Characterizing the Citation Graph as a Self-Organizing Networked Information Space

Yuan An[1], Jeannette C.M. Janssen[2], and Evangelos E. Milios[1]

[1] Faculty of Computer Science, Dalhousie University
Halifax, NS, Canada
{yuana,eem}@cs.dal.ca

[2] Dept. of Math. and Stats., Dalhousie University
Halifax, NS, B3H 3J5 Canada
janssen@mscs.dal.ca

Abstract. Bodies of information available through the Internet, such as digital libraries and distributed file-sharing systems, often form a self-organizing networked information space, i.e. a collection of interconnected information entities generated incrementally over time by a large number of agents. The collection of electronically available research papers in Computer Science, linked by their citations, form a good example of such a space. In this work we present a study of the structure of the citation graph of computer science literature. Using a web robot we build several citation graphs from parts of the digital library *ResearchIndex*. After verifying that the degree distributions follow a power law, we apply a series of graph theoretical algorithms to elicit an aggregate picture of the citation graph in terms of its connectivity. The results expand our insight into the structure of self-organizing networked information spaces, and may inform the design of focused crawlers searching such a space for topic-specific information.

1 Introduction

Thanks to the expansion and growing popularity of the Internet, a rapidly increasing amount of information is available electronically and in networked form. Thanks to the open and distributed nature of the World Wide Web, bodies of information are often created in a self-organizing way: the information entities are created by independent agents, and each agent links its information to the entities of a limited number of other agents. Hence we can refer to such bodies of information as *self-organizing networked information spaces*. The body of web pages and their hyperlinks forms a canonical example of such an information space, but the same principle applies, for example, to digital libraries and distributed file-sharing systems.

An interesting property of networked information spaces is that information is not only encoded in the entities themselves, but also in the link structure. Many properties of the entities can be inferred from the link structure. Graph-theoretic methods to study this link structure have therefore become popular. Since self-organizing networked information spaces are the product of a similar

H. Unger, T. Böhme, and A. Mikler (Eds.): I^{2}CS 2002, LNCS 2346, pp. 97–107, 2002.

generative process, one would expect that their link structure will have certain common characteristics. Previous studies have focused mainly on the World Wide Web; we present here a study of an electronic library of scholarly papers in Computer Science.

A body of scientific literature can be seen as a networked information space. Here the information entities are the scientific papers, and they are linked together by citation relations. The link structure of this networked information space can be represented by a directed graph, which is commonly referred to as the *citation graph*. Each node of the citation graph represents a paper, and a directed link from one node to another implies that the paper associated with the first node cites the paper associated with the second node.

Citation graphs representing scientific papers contain valuable information about levels of scholarly activity and provide measures of academic productivity. A citation graph has the potential of revealing interesting information about a particular scholarly research topic: it may be possible to infer research areas and their evolution over time, measure relations between research areas and trace the influence of ideas that appear in the literature.

In this paper we report the results of examining various aspects of connectivity of the citation graph of computer science literature with graph theoretic algorithms. To build the citation graph, we implemented a web robot to query the online computer science library *ResearchIndex*.

Research in bibliometrics has long been concerned with the use of citations to produce quantitative estimates of the importance and impact of individual scientific publication and journals. The best-known measure in this field is Garfield's impact factor [7]. The impact factor is a ranking scheme based fundamentally on a pure counting of the in-degree of nodes in the citation graph. Redner [9] has focused on the statistical distribution of the number of citations of the scientific literature. Chen [4,5] developed a set of methods that extends and transforms traditional author co-citation analysis by heuristically extracting structural patterns from scientific literature for visualization as a 3D virtual map.

As to structural analysis of other networked information spaces, Broder et al. [2] studied various properties of Web graph including its diameter, degree distributions, connected components, and macroscopic structure, proposing a bow tie model of the Web. Earlier work, exploring the scaling properties of the Web graph, has been done by Barabasi [1]. More recent work, comparing properties of the Web at various levels, can be find in [6]. Exploiting the link topology of networked information space for information discovery has been recently proposed for the Web [3].

