# A Geographical Analysis of Knowledge Production in Computer Science

Guilherme Vale Menezes, Nivio Ziviani,
Alberto H. F. Laender, Virgílio Almeida

Computer Science Department
Federal University of Minas Gerais
31270-901 Belo Horizonte, Brazil
{gmenezes, nivio, laender, virgilio}@dcc.ufmg.br

## ABSTRACT

We analyze knowledge production in Computer Science by means of coauthorship networks. For this, we consider 30 graduate programs of different regions of the world, being 8 programs in Brazil, 16 in North America (3 in Canada and 13 in the United States), and 6 in Europe (2 in France, 1 in Switzerland and 3 in the United Kingdom). We use a dataset that consists of 176,537 authors and 352,766 publication entries distributed among 2,176 publication venues. The results obtained for different metrics of collaboration social networks indicate the process of knowledge production has ăchanged differently for each region. Research is increasingly done in teams across different fields of Computer Science. The size of the giant component indicates the existence of isolated collaboration groups in the European network, contrasting to the degree of connectivity found in the Brazilian and North-American counterparts. We also analyzed the temporal evolution of the social networks representing the three regions. The number of authors per paper experienced an increase in a time span of 12 years. We observe that the number of collaborations between authors grows faster than the number of authors, benefiting from the existing network structure. The temporal evolution shows differences between well-established fields, such as Databases and Computer Architecture, and emerging fields, like Bioinformatics and Geoinformatics. The patterns of collaboration analyzed in this paper contribute to an overall understanding of Computer Science research in different geographical regions that could not be achieved without the use of complex networks and a large publication database.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

## General Terms

Experimentation

## Keywords

Coauthorship Networks, Computer Science, Collaboration Social Networks

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
*WWW 2009*, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

## 1. INTRODUCTION

A social network is a collection of people, each of whom is acquainted with some subset of the others [13]. According to Social Science terminology, individuals or groups are called *actors*, and relationships between them are called *ties* [15]. Ties may represent friendship, acquaintance, collaboration, affiliation, or even the transmission of diseases between actors. Social network analysis help us understand the social behavior of these actors.

In scientific *collaboration social networks*, actors are *authors* and a tie exists between two authors if they have already collaborated on the production of some work in a given period of time. In this way, network measures can be employed to obtain information regarding a scientific community specific characteristics, or to compare two or more communities according to these characteristics.

It is important to stress that a collaboration among two authors implies in a certain affinity between them: authors collaborate if they are interested in the same area, if they are affiliated to the same organization, or if they at least speak the same language. Accordingly, if authors publish *intensely* with each other, meaning they have many instances of collaboration in a given period of time, it can be said that they have great affinity.

We can also analyze a social network from a temporal perspective, since in some networks actors and ties may be created or destroyed at any point in time. In this sense, a social network *evolves* over time and its evolving characteristics can be measured and studied during this process. For instance, a collaboration social network has to incorporate the publishing of new papers by adding new authors and collaborations into the network. Each paper has a publishing date, and thus any period of collaboration can be set for the study of collaboration network characteristics.

During the last decades, the advent and popularization of the Web has helped prosper the study of social networks, creating opportunities for the study of social networks in an unprecedented scale. Online social networks such as Facebook[1], MySpace[2], Flickr[3], Orkut[4], among many other examples, are today commonly found and each one have millions of users. Scientific collaboration networks may also be

[1]http://www.facebook.com/.
[2]http://www.myspace.com/.
[3]http://www.flickr.com/.
[4]http://www.orkut.com/.

easily obtained from existing Web services, such as DBLP[5], CiteSeer[6], Google Scholar[7], and Microsoft Libra[8].

This paper analyzes characteristics of collaboration social networks in Computer Science communities, formed by researchers and their publications. The study relies on datasets collected from DBLP. It includes three phases. First, we study measurements of collaboration networks from three regions of the world, namely a network of 8 Brazilian graduate programs, a network of 3 Canadian and 13 US graduate programs, and a network of 2 French, 1 Swiss and 3 British graduate programs. Second, we carry out a temporal analysis of three networks and four different Computer Science fields, over a time span of 12 years. Finally, we analyze a graph of interrelationships among 30 Computer Science subfields.

