DOI 10.1007/s10115-003-0128-3
Springer-Verlag London Ltd. © 2004
Knowledge and Information Systems (2004) 6: 664–678

Knowledge and
Information Systems

# Characterizing and Mining the Citation Graph of the Computer Science Literature

Yuan An[1], Jeannette Janssen[2], Evangelos E. Milios[3]

[1]Department of Computer Science, University of Toronto
[2]Department of Mathematics and Statistics, Dalhousie University
[3]Faculty of Computer Science, Dalhousie University

**Abstract.** Citation graphs representing a body of scientific literature convey measures of scholarly activity and productivity. In this work we present a study of the structure of the citation graph of the computer science literature. Using a web robot we built several topic-specific citation graphs and their union graph from the digital library ResearchIndex. After verifying that the degree distributions follow a power law, we applied a series of graph theoretical algorithms to elicit an aggregate picture of the citation graph in terms of its connectivity. We discovered the existence of a single large weakly-connected and a single large biconnected component, and confirmed the expected lack of a large strongly-connected component. The large components remained even after removing the strongest authority nodes or the strongest hub nodes, indicating that such tight connectivity is widespread and does not depend on a small subset of important nodes. Finally, minimum cuts between authority papers of different areas did not result in a balanced partitioning of the graph into areas, pointing to the need for more sophisticated algorithms for clustering the graph.

**Keywords:** Citation graph; Graph connectivity; Networked information spaces; Power law; Small worlds

## 1. Introduction

A body of scientific literature can be seen as a networked information space, a collection of information entities connected by a link structure. 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

*Received 7 November 2001*
*Revised 2 January 2003*
*Accepted 26 February 2003*
*Published online 6 February 2004*

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 to reveal 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.

As a first step in this direction, we investigate in this article a citation graph in a manner similar to the investigations carried out on the World Wide Web. Our study was carried out in three stages. First, we confirm that the in-degree distribution of our citation graphs follows a power law distribution, as discovered by other studies of the citation structure of the science literature, and by studies of the World Wide Web. Second, we investigate the connectivity properties of the citation graph by applying a series of graph theoretic algorithms to compute the weakly connected components, strongly connected components and biconnected components. An understanding of the connectivity properties of the citation graph is a prerequisite for addressing the problem of clustering the citation graph into subject areas. It also parallels similar studies on the World Wide Web. Third, we attempt to separate the citation graph into subject areas. To this end, we compute the global minimum cut, and the minimum cuts separating authority papers in different areas. We also compute the shortest paths between pairs of papers. An aggregate picture of the citation graph in terms of its connectivity emerges out of the results. This picture points out that the citation graph is tightly connected, and finding communities is a non-trivial problem.

To build citation graphs for our experimental investigations, we implemented a web robot to query the online computer science library *ResearchIndex*. We built the citation graph around three different areas within computer science: Neural Networks, Software Engineering and Automata. The areas were chosen to be distinct from each other, and to contain one area (Neural Networks) familiar to the authors. Although there is no theoretical guarantee that the topics chosen are representative of the big picture, we found that the union graph of the three areas and the individual area graphs show remarkably similar behaviour under all the connectivity tests that we conducted.

The following considerations motivated our study. Understanding the link topology of the citation graph using graph-theoretic tools may:

1. facilitate knowledge discovery relying on link information such as similarity calculation, and finding communities.
2. help in citation graph visualization.
3. help evaluate the evolution of specialities or research themes over time.

