		26	0.1	4		
10		25	0.1	4		

SCCs: strongly connected components.

early citing’: (1) citing in-preparation research, (2) citing under-review research and/or (3) citing in-press article (due to the time lag between the point of time when an article is online available and that when it is assigned to a certain issue/chapter of a journal/monograph/conference proceeding). The reasons why the aforementioned three situations occur are as follows: (1) cited authors must have uploaded their manuscript on pre-print platforms, (2) authors in the citing and cited publications probably know each other and/or (3) there are some intersections between citing and cited authors (self-citation).

The investigation on loops in citation networks brings many interesting implications. From the perspective of academic ethnic, the occurrence of these loops probably results from one publication citing another under-review publication, essentially a ‘super-early’ citation. However, under-review publications might have critical drawbacks in the literature review, empirical studies, hypotheses, arguments or conclusions, as they are not peer-reviewed by referees of

Table 3. Number of publications available with different levels of self-citations (authors, institutions and journals).

Data set	WOS-CS	WOS-P
Number of publications in total	4011	58,315
Number of publications with author information	4011	58,314
Number of publications with authors' institution information	2888	38,721
Number of publications with journal information	3846	50,498

WOS-CS : Web of Science-computer science; WOS-P: Web of Science-physics.

Table 4. Author self-citations and loops.

Data set	WOS-CS	WOS-P
Share first authors	44.4%	45.1%
Share non-first authors	26.5%	35.4%
Do not share co-authors	29.1%	19.5%

WOS-CS: Web of Science-computer science; WOS-P: Web of Science-physics.

Table 5. Authors' institution self-citations and loops.

Data set	WOS-CS	WOS-P
Share at least one institution	81.8%	83.2%
Do not share institutions	18.2%	16.8%

WOS-CS: Web of Science-computer science; WOS-P: Web of Science-physics.

Table 6. Journal self-citations and loops.


Data set	WoS-CS	WOS-P
Share journal	97.2%	97.2%
Do not share journal	2.8%	2.8%

WOS-CS: Web of Science-computer science; WOS-P: Web of Science-physics.

journals/conferences. Such citing relationships need to be reconsidered and carefully reviewed by domain experts. Loops in publication citation networks could potentially be used as a starting point of detecting errors and mistakes of research.

Meanwhile, according to the finding of this study, self-citations contribute most to the formation of loops. Zhao et al. [24] concluded that self-citations tend to serve as 'substantial citations' (p. 949) than other external citations. Therefore, these self-citations might play an important role, such as serving as fundamental literature and supports, in the citing publications. Nevertheless, the motivation of these 'super-early' citations as well as other behaviour-level details is supposed to be discussed and researched more in-depth.

Based on Figures 3(c) and 4(c), although recent years have witnessed the increasing of the number of loops in citation networks, we should still note that the loops are rare compared with the total number of publications in each year, as argued by Leicht et al. [19]. Hence, the increase in the self-citations will not challenge the academic world.

Loops in publication citation networks might also trigger some further discussions and implications in co-citation [25–27], bibliographic coupling analyses [28,29] and their combined studies [8,30]. The SCC pattern '

Journal of Information Science, 46(6) 2020, pp. 837–848 © The Author(s), DOI: 10.1177/0165551519871826