Computer Science is a field that has distinct features when compared to other traditional research fields, such as Biology, Physics, Chemistry or Mathematics. A Computer Science paper can become obsolete due to new technological breakthroughs in a matter of months. Furthermore, the field is a relatively new one and there is an ongoing effort to fit it in academic institutions' plans. As a consequence, it is important to understand intrinsic characteristics of Computer Science departments and researchers. This paper analyzes the knowledge production process in Computer Science in different geographical regions, using a social network approach.

This paper is organized as follows. Section 2 shows some related work. Section 3 lists the graduate programs addressed in our study and describes the data gathering process. In sequence, Section 4 describes the characteristics of the networks for the three sets of countries and Section 5 discusses the results of the temporal evolution analysis of these three networks. A study of the interrelationship between subfields is presented in Section 6. Lastly, Section 7 presents our conclusions.

## 2. RELATED WORK

In this section, we present a brief overview of some recent work on coauthorship networks. Newman [13] compares graph characteristics of several scientific communities, including Biomedicine, Physics and Computer Science. In general, their clustering coefficient is high and their characteristic path length is short. The digital library community is studied in [10], in which the authors conduct connected component measurements, clustering coefficient measurements, among others. A similar analysis for the Data and Knowledge Engineering journal, the software reverse engineering community, and the SIGMOD conference is presented in [4], [6], and [12] respectively.

Barabasi *et al* [1] analyze coauthorship graphs using a database containing relevant journals in Mathematics and Neuroscience for an 8-year period (1991−1998). The authors infer the dynamic and the structural mechanisms that govern the evolution and topology of this two scientific fields. The results indicate that the network is scale-free and that the network evolution is governed by preferential attachment, affecting both internal and external links.

[5]http://www.informatik.uni-trier.de/∼ley/db/.
[6]http://citeseer.ist.psu.edu/.
[7]http://scholar.google.com/.
[8]http://libra.msra.cn/.

Böner *et al* [2] analyze the impact of coauthorship teams based on the number of publications and their citations on a local and global scale. The authors use a weighted graph representation that encodes coupled author-paper networks as a weighted coauthorship graph. This weighted graph representation is applied to a dataset comprising of 614 articles published by 1,036 unique authors between 1974 and 2004. The reference characterizes the properties and evolution of a new subfield of science: Information Visualization.

Wuchty *et al* [19] study the evolution of the number of authors (team sizes) in publications from science and engineering, social sciences, the arts and humanities, and patents, showing that teams increasingly dominate solo authors in the production of knowledge in the four datasets.

In [8], the authors use the same data source adopted in this paper to study the quality of the top 8 Brazilian graduate programs in Computer Science. They compare the scientific production of the top Brazilian programs with that of reputable North-American and European programs. Furthermore, they also observe a ratio of more than 2 conference papers for each journal article in all programs in Brazil, North-America and Europe, which appears to be an important characteristic of the field.

## 3. DATA GATHERING

Our study addresses the Computer Science graduate programs from the same 30 institutions reported in [8], being 8 from Brazil, 16 from North America (3 from Canada and 13 from the United States), and 6 from Europe (2 from France, 1 from Switzerland and 3 from the United Kingdom). The sample was based on the ease of accessing information about the institutions, but all are included in top positions of existing Computer Science program rankings. They are listed below.

- **Brazil:** Federal University of Minas Gerais, Federal University of Pernambuco, Federal University of Rio de Janeiro, Federal University of Rio Grande do Sul, Pontifical Catholic University of Rio de Janeiro, University of Campinas, University of São Paulo at São Paulo, and University of São Paulo at São Carlos.
- **Canada:** University of British Columbia, University of Toronto, and University of Waterloo.
- **France:** École Polytechnique and Université Pierre et Marie Curie − Paris VI.
- **Switerland:** ETH Zürich.
- **United Kingdom:** Cambridge University, Imperial College, and Oxford University.
- **United States:** Brown University, California Technology Institute, Carnegie Mellon University, Cornell University, Harvard University, University of Illinois at Urbana-Champaign, Massachusetts Institute of Technology, Princeton University, Stanford University, University of California at Berkeley, University of Texas at Austin, University of Washington, and University of Wisconsin.