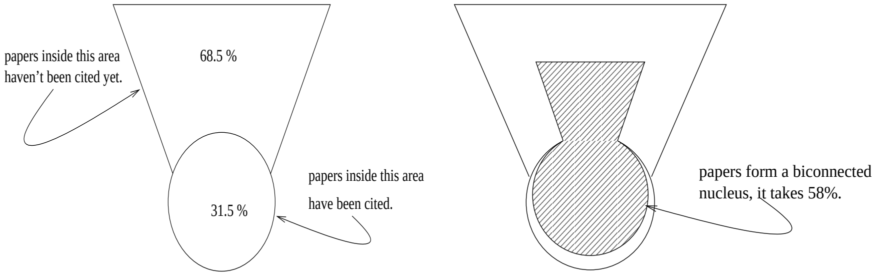
The following considerations motivated our study. Understanding the link topology of the citation graph using graph-theoretic tools may facilitate knowledge discovery relying on link information such as similarity calculation and finding communities, help in citation graph visualization, and help evaluate the evolution of specialties or research themes over time. Moreover, comparing the structure of a citation graph with that of other networked information spaces such as the Web will increase our understanding of the factors that influence the link structure of such spaces. This sort of understanding will lead to improved methods for Web navigation and data mining.

1.1 Main Results

We performed three sets of experiments on our collection of citation graphs obtained from different research areas. The main result of our analysis was that these citation graphs showed remarkably similar behavior in each of our experiments. Moreover, the results of the experiments performed on the union of the three graphs was again similar to that of each of its parts, indicating the self-similarity of the citation graph.

We first constructed a robot for querying *ResearchIndex* [8], and using the robot we built a collection of three local citation graphs by starting with papers from three different topics. We also merged the three graphs into the union graph: the combined citation graph of the three individual ones.

The first set of experiments computed the in-degree distributions and demonstrated that they follow a power law. Specifically, the fraction of articles with k citations is proportional to $1/k^e$, where the exponent e is close to 1.7 for each of the four graphs. We also investigated the average shortest path length between nodes, concluding that, if direction of the links is ignored, the citation graph classifies as a *small-world network*.



(a) 68.5% of the nodes have no incoming link

(b) 58% of the nodes in the giant Weakly Connected Component(WCC) account for a big Biconnected Component(BCC)

Fig. 1. The connectivity of the citation graph

The second set of experiments investigated the connectivity of the citation graph. It was found that approximately 90% of the nodes form a single Weakly Connected Component (WCC) if citations are treated as undirected edges. Within this giant WCC, almost 68.5% of the nodes have no incoming link, suggesting that 68.5% of the publications in the giant WCC have not been cited (yet). See Figure 1(a) for a representation of this result. Furthermore, within the giant WCC, around 58% of its publications form a large Biconnected Component(BCC), and almost all the remaining nodes of the giant WCC fall into trivial BCCs, each of which consists of a single distinct node. The aggregate picture that emerges is shown in Figure 1(b).

2 Measurements on the Citation Graph

2.1 Building the Citation Graph

The first step is to extract citation graphs from a citation database. *ResearchIndex*[8] is a Web-based digital library and citation database of computer science literature and provides us easy access to citation information for our study. We constructed a Web robot for querying *ResearchIndex* autonomously. We chose three areas within computer science as starting points: *Neural Networks, Automata and Software Engineering.*

Our procedure for creating the citation graphs started from a base set, obtained via keyword search, containing thousands of nodes that are not necessarily connected. We then expanded this base set by following incoming links and outgoing links of the nodes in the base set. The crawling process was terminated when space and time limitations were reached. About 100,000 papers were parsed for each topic.