## 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 behaviour in each of our experiments. Moreover, the results of the experiments performed on the union of the three graphs were 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* (Lawrence, Bollacker, and Giles 2001), and using the robot we built a collection of three local citation

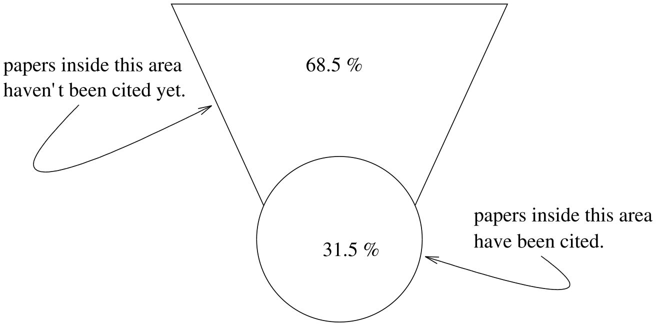


**Fig. 1.** The connectivity of the citation graph: 68.5% of the nodes have no incoming link.

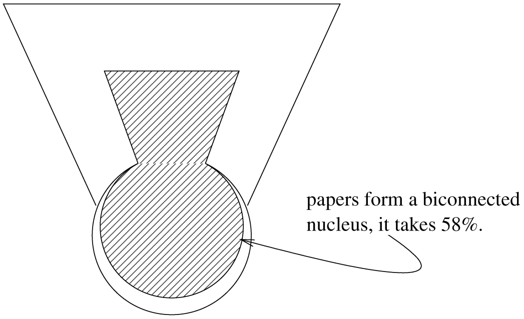


**Fig. 2.** The connectivity of the citation graph: 58% of the nodes in the giant Weakly Connected Component(WCC) account for a big Biconnected Component(BCC).

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.

Papers whose full text (including references) were not available on ResearchIndex were not included in the citation graph. This is an assumption we made to simplify graph analysis. Otherwise we would have to consider two distinct categories of nodes, thus severely complicating the analysis. This limitation is not unique to ResearchIndex and our graphs, it is inherent in any citation index. We informally examined the Science Citation Index(SCI), and we found a similar proportion of papers published in venues not indexed there.

The first set of experiments computed the in-degree distributions and confirmed previous studies (Redner 1998) demonstrating 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.

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 Fig. 1 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 Fig. 2.

The third set of experiments applied the Minimum cut Maximum flow algorithm for finding cluster information. The results indicated that we need more sophisticated tools for the clustering problem in the citation graph.

### 1.2. Related Work

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 (Garfield 1972). The impact factor is a ranking scheme based fundamentally on a pure counting of the in-degree of nodes in the citation graph. Redner (Redner 1998) has focused on the statistical distribution of the number of citations of the scientific literature. Chen (Chen 1999; Chen 2001) 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 large scale networks, Broder et al. (Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, and Wiener 2000) 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 (Barabasi and Albert 1999). Exploiting the link topology of large-scale networks for information discovery has been recently proposed for the Web (Chakrabarti, Dom, Gibson, Kleinberg, Kumar, Raghavan, Rajagopalan, and Tomkins 1999). This work starts with a query and the response set of an index-based search engine to it, which is viewed as the "root" set. The base set consists of all the pages that link to or are linked from the root set (excluding links within the same domain). Finally, authorities and hubs are computed in the root set, and the pages with highest authority and hub values are returned. In a related project, the authors attempt to identify communities by enumerating complete directed bipartite subgraphs of the Web, and then expanding them through treating them as the root set in the previous algorithm.

## 2. Measurements on the Citation Graph

The first step is to extract citation graphs from a citation database. *ResearchIndex* (Lawrence, Bollacker, and Giles 2001) 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. A measure of topic drifting could be a more plausible stopping criterion. However, this was not an option in our study, as we did not have access to the text of the articles. Therefore we should treat the titles of the three subgraphs we constructed (Neural Networks, Software Engineering and Automata) with an awareness that they represent three graphs built with these topics as starting points, and there is no guarantee

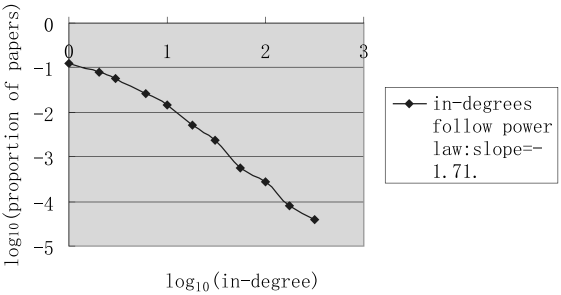


**Fig. 3.** The in-degree distribution in the union citation graph in computer science literature subscribes to the power law with exponent $= 1.71$.

that their contents fully cover or are restricted only to these areas. It is reasonable to expect, however, that the process led to a union graph that includes several areas. Defining the areas is in itself a problem that depends on the judgment of experts and is difficult to define formally. Examining the topical cohesion of scientific document collections is beyond the scope of this article, and it is a direction we are currently pursuing. A study of the full citation graph would be more interesting, as setting a fixed size for stopping the search may not fully capture a topic, if it is larger than the size explored, or it may drift into other topics.