The data gathering process for our study involved three main steps. In the first step, we extracted from the home pages of the respective institutions the names of the faculty

members of the 30 graduate programs. In the second step, using the names of these faculty members, we collected from DBLP their respective pages, the pages of their coauthors, and the pages of the coauthors of these coauthors, extracting from them the corresponding publication data. All this data was stored in a relational database in order to provide us with a flexible querying environment. Finally, in the third step, we associated each publication venue (conference or journal) in the database to a specific Computer Science subfield. For this, we used a list of 30 Computer Science subfields that reflect the special interest groups of the Brazilian Computer Society, as shown in Table 1.

This data gathering process summed up 176,537 authors and 352,766 publications (conference papers and journal articles), distributed among 2,176 distinct venues. The data was collected from the DBLP repository on June 27, 2007, and refer to articles published between 1954 and 2007. These numbers are summarized in Table 2. Figure 1 shows the number of authors and the number of publications over a period of 12 years, from 1994 to 2006. Notice that the number of papers grows faster than the number of authors in the period, since older authors continue to establish connections over time.

**Table 1: Computer Science Subfields**

| | |
|---|---|
| Algorithms and Theory | Applied Computing |
| Artificial Intelligence | Bioinformatics |
| Circuit Conception | Computer Architecture |
| Computer Graphics | Computer Networks and Distributed Systems |
| Computer Vision | Data Mining |
| Databases | Embedded Systems and Real-Time Systems |
| Formalisms, Logics and Semantics | Games and Entertainment |
| Geoinformatics | Human-Computer Interaction |
| Informatics in Education | Information Retrieval |
| Information Systems | Machine Learning |
| Modelling and Simulation | Natural Language Processing |
| Operation Systems | Operational Research and Optimization |
| Programming Languages | Robotics, Automation and Control |
| Security and Privacy | Software Engineering |
| Ubiquitous Computing | Web, Hypermedia Systems, Multimedia |

# 4. COAUTHORSHIP NETWORKS

In this section we describe the three coauthorship networks, generated from data gathered from DBLP, and present some statistics about them. Before that, we discuss some social network fundamentals.

**Table 2: Data Summary**

| | |
|---|---|
| Programs | 30 |
| Authors | 176,537 |
| Publications | 352,766 |
| Venues | 2,176 |

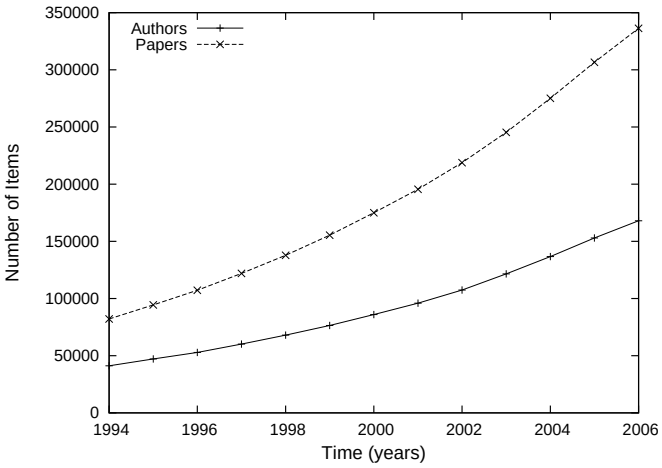


**Figure 1: Number of Authors and Number of Papers over Time**

## 4.1 Fundamentals

Collaboration social networks are modeled as undirected, unweighted graphs, in which there is an edge between two actors if they have collaborated at least once during a certain period of time. In a coauthorship network, nodes represent authors, and two authors are connected by an edge if they have coauthored one or more papers, as shown in Figure 2.