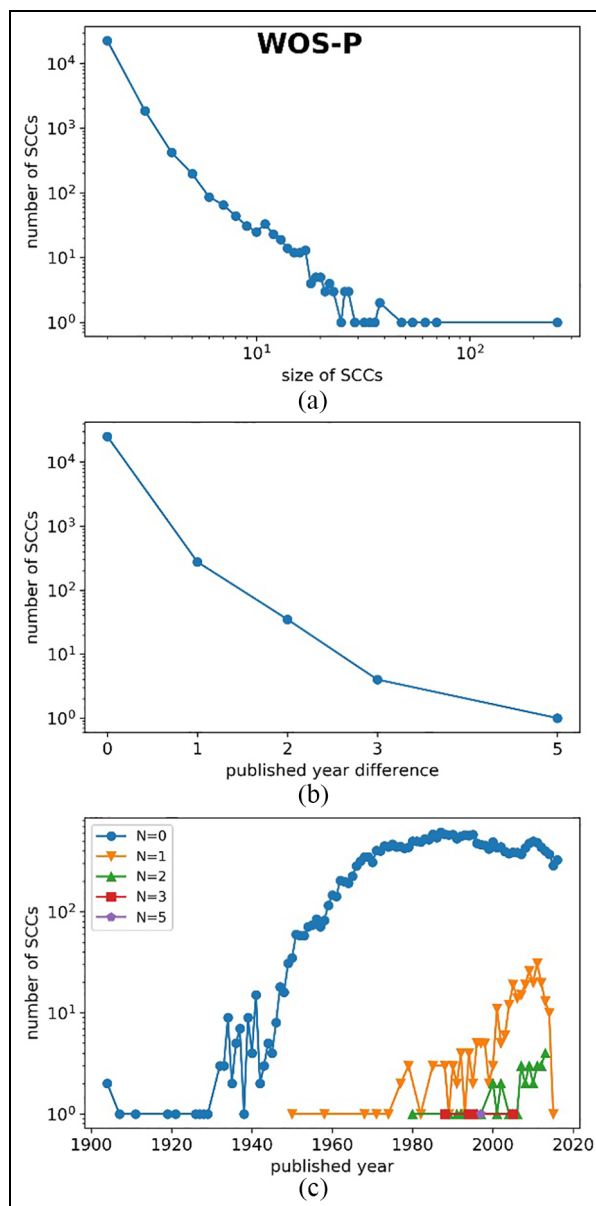



Figure 4. Loops in the citation network of physics: (a) distribution of the size of strongly connected components (SCCs), (b) time span of the largest year difference between all publications in an SCC and (c) SCC distribution of publications in different years, in which N refers to the published year difference of the oldest and youngest publications in the SCC.

aggregating publication-level indirect citations remains to be discussed, for example, the amount of information gained in the raw co-citation matrix after considering loops in citation networks.

This article has several limitations. For instance, we only investigate a natural science discipline and an engineering discipline; therefore, it is difficult to generalise our current results to all disciplines, such as art and humanities fields. Yet, given the extreme dominance of its occurrence in our empirical domains, we guess the motif of ‘’ will remain the most frequent structure among all motifs even if we duplicate our empirical study to other domains. Meanwhile, we admit that our strategy of selecting computer science and physics publications is biased, as only the ‘subject’ field in our WoS database is considered. In the future, we will employ the strategy of Sinatra et al. [33] to improve this. Moreover, we only consider citing relationships recorded in the database but not any full text-based features [34,35]. Future related work should involve citing sentences (also called ‘citances’ [36]) to investigate how they

form a loop in a given citation network from a retrospective view. Furthermore, when an article cited another one, the version it cited might also be interesting to explore in the future.

Acknowledgement

The authors acknowledge the Indiana University Pervasive Technology Institute for providing KARST, a high-performance computing system in Indiana University [37] that has contributed to the research results reported within this article. The authors thank Matthew Alexander Hutchinson for providing help on processing the empirical data and Ying Ding for fruitful discussions. The authors are grateful to two anonymous reviewers for their insightful suggestions.

Author contributions

Y.B. and Y.H. contributed equally to this article.


Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This article is financially supported by the major program of the Social Science Foundation of China (No. 17ZDA292). This research was supported – in part – by the Lilly Endowment Inc. through its support for the Indiana University Pervasive Technology Institute, and – in part – by the Indiana METACyt Initiative. The Indiana METACyt Initiative at Indiana University was also supported – in part – by the Lilly Endowment Inc.

ORCID iDs

Yi Bu  <https://orcid.org/0000-0003-2549-4580>

Wei Lu  <https://orcid.org/0000-0002-0929-7416>

Note

1. There are two elements in a network (graph), namely, nodes (also called vertices or objects) and edges (also called ties or links). For instance, in Figure 1, A–I are all nodes while links such as $A \rightarrow B$ is an edge. If all edges are connected from one node to another, this network is called a *directed network*. A citation network is typically directed because an edge in a citation network represents a specific citing relationship from a citing publication to a cited publication. However, an *acyclic network* is defined as a network in which there is no cycle (e.g. $A \rightarrow B \rightarrow C \rightarrow B \rightarrow A$ is a typical cycle in the network shown in Figure 1) in the network. An *acyclic directed network* is, therefore, defined as the intersection of direct networks and acyclic networks. That is to say, if a network is both a directed network and an acyclic network, it is an acyclic directed network.

References

- [1] Calero-Medina C and Noyons EC. Combining mapping and citation network analysis for a better understanding of the scientific development: the case of the absorptive capacity field. *J Inform* 2008; 2: 272–279.
- [2] Dawson S, Gašević D, Siemens G et al. Current state and future trends: a citation network analysis of the learning analytics field. In: *The fourth international conference on learning analytics and knowledge*, Indianapolis, IN, 24 March–28 March 2014, pp. 231–240. New York: ACM.
- [3] Huang Y, Bu Y, Ding Y et al. Number verses structure: towards citing cascades. *Scientometrics* 2018; 117: 2177–2193.
- [4] Huang Y, Bu Y, Ding Y et al. Direct citations between citing publications, 2019, <https://arxiv.org/abs/1811.01120>
- [5] Kajikawa Y and Takeda Y. Citation network analysis of organic LEDs. *Technol Forecast Soc* 2009; 76: 1115–1123.
- [6] Yan E and Ding Y. Scholarly network similarities: how bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks relate to each other. *J Am Soc Inf Sci Tec* 2012; 63: 1313–1326.
- [7] Min C, Ding Y, Li J et al. Innovation or imitation: the diffusion of citations. *J Assoc Inf Sci Tech* 2018; 69: 1271–1282.
- [8] Bu Y, Ni S and Huang W-B. Combining multiple scholarly relationships with author cocitation analysis: a preliminary exploration on improving knowledge domain mappings. *J Informetr* 2017; 11: 810–822.
- [9] Ding Y. Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. *J Informetr* 2011; 5: 187–203.