The above process leads to the formation of three raw citation graphs 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.

## 2.1. 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 $\gamma$ in the power law for in-degrees is a remarkably consistent 1.7.

Figure 3 is a log-log plot of the binned in-degree distribution of the union citation graph for extracting the exponent $\gamma$. 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 Fig. 4. It peaks at 16, and after 18 it follows a power 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

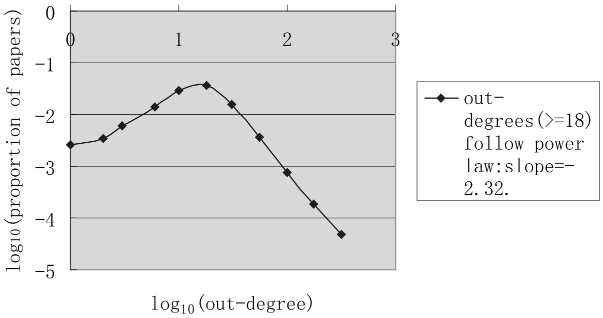


**Fig. 4.** The out-degree distribution in the union citation graph.

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.

## 2.2. Diameter

We turn next to the diameter measurement of citation graphs. In this study, the diameter is defined as the maximum shortest path between any pair of nodes in the graph. More formally, the diameter is the maximum over all ordered pairs $(u, v)$ of the length of the shortest path from $u$ to $v$ in the citation graph, where $u$ and $v$ are nodes of the 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 the shortest path when treating edges as undirected.

Before we measured diameters, we tested the connectivity of the citation graph as an undirected graph. The results 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 Sect. 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.

We observe that the citation graph, if we ignore the direction of edges, is a 'small world' (in a sense similar to (Watts and Strogatz 1998)) with an undirected diameter of around 18. The result is consistent at a variety of scales and topics.

A completely different picture emerges when we consider 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). We attribute the lack of connectivity to the temporal nature of the citation graph. In almost all cases, references

**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 |

can only be made to papers that appeared previously, and therefore directed cycles are unlikely. (Some directed cycles arise in special circumstances, see Sect. 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 an 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

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 behaviour is observed for the union graph, suggesting a certain degree of self-similarity.

### 3.2. Strongly Connected Components

We turn next to the extraction of the *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

**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 |

**Table 3.** The results of Strongly Connected Component experiments on different citation graphs: there exist many small SCCs, among them there are 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 |

such that for all pairs of vertices $(u, v)$ of the subgraph, there exists a directed path (dipath) from $u$ to $v$. In the context of the citation graph, a dipath from $u$ to $v$ means that article $u$ directly cites article $v$ or article $u$ cites an intermediate article $w$, $w$ cites the next intermediate article and so on, until it reaches article $v$ indirectly. Since there is a temporal direction between citing article and cited article, if article $u$ directly or indirectly cites article $v$, then $v$ would not cite back to $u$. 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 three sizable SCCs in each of the citation graphs, as well as a few very small SCCs. The results drawn from the experiments are shown in Table 3.