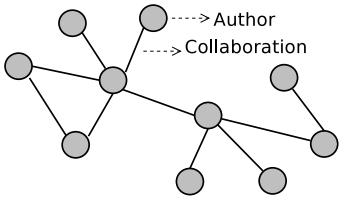


**Figure 2: A Coauthorship Network**

The modelling of a coauthorship network as a graph allows the use of some interesting measures. A possibility is the connected component analysis, which consists in measuring the size of the *giant component* (or the largest connected component) [15]. A *connected component* is a maximal connected subgraph. Two vertices are in the same connected component if and only if there exists a path between them. The measure consists in computing the fraction of graph vertices that are part of the giant component.

The *distance* between two nodes of a network is the length of the shortest path between them. The *average distance* over all pairs of nodes represents the average separation of

the coauthorship network and characterizes its interconnectedness. Another measure is the *graph diameter*, which is the longest distance between any two vertices of the graph. If we think of an edge as a means for conducting information between actors, the information can be spread more rapidly among actors in a graph with smaller diameter.

A measure related to the graph diameter is the graph *characteristic path size* [15] (or average path size) calculated by computing the shortest path between every pair of vertices and calculating their average. A graph with a smaller characteristic path size can also conduct information faster than a graph with a long characteristic path size.

Social networks typically have a small characteristic path size, which classify them as *small-world* networks. The small world phenomenon was identified for the first time in [11], in which the author obtained evidence that any two individuals in the United States are separated by a path formed by 6 individuals, on average. This phenomenon was later called the *six degrees of separation* phenomenon. In this context, a small characteristic path size is orders of magnitude smaller than the number of graph vertices.

In the context of a coauthorship social network, the *clustering coefficient* measures the degree of transitivity in the *publishes with* relationship that defines the graph. As an example, if the clustering coefficient is high, there is a high probability that a relationship between authors A and B, and another relationship between authors B and C will induce a relationship also between authors A and C. In other words, the clustering coefficient of a node in the coauthorship network indicates how much an author's collaborators are willing to collaborate with each other. In this sense, the clustering coefficient indicates the existence of ordering in a local level [17]. More formally, the clustering coefficient of a node $i$ can be defined as

$$C_i = \frac{\text{Number of triangles connected to vertex i}}{\text{Number of triples centered on vertex i}}.$$

The global clustering coefficient is the average clustering coefficient of all nodes in the network, which can take values between 0 and 1 [18].

The *assortative mixing* of a network refers to the preference of vertices with high degree to be connected to other vertices with high degree. The assortative mixing of a network can be quantified using a connected degree-degree correlation function [3], in which a positive number indicates an assortative network, a negative number indicates a disassortative network (i.e., vertices with high degree connect to other vertices with low degree), and 0 indicates a network without assortativity. Social networks are known to be assortatively mixed, while technological and biological networks tend to be disassortative [14].

## 4.2 Statistics of the Three Networks

The three generated networks represent Brazilian, North-American and European programs, from now on referred to as Br, Ca-US and Fr-Sw-UK networks, respectively. The basic statistics of these networks are summarized in Table 3. The largest network is the Ca-US network with 1,008 authors and 40,039 papers over a 12-year period. The number of papers per author is much larger for researchers in US and Canada, dwarfing the two other networks. The number of papers per author was calculated by averaging the number of papers for each author in the dataset. Notice that this is different from simply calculating the division between the number of papers and the number of authors, since each paper can have more than one author. The Ca-US network has the highest number of papers per author (45.89), much higher than the the Fr-Sw-UK (19.85) and the Br (16.06) measurements.

The number of authors per paper exhibit a similar pattern for the three networks, ranging from 2.87 to 3.21. It is worthnoting that these numbers vary significantly across different subject areas. In [16], the number of author per papers in Biology, Physics and Mathemetics vary from 5.1 to 6.9, reflecting differences in the way research is done in those fields. The number of callaborators an author has in the Br and Fr-Sw-Ok networks are very close (18.64 and 17.78, respectively), whereas in the Ca-Us network this number is more than two times higher (42.11). This is presumably a result of the way computer scientists work in US and Canada, where there exists many links among universities and companies.