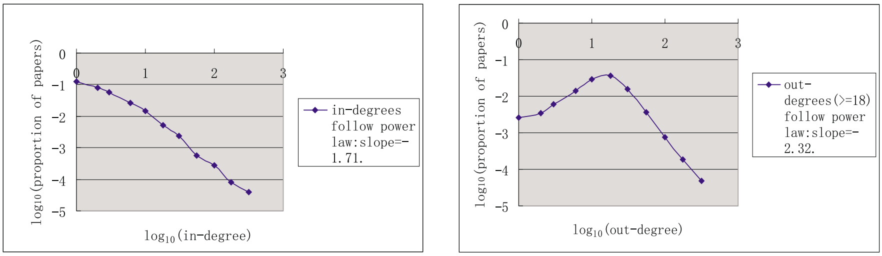
The above process leads to the formation of three raw citation graphs for each of the selected topics and their union graph. We note that there are two types of articles in the raw citation graphs: the first type of article is fully available in *ResearchIndex*, including its full text and references; the second type of article is brought into *ResearchIndex* by a reference of other papers, but its text and references are not in *ResearchIndex.* The second type only contributes part of the information to the citation graph. In the experiments reported in this article, the citation graphs used were obtained from the raw citation graphs by removing all articles of the second type. The measurements we extracted from the citation graphs we built included in- and out-degree distributions (involving only the articles in the citation graphs, which are a subset of the citing and cited articles respectively) and diameters.

2.2 Degree Distributions

We begin by considering the in-degrees of nodes in the citation graph. We observed that the in-degree distributions follow a power law; i.e. the fraction of papers with in-degree i is proportional to $1/i^{\gamma}$ for some $\gamma > 1$. Our experiments on all citation graphs built from the different topics as well as the union citation graph confirmed this result at a variety of scales. In all these experiments, the value of the exponent γ in the power law for in-degrees is a remarkably consistent 1.7.

Figure 2(a) is a log-log plot of the binned in-degree distribution of the union citation graph for extracting the exponent γ. The value $\gamma = 1.71$ is derived from the slope of the line providing the best linear fit to the data in the figure.

The out-degree distribution in the union citation graph follows a more complex distribution, shown in 2(b). It peaks at 16, and after 18 it follows a power



(a) The in-degree distribution subscribes to a power law with exponent=1.71

(b) The out-degree distribution

Fig. 2. In- and outdegree distribution in the union citation graph

law distribution with exponent 2.32. This outcome is not surprising, as there are very few papers, typically tutorial in nature, with a large number of references, while the majority of the papers have references in the range of 20 to 50. It should be noted that the out-degree of a paper in our citation graph is less than its number of references, since we only include in the citation graph the papers that are fully available in the *ResearchIndex* database. This affects older papers more, since their references are less likely to be available in electronic form.

2.3 Diameter

We turn next to the diameter measurement of citation graphs. In this study, the diameter is defined as the maximum over all ordered pairs(u, v) of the length of the shortest path from u to v in the citation graph. We measured two types of diameter for the citation graph: directed diameter and undirected diameter. Directed diameter is measured by the directed shortest path or dipath, while undirected diameter is obtained by treating edges as undirected.

Our connectivity tests revealed that the citation graph is not connected. This means that there are nodes which cannot be reached by a path from other nodes, implying that the diameter is infinite. However, the tests also revealed that $\approx 80\% - 90\%$ of the nodes are in one giant connected component, while the rest form a few very small components. Details are described in Section 3. We therefore considered the diameter of this giant connected component as the undirected diameter of the graph.

The diameters obtained by applying Dijkstra's shortest path algorithm on the giant connected components of the citation graphs built for the three topics and their union are shown in Table 1.

Ignoring the orientation of edges, we observe that the citation graph is a 'small world' with an undirected diameter of around 18. The result is consistent at a variety of scales and topics.