In order to know which publications formed the SCCs, i.e., how the directed cycles were generated in those citation graphs, we extracted some SCCs from citation graphs and searched articles of these SCCs directly in *ResearchIndex*'s database to find their titles, abstracts, authors, journals and published years. Our study shows that several types of publications formed SCCs: (1) publications written by the 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, whose authors are in the same institute, dealing with the same speciality 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 graphs. (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

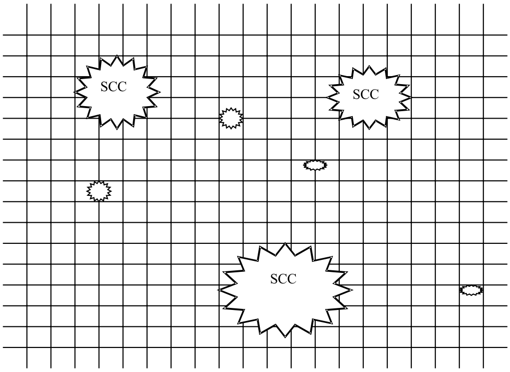


**Fig. 5.** The directed connectivity of a citation graph: a number of small SCCs embedded in a background net; the background net is a directed acyclic graph.

graph; this is the reason for the existence of 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 Fig. 5. 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.

## 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.

## 3.4. Aggregate Picture

From the application of algorithms to detect connected components, an aggregate picture of the citation graph as an undirected graph emerged. First of all, 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,

**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 |

**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 |

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. Secondly, the results of the WCC experiment indicate that $\approx$ 90% of the nodes form a giant weakly connected component. The single giant WCC can be divided into two parts: one part contains almost 68.5% of the nodes without any incoming link, suggesting that 68.5% of publications have not been cited yet, another part contains the rest of publications with at least one citation. Finally, in the giant WCC, around 58% of nodes form a big Biconnected Component(BCC) and act as a biconnected nucleus, with a few small BCCs connected to this nucleus by cut vertices, and all the rest of nodes fall into trivial BCCs each of which consists of a single distinct node connected to this nucleus or some other small pieces. The aggregate picture is shown in Figs. 1 and 2.

## 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 Tables 4 and 5.

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 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 the 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. A topic for further research is to conduct a more thorough investigation of the connectivity properties (for example average in-degree) of the graphs resulting from the removal of hub/authority or non-hub/non-authority nodes.

## 5. Minimum Cuts

A question related to understanding the structure of the citation graph is that of finding thematically cohesive communities. So far, our study of various types of connected components has resulted in a well-connected citation graph with a giant biconnected nucleus. The next question is whether there is any further structure within this nucleus. We attack this problem using minimum cut algorithms, both for global minimum cut and for minimum cuts between specific pairs of nodes.

Mathematically, an (edge) cut $C$ of graph $H = (V, E)$ is a set of edges which, when removed, disconnect the graph. The size of a cut is the number of edges in the cut. Given an edge weight function $w : E \to R$, a minimum cut is a cut whose total weight is minimum.

Our min-cut experiments focus on the giant connected component of union citation graph. After extracting the giant connected component $H = (V, E)$ from the union citation graph, we assign edge weight $w(e) = 1$ for all $e \in E$, and then apply the global minimum cut algorithm on $H$. In order to explore the interior structure of graph $H$, we implemented the min-cut exploratory procedure as follows:

```
procedure Explore_Min_Cut ($H = (V, E)$)
   while $|H| > 0$
     compute min-cut $C$ of $H$;
     calculate edge weight over crossing edge set $F$;
     let $H_1 = (C, E_1)$ be graph induced by $C$;
     let $H_2 = (V - C, E_2)$ be graph induced by $V - C$;
     Explore_Min_Cut($H_1$);
     Explore_Min_Cut($H_2$);
   end while;
```

Most of the resulting cuts found by the above procedure are trivial cuts, each of which separates only one node from the rest of the graph.

When the procedure *Explore_Min_Cut(H)* is applied to graph $H$ recursively until the fragments become trivially small, we find that 99% of the cuts are trivial.

From these experiments we conclude that the interior link structure of the citation graph is dense; There is *no* explicit community information discernible through global minimum cuts. We conclude that we need more sophisticated tools for finding communities in citation graph.