**Table 3: Statistics for the Br, Ca-US and Fr-Sw-UK Networks**

| | Br | Ca-US | Fr-Sw-UK |
|---|---|---|---|
| Number of Authors | 357 | 1,008 | 488 |
| Number of Papers | 4,405 | 40,039 | 8,764 |
| Papers per Author | 16.06 | 45.89 | 19.85 |
| Authors per Paper | 3.21 | 2.87 | 2.77 |
| Average Collaborators | 18.64 | 42.11 | 17.78 |
| Giant Component | 78.15% | 78.27% | 26.17% |
| Average Giant Component for Isolated Programs | 67,41% | 56,46% | 30,01% |
| Average Path Length | 6.47 | 4.42 | 6.18 |
| Diameter | 16 | 12 | 15 |
| Clustering Coefficient | 0.30 | 0.20 | 0.38 |
| $\alpha_{DegreeDistribution}$ | 1.51 | 1.77 | 1.89 |
| Assortativity | 0.25 | 0.35 | 0.38 |

Now we present a comparative analysis of structural features of the three networks, by comparing measurements that characterize the network structure. The number of nodes of the the giant component plays an important role in coauthorship networks, since it represents the portion of the authors that are connected via collaboration. Any author in the giant component can be reached from any other by traversing a path of intermediate coauthors. The figure of 78% for the size of giant component in Ca-US and Br indicates that the majority of Computer Science authors in these regions are connected, avoiding the proliferation of small isolated communities.

In contrast, the giant component size in Fr-Sw-UK is only 26%. One possible explanation is the existence of graduate programs from different countries (with different languages) in this network, which makes the integration between dif-

ferent communities more difficult. However, we have also measured the size of the giant component inside isolated graduate programs, disconsidering relationships with other programs, and we have found that the average size of the giant component inside the programs in the Fr-Sw-UK network is also considerably smaller when compared with Br and Ca-US networks. This may be an indication that the formation of isolated communities is an intrinsic characteristic of scientific collaboration in the Fr-Sw-UK network.

According to [16], the average distance over all pairs of nodes in Biology is 4.6 whereas in Mathematics is 7.6. Likewise, our three networks show an average distance that varies from 4.4 to 6.4, much smaller than their size, indicating the presence of "small world" effects [11]. The diameter of the three networks, i.e., the largest shortest path of the networks, is between 12 and 16.

The average clustering coefficient over the Ca-US network is 0.20, whereas the clustering coefficient in the Fr-Sw-UK network is higher, at 0.38. Both values are typical of social networks [15]. The low clustering coefficient of the Ca-US network indicates a high level of collaboration between authors, which is likely a sign of both the increasing in the number of nodes and the multidisciplinarity of the authors, since it is likely that two collaborators that do not collaborate among themselves belong to different fields.

The presence of positive assortative mixing (i.e., positive degree correlations) between adjacent nodes suggests that coauthorship networks can be largely understood in terms of the organization of nodes into communities of collaboration, a feature that can explain, to some extent, the observed values for the clustering coefficient and assortativity.

A network metric that has been much studied for various networks is the degree distribution, $P(k)$, giving the fraction of nodes that have $k$ edges. Figure 3 characterizes the degree $k$ distribution for the Ca-US network as a power law $P(k) \propto 1/k^{\alpha}$, with exponent $\alpha = 1.77$. Br and Fr-Sw-UK networks follow similar distributions with exponents $\alpha = 1.51$ and 1.89, respectively. Networks for which $P(k)$ has a power-law tail are known as scale-free networks, indicating that a small number of authors concentrate a very large number of collaborations whereas a large number of authors have a few coauthors. The largest is $\alpha$, the largest is the difference between the more frequent and the less frequent degree values.