Table 1. The diameters of citation graphs built from different topics as well as union citation graph. Topic: N.N.: Neural Networks, S.E.: Software Engineering

	graph size	directed diameter	undirected diameter
citation graph–N.N.	23,371	24	18
citation graph–Automata	28,168	33	19
citation graph–S.E.	19,018	22	16
union citation graph	57,239	37	19
average		29	18

In contrast, we do not obtain such a 'small world' property in the *directed* citation graph. Our statistical study shows that the probability of having a directed path between any pair of nodes is only *2%*. The directed diameter was calculated by taking the maximum only over those pairs of nodes that are connected by a directed path. This diameter turned out to be around 30 (see Table 1). This is an outcome of the temporal nature of the citation graph. In almost all cases, references can only be made to papers that appeared previously, and therefore directed cycles are unlikely. (Some directed cycles arise in special circumstances, see Section 3.2)

3 Reachability and Connected Components

We now consider the connectivity of our citation graphs of computer science literature. This involves examining the various types of its connected components and reachability of nodes. Given a citation graph $G = (V, E)$, we will view G both as a directed graph as well as and undirected graph (the latter by ignoring the direction of all edges). We now ask how well connected the citation graph is. We apply a set of algorithms that compute reachability information and structural information of directed and undirected citation graphs: *Weakly Connected Components (WCC), Strongly Connected Components (SCC)* and *Biconnected Components (BCC).*

3.1 Weakly Connected Components

The first of our connectivity experiments showed that the citation graph is not, in general, connected. This can be explained in the context of our construction of the citation graphs: we started building each citation graph from a base set containing a number of documents which are not necessarily connected, and while the expansion of the base set serves to connect many of these documents, others remain in small isolated components. Moreover, our cleaning up process of removing those articles, whose text and references are not available, produced more isolated components.

Mathematically, a *Weakly Connected Component(WCC)* of an undirected graph $G = (V, E)$ is a maximal connected subgraph of G. A WCC of a citation graph is a maximal set of articles each of which is reachable from any other if links may be followed either forwards or backwards. In the context of a citation graph, links stand for the citations from one article to other articles cited in the former one. The WCC structure of a citation graph gives us an aggregate picture of groups of articles that are loosely related to each other.

The results drawn from the weakly connected component experiments on citation graphs are shown in Table 2. The results reveal that the citation graph is well connected–a significant constant fraction $\approx 80\% - 90\%$ of all nodes fall into one giant connected component. It is remarkable that the same general results on connectivity are observed in each of the three topic subgraphs. In turn, the same behavior is observed for the union graph, suggesting a certain degree of self-similarity. (The term *self-similarity* is used here, as in [6], to denote similar statistical behavior at several levels.)

Table 2. The results of Weakly Connected Component experiments on different citation graphs: the majority ($\approx$ 90%) of articles are connected to each other if links are treated as without directions.citation graph:N.N stands for Neural Networks; S.E. stands for Software Engineering

	graph size	size of largest WCC	percentage of largest WCC	size of second largest WCC
citation graph–N.N.	23,371	18,603	79.6%	21
citation graph–Automata	28,168	25,922	92%	20
citation graph–S.E.	19,018	16,723	87.9%	12
union citation graph	57,239	50,228	87.8%	21

3.2 Strongly Connected Components

We turn next to the extraction of *Strongly Connected Component(SCC)* of the connected components of the three topical citation graphs and their union graph. A *Strongly Connected Component(SCC)* of a directed graph is a maximal subgraph such that for all pairs of vertices (u, v) of the subgraph, there exists a directed path (dipath) from u to v. An article cannot cite articles that have not been written yet, so if article u directly or indirectly cites article v, then v must be older than u, so, under normal circumstances, v will not cite u. As a result, we might expect that there is no SCC in the citation graph. But contrary to our expectation, the results of SCC experiments on the collection of citation graphs reveal that there exist one to three sizable SCC's in each of the citation graphs, as well as a few very small SCC's. The results drawn from the experiments are shown in Table 3.