Since the global minimum cut approach does not give us communities, we turn to the computation of minimum cuts between specific nodes. If two nodes are selected to belong to different topics, then possibly the minimum cut between them might separate the papers belonging to the two topics. To investigate this hypothesis we selected authority papers (papers with large in-degree) belonging to the topics of Neural Networks, Automata and Software Engineering, and computed the minimum cuts of the union graph between pairs of authority papers. For the minimum cut computation, we made two modifications to the (directed) union graph of 50,228 nodes:

1. We added the reverse edges to all the edges of the graph, so as to effectively treat it as an undirected graph.
2. We added node capacities equal to 1, by the following construction. We replace each node $v$ by two nodes $v_{in}$ and $v_{out}$ and an edge from $v_{in}$ to $v_{out}$, and we connect all the edges into $v$ to $v_{in}$ and all the edges out of $v$ to $v_{out}$. All edges have capacity 1, thus allowing us to associate capacities 1 to all the nodes, as well.

The resulting graph has 100,456 nodes.

From these experiments we obtain highly unbalanced partitions of the union graph. The cut sizes are similar to the in-degree of the nodes, and the smaller partition contains at most a few hundred nodes while the larger partition contains the rest of the nodes (approximately 99,000).

These results further confirm the high overall connectivity of the citation graph. Generally, the number of different paths between the source and sink papers is much higher than the degree of these papers. This effect is much more prominent than the difference in connectivity in the interior and the exterior of a community of nodes corresponding to one subject area, which was the property we hoped to exploit. In future work, the minimum cut method should be biased towards more balanced cuts, to avoid the highly unbalanced partitions that we obtain with the unmitigated method (Even, Naor, Rao, and Schieber 1999).

## 6. Conclusion

The Citation Graph of Computer Science is a directed graph whose nodes are articles and whose edges are references that appear in the node of origin. 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. In this article, we reported the results of examining the citation graph with graph theoretic algorithms. We constructed a web robot querying the computer science digital library *ResearchIndex* and built the citation graph for three different areas within computer science. We verified that the in-degree distribution follows the power law, and we applied a series of graph theoretic algorithms. We extracted Weakly Connected Components, Strongly Connected Components, Biconnected Components, and we computed Global Minimum Cut, Minimum cuts between authority papers in different areas, and shortest paths between pairs of papers. Based on the results, we elicit an aggregate picture of the citation graph in terms of its connectivity. Our attempts to extract research communities with the above standard graph theoretic algorithms and no further heuristics were not successful. The citation graph retained a large well-connected component, pointing to the need for the application of more sophisticated balanced graph partitioning algorithms. It should be pointed out that the balanced graph partitioning problem is NP-hard (Garey, Johnson, and Stockmeyer 1976), but local search strategies have been proposed for them (Kernighan and Lin 1970; Fiduccia and Mattheyses 1982).

Our results suggest that the citation graph shares the same self-similarity (or fractal) properties as the World Wide Web (Dill, Kumar, McCurley, Rajagopalan, Sivakumar and Tomkins 2001). This implies that our conclusions are likely to be applicable to the entire citation graph.

A number of interesting further directions of research are suggested by our study. A major open question is how to autonomously partition the citation graph into communities. 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, and a precise definition of a community in graph-theoretic terms, 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 space.

Fundamental questions about the temporal evolution of the citation graph should be addressed. A study of the temporal evolution of the local link structure of citation graphs can be used for predicting research trends or for studying the life span of specialties and communities. Dynamic models for the citation graph could be developed based on such a study, and such models can, in turn, serve as a tool for prediction and experimentation.

The issue of topic drifting that is present in the citation graphs we extracted needs to be addressed by building a framework that permits the investigation to be carried out on the full citation graph.

A last suggestion for further research is the use of insight about 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.

## About the authors

*Yuan An* completed his Master's in the Faculty of Computer Science of Dalhousie University, and he is pursuing his Ph.D. in the Department of Computer Science, University of Toronto. He can be reached by e-mail yuana@cs.toronto.edu.

*Jeannette Janssen* is a member of the faculty of the Department of Mathematics and Statistics, Dalhousie University. Her current research interests are Combinatorial Optimization, Graph Colouring, Frequency Assignment. She can be reached by e-mail janssen@mscs.dal.ca.