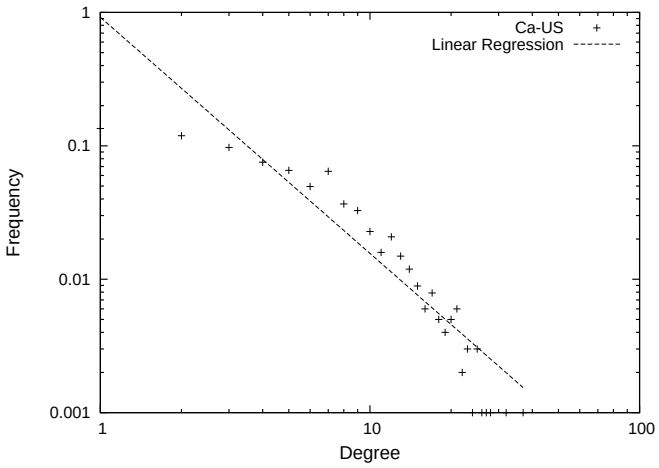


**Figure 3: Degree Distribution for Ca-US Network (Log-Log Scale)**

## 5. TEMPORAL EVOLUTION

We performed an analysis of the Br, Fr-Sw-UK and Ca-US networks over a span of 12 years, from 1994 to 2006. Our data was obtained in June 2007, which means the last complete year in our dataset is 2006. We considered that a period of 12 years is enough to show the recent trends in the evolution of these networks.

### 5.1 Evolution of the Three Networks

Our first step was to measure the cumulative number of vertices over time in the three networks. The number of vertices for each year is the union of the active vertices for that year and for every year that precedes it. An author that published a paper during a given year is represented by a vertice in the network relative to that year. The results are shown in Figure 4. Notice that the number of vertices in the Ca-US network is always greater than the number of vertices in the two other networks. In addition, in 1998 the Br and Fr-Sw-UK curves cross over, reflecting a faster increase in the number of vertices in the Br network between 1996 and 2000. One possible reason is the increased effort in the assessment of the quality of the Brazilian programs in the period, lead by CAPES[9], a Brazilian Ministry of Education's agency, which encouraged Brazilian authors to publish in international venues. The popularization of the Web may also have had an influence in this increase. The cumulative number of edges for the three networks is shown in Figure 5, which shows a close resemblance to Figure 4.

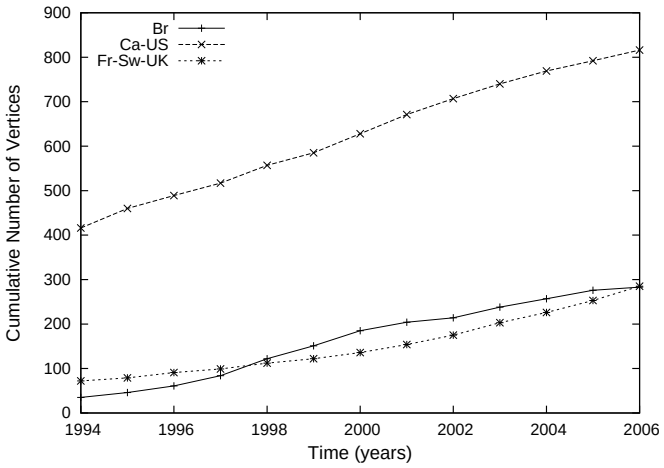


**Figure 4: Cumulative Number of Vertices for Br, Fr-Sw-UK and Ca-US Networks in the Period**

Figure 6 plots the number of vertices versus the number of edges for the three networks. Each point of the log-log graph represents the number of edges and vertices at year $t$. Notice that the number of edges grows faster than the number of vertices with time, which shows that the formation of new relations benefits not only from the inclusion of new vertices in the network, but also from the existing structure of the network, formed by vertices previously inserted. This faster