- [10] Radicchi F, Fortunato S, Markines B et al. Diffusion of scientific credits and the ranking of scientists. *Phys Rev E* 2009; 80: 056103.
- [11] Small H. Update on science mapping: creating large document spaces. *Scientometrics* 1997; 38: 275–293.
- [12] Wang Y and Bowers AJ. Mapping the field of educational administration research: a journal citation network analysis. *J Educ Admin* 2016; 54: 242–269.
- [13] Chalupsky H. Unsupervised link discovery in multi-relational data via rarity analysis. In: *The third IEEE international conference on data mining*, Melbourne, FL, 19–22 November 2003, pp. 171–178. New York: IEEE.
- [14] Hu B, Dong X, Zhang C et al. A lead-lag analysis of the topic evolution patterns for preprints and publications. *J Assoc Inf Sci Tech* 2015; 66: 2643–2656.
- [15] Lin SD and Chalupsky H. Using unsupervised link discovery methods to find interesting facts and connections in a bibliography dataset. *ACM SIGKDD Explor Newslett* 2003; 5: 173–178.
- [16] Bilke S and Peterson C. Topological properties of citation and metabolic networks. *Phys Rev E* 2001; 64(3): 036106.
- [17] De Nooy W, Mrvar A and Batagelj V. *Exploratory social network analysis with Pajek: revised and expanded edition for updated software*. Cambridge: Cambridge University Press, 2018.
- [18] Madani F, Zwick M and Daim T. Keyword-based patent citation prediction via information theory. *Int J Gen Syst* 2018; 47(8): 821–841.
- [19] Leicht EA, Clarkson G, Shedden K et al. Large-scale structure of time evolving citation networks. *Euro Phys J B* 2007; 59: 75–83.
- [20] Thulasiraman K and Swamy MN. *Graphs: theory and algorithms*. New York: John Wiley & Sons, 2011.
- [21] Nuutila E and Soisalon-Soininen E. On finding the strongly connected components in a directed graph. *Inform Process Lett* 1994; 49(1): 9–14
- [22] Tarjan R. Depth-first search and linear graph algorithms. *SIAM J Comput* 1972; 1(2): 146–160.
- [23] Milo R, Shen-Orr S, Itzkovitz S et al. Network motifs: simple building blocks of complex networks. *Science* 2002; 298(5594): 824–827.
- [24] Zhao D, Strotmann A and Cappello A. In-text function of author self-citations: implications for research evaluation practice. *J Assoc Inf Sci Tech* 2018; 69(7): 949–952.
- [25] Bu Y, Liu TY and Huang WB. MACA: a modified author co-citation analysis method combined with general descriptive meta-data of citations. *Scientometrics* 2016; 108(1): 143–166.
- [26] McCain KW. Mapping authors in intellectual space: a technical overview. *J Am Soc Inf Sci* 1990; 41(6): 433–443.
- [27] Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J Am Soc Inf Sci* 1973; 24(4): 265–269.
- [28] Zhao D and Strotmann A. Evolution of research activities and intellectual influences in information science 1996–2005: introducing author bibliographic-coupling analysis. *J Am Soc Inf Sci Tech* 2008; 59(13): 2070–2086.
- [29] Zhao D and Strotmann A. The knowledge base and research front of information science 2006–2010: an author cocitation and bibliographic coupling analysis. *J Assoc Inf Sci Tech* 2014; 65(5): 995–1006.
- [30] Bu Y, Waltman L and Huang Y. A multidimensional perspective on the citation impact of scientific publications, 2019, <https://arxiv.org/abs/1901.09663>
- [31] McCain KW. Mapping economics through the journal literature: an experiment in journal cocitation analysis. *J Am Soc Inf Sci* 1991; 42(4): 290–296.
- [32] White HD and McCain KW. Visualizing a discipline: an author co-citation analysis of information science, 1972–1995. *J Am Soc Inf Sci* 1998; 49(4): 327–355.
- [33] Sinatra R, Deville P, Szell M et al. A century of physics. *Nat Phys* 2015; 11(10): 791.
- [34] Ding Y, Zhang G, Chambers T et al. Content-based citation analysis: the next generation of citation analysis. *J Assoc Inf Sci Tech* 2014; 65(9): 1820–1833.
- [35] Lu W, Huang Y, Bu Y et al. Functional structure identification of scientific documents in computer science. *Scientometrics* 2018; 115(1): 463–486.
- [36] Nakov PI, Schwartz AS and Hearst M. Citances: citation sentences for semantic analysis of bioscience text. In: *Proceedings of the SIGIR*, Sheffield, 25–29 July 2004, vol. 4, pp. 81–88. Sheffield: University of Sheffield.
- [37] Stewart CA, Welch V, Plale B et al. Indiana University Pervasive Technology Institute, 2017, <https://doi.org/10.5967/K8G44NGB>