*Evangelos E. Milios* is a member of faculty and Director of the Graduate Program of the Faculty of Computer Science, Dalhousie University. His current research activity is centred on Software agents for Web information retrieval. He can be reached by e-mail eem@cs.dal.ca.

## References

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Broder AZ, Kumar SR, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web: experiments and models. In Proc 9th WWW Conf, pp 309–320

Chakrabarti S, Dom BE, Gibson D, Kleinberg J, Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Mining the link structure of the World Wide Web. IEEE Comput 32:60–67

Chen C (1999) Visualising semantic spaces and author co-citation networks in digital libraries. Inf Proc Manage 35:401–420

Chen C (2001) Visualising a knowledge domain's intellectual structure. IEEE Comput 34:65–71

Dill S, Kumar R, McCurley K, Rajagopalan S, Sivakumar D, Tomkins A (2001) Self-similarity in the web. In 27th Int Conf on Very Large Databases (VLDB2001)

Even G, Naor J, Rao S, Schieber B (1999) Fast approximate graph partitioning algorithms. SIAM J COMPUT. Soc Ind Appl Math 28(6):2187–2214

Fiduccia CM, Mattheyses RM (1982) A linear-time heuristic for improving network partitions. In Proc of the 19th IEEE Des Automation Conf, pp 175–181

Garey M, Johnson D, Stockmeyer L (1976) Some simplified NP-complete graph problems. Theor Comput Sci 1:237–267

Garfield E (1972) Citation analysis as a tool in journal evaluation. Science 178:471–479

Kernighan BW, Lin S (1970) An efficient heuristic procedure for partitioning graphs. The Bell Syst Tech J 29(2):291–307

Lawrence S, Bollacker K, Giles CL (2001) ResearchIndex. NEC Research Institute, http://citeseer.nj.nec.com (accessed on Sep 30, 2001)

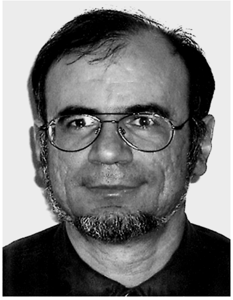
Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Euro Phys J B 4:131–134

Watts D, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

## Author Biographies

**Yuan An** is currently a Ph.D. student in the Department of Computer Science at the University of Toronto working with Prof. John Mylopoulos. He obtained his Master's degree in Computer Science from Dalhousie University, Canada, under the supervision of Prof. Evangelos Milios and Prof. Jeannette Janssen. He holds M. Eng. and B. Eng. degrees from Tsinghua University, China. His research interests include Knowledge Discovery and Management for large and heterogeneous data sets, Conceptual Modelling, and Database techniques on the Web. His current research is on Semantic Encapsulation on supporting Model Interoperation.

**Jeannette Janssen**'s research area is applied graph theory. She has worked on the problem of frequency assignment in cellular and digital broadcasting networks. Her current interest is in graph theory applied to the World Wide Web and other networked information spaces. Dr. Janssen did her Master's studies at Eindhoven University of Technology in the Netherlands, and her doctorate at Lehigh University, USA. She is currently an associate professor at Dalhousie University, Canada.



**Evangelos Milios** received a diploma in Electrical Engineering from the National Technical University of Athens, and Master's and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology. He held faculty positions at the University of Toronto and York University. He is currently a Professor of Computer Science at Dalhousie University, Canada, where he was Director of the Graduate Program. He has served on the committees of the ACM Dissertation Award, and the AAAI/SIGART Doctoral Consortium. He has worked on the interpretation of visual and range signals for landmark-based positioning, navigation and map construction in single- and multiagent robotics. His current research activity is centred on Networked Information Spaces, Web information retrieval, and aquatic robotics. He is a Senior Member of the IEEE.

*Correspondence and offprint requests to*: Evangelos E. Milios, Faculty of Computer Science, Dalhousie University, 6050 University Avenue, Halifax, Nova Scotia B3H 1W5, Canada. Email: eem@cs.dal.ca