[9]http://www.capes.gov.br

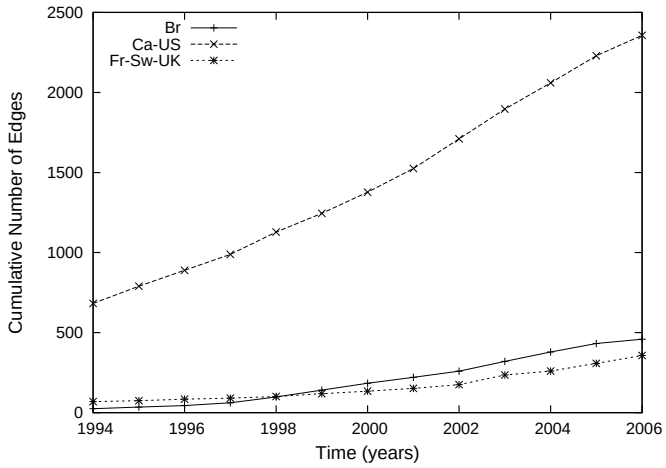


**Figure 5: Cumulative Number of Edges for Br, Fr-Sw-UK and Ca-US Networks in the Period**

growth is more evident in the Ca-US network, with inclination $\alpha = 1.84$, followed by the Br network, with inclination $\alpha = 1.39$, and by the Fr-Sw-UK network, with inclination $\alpha = 1.23$. In particular, the densification of the networks is not arbitrary; as the coauthorship evolves over time, they follow a version of the relation $e(t) \sim n(t)^{\alpha}$, where $e(t)$ and $n(t)$ denote the number of edges and nodes of the graph at time t and $\alpha$ is the exponent. In this case, $\alpha$ is larger than 1, indicating a deviation from linear growth, which means that the number of relations grows faster than the number of authors [9].

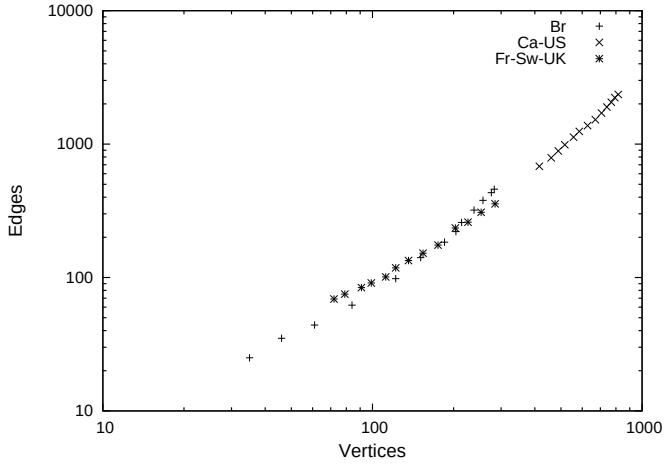


**Figure 6: Evolution of Number of Edges versus Number of Vertices for the Br ($\alpha = 1.39$), Ca-US ($\alpha = 1.84$) and Fr-Sw-UK ($\alpha = 1.23$) Networks (Log-Log Scale)**

The relative giant component size for Br, Ca-US and Fr-Sw-UK networks can be seen in Figure 7. The moderated growth of the Ca-US giant component (81% in 1994 to 96% in 2006) contrasts with the fast growth of the Br and Fr-Sw-UK giant component in the period. In special, the growth of the giant component in the Br network has been steep starting from 1999 (28% to 98%). The networks considered in the evolution study are formed by active faculty members in the period, which means there are no isolated vertices in the networks. Furthermore, in our study we chose only highly-ranked graduate programs, which justifies the giant component covering almost the whole network.

The Fr-Sw-UK giant component covers only 42% of the network in 2006, which may reflect the fact that its programs are from different countries, with different languages. However, a smaller giant component is also evident inside isolated programs, as shown in Table 3. As stated before, this may indicate that this event may not be only due to the diversity of countries, but also from an intrinsic characteristic of these programs.

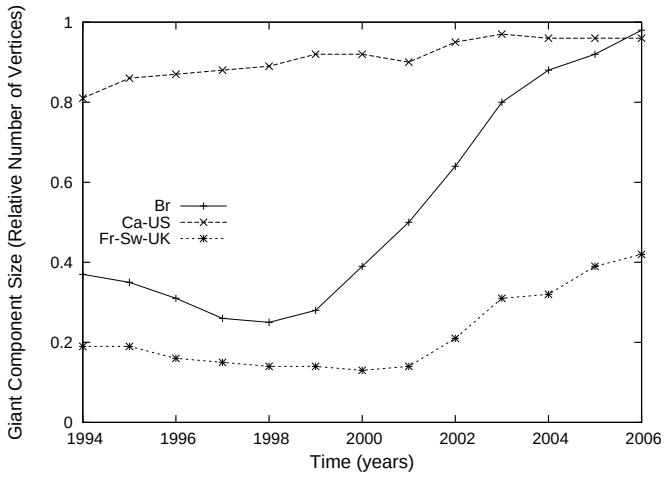


**Figure 7: Giant Component for Br, Ca-US and Fr-Sw-UK Cumulative