In order to know how the directed cycles were generated in those citation graphs, we extracted some SCCs from citation graphs and searched the corresponding articles of these SCCs directly in *ResearchIndex*'s database. Our study shows that several types of publications formed SCCs: (1) publications written by same authors tend to cite each other, they usually produce self-citations, (2) publications which are tightly relevant tend to cite each other, e.g., publications,

Table 3. The results of Strongly Connected Component experiments on different citation graphs: there exist many small SCCs, among them there are one -three bigger SCC(s), the rest are even smaller comparing those bigger ones. citation graph:N.N stands for Neural Networks; S.E. stands for Software Engineering

	graph size	size of largest SCC	size of second largest SCC	size of third largest SCC
citation graph–N.N.	18,603	144	14	10
citation graph–Automata	25,922	192	29	24
citation graph–S.E.	16,723	17	11	8
union citation graph	50,228	239	155	60

whose authors in same institute, dealing with same specialty and getting published concurrently are highly relevant and tend to cite each other, (3) publications which were published in several different forms, such as journals, conference proceedings or technical reports, at different times often formed directed cycles with other publications. The different forms of the publication were considered as one node during our creation process of the citation graph. (4) books or other publications which were published in several editions at different times, where the newer editions contained more recent references, often acted as jump points in the citation graph. The jump points formed by publications of type (4) caused large directed cycles in the citation graph; this is the reason of the existence of one to three bigger SCCs. Types (1)–(3) of articles usually gave rise only to small SCCs containing 2–5 articles.

A conceptual map arising from the analysis of the results of the SCC experiment on the union citation graph is depicted in Figure 3. A number of small SCCs are embedded in a well connected background net. This background net is a directed acyclic structure, i.e., there is no directed cycle in the background net.

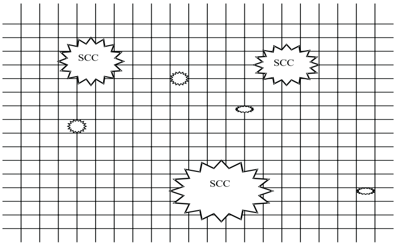


Fig. 3. The directed connectivity of a citation graph: a number of small SCCs embedded in a background net;the background net is a directed acyclic graph

3.3 Biconnected Components

We now turn to a stronger notion of connectivity in the undirected view of the citation graph, that of biconnectivity. A *Biconnected Component(BCC)* of an undirected graph is a maximal subgraph such that every pair of vertices is biconnected. Two vertices u and v are biconnected if there are at least two disjoint paths between u and v, or, equivalently, if u and v lie on a common cycle. Any biconnected component must therefore lie within a weakly connected component. Applying the biconnected component algorithm on the giant connected components of citation graphs, we find that each giant connected component of each citation graph contains a giant biconnected component. The giant BCC acts as a central biconnected nucleus, with small BCCs connected to this nucleus by cut vertices, and other single trivial nodes connected to the nucleus or a small BCC.

The numerical analysis of sizes of BCCs indicated that $\approx 58\%$ of all nodes account for the giant biconnected nucleus, the rest $\approx 40\%$ of the nodes are in trivial BCCs each of which consists of single distinct node, and the remaining $\approx 2\%$ of the nodes fall into a few small BCCs.

4 Does Connectivity Depend on Some Key Articles?

We have observed that the citation graph is well connected–90% of the nodes form a giant connected component which in turn contains a biconnected nucleus with 58% of all nodes. The result that the in- distributions follow a power law indicates that there are a few nodes of large in-degree. Moreover, our analysis of the out-degrees implies that there are also some nodes with large out-degree. We are interested in determining whether the widespread connectivity of the citation graph results from a few nodes of large in-degree acting as “authorities” or a few nodes of large out-degree acting as “hubs”. We test this connectivity by removing those nodes with large in-degree or out-degree, and computing again the size of the largest WCC. The results are shown in Table 4 and Table 5.

Table 4. Sizes of the largest Weakly Connected Components(WCCs) when nodes with in-degree at least k are removed from the giant connected component of union citation graph

size of graph	50,228							
k	200	150	100	50	10	5	4	3
size of graph after removing	50,222	50,215	50,152	49,775	46,850	43,962	42,969	41,246
size of largest WCC	50,107	49,990	48,973	43,073	26,098	14,677	9,963	1,140

These results show that the widespread connectivity does not depend on either hubs or authority papers. Indeed, even if all links to nodes with in-degree 5 or higher are removed (certainly including links to every well-known article

Table 5. Sizes of the largest Weakly Connected Components(WCCs) when nodes with out-degree at least k are removed from the giant connected component of union citation graph

size of graph	50,228							
k	200	150	100	50	10	5	4	3
size of graph after removing	50,225	50,225	50,224	50,205	48,061	43,964	42,238	39,622
size of largest WCC	50,202	50,202	50,198	50,131	46,092	37,556	33,279	26,489

on computer science), the graph still contains a giant Weakly Connected Component(WCC). Similarly, if all links to nodes with out-degree 3 or higher are removed, the graph is still well connected. We conclude that the connectivity of citation graph is extremely resilient and is not due to the existence of hubs and authorities, which are embedded in a graph that is well connected without their contributions.

5 Discussion

We have reported the results of our examination of the self-organizing networked information space formed by electronically available scientific articles in Computer Science. The link structure of this space (referred to as the citation graph) can potentially be used in a variety of ways, for example to infer research areas and their evolution over time, measure relations between research areas, and trace the influence of ideas that appear in the literature.

For our analysis of the citation graph we applied graph-theoretic algorithms. We verified that the in-degree distribution follows a power law, a characteristic observed in various experimental studies to hold for other networked information spaces. We also studied the connectivity by extracting weakly and strongly connected components, as well as biconnected components. The aggregate picture emerging here differs from that of the Web, since citations, unlike hyperlinks, generally are restricted by the time in which a paper was written (older papers cannot reference newer papers). We measured the diameter of the graph, and verified that it is lower than would be expected of a random graph of comparable sparsity, classifying a citation graph as a "small world network". We also found evidence that the citation graph is quite robust in terms of connectivity; when nodes with low degree were removed, the graph still stayed mostly connected. In general, we found that the citation graph displays many of the characteristics of other networked information spaces, though it differs in some aspects due to the specific, time-dependent nature of citations. A suggestion for further research is the use of the observed characteristics of the citation graph to develop tools for better navigation, mining and retrieval in networked information spaces, such as the World Wide Web or corporate intranets.

In a follow-up of this study, we computed minimum cuts between authority papers in different areas, with the hope that this would enable us to separate the different research communities represented by the graph. These naive attempts were largely unsuccessful. The extraction of communities from the citation graph is an important area of further study. Our intuition and experience tells us that papers on a specific research topic must be more densely interconnected than random groups of papers. Hence research topics or "communities" should correspond to locally dense structures in the citation graph. However, our work shows that the connectivity of citation graphs as a whole is such that it is not possible to extract such communities with straightforward methods such as minimum cut. More sophisticated methods are needed if we wish to succeed in mining the community information encoded in the link structure of a citation graph or other networked information spaces.

Another important subject of further study is the evolution of the citation graph over time. Knowledge about the temporal evolution of the local link structure of citation graphs can be used to predict research trends or to study the life span of specialties and communities. Such knowledge can also be used for the development of dynamic models for the citation graph. Such models can, in turn, give insight into the self-organizing processes that created the citation graph, and serve as a tool for prediction and experimentation.

References

1. A-L.Barabasi and R.Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.
2. A.Z. Broder, S.R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web: experiments and models. In *Proc. 9th WWW Conf.*, pages 309–320, 2000.
3. Soumen Chakrabarti, Byron E. Dom, David Gibson, Jon Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Mining the Link Structure of the World Wide Web. *IEEE Computer*, 32:60–67, 1999.
4. Chaomei Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35:401–420, 1999.
5. Chaomei Chen. Visualising a knowledge domain's intellectual structure. *IEEE Computer*, 34:65–71, 2001.
6. S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. In *Proceedings of the 27th VLDB conference*, Roma, Italy, 2001.
7. E.Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972.
8. Steve Lawrence, Kurt Bollacker, and C. Lee Giles. *ResearchIndex.* NEC Research Institute, `http://citeseer.nj.nec.com` (accessed on Sep.30, 2001), 2001.
9. S.Redner. How Popular is Your Paper? An Empirical Study of the Citation Distribution. *European Physical Journal B*, 4:131–134, 1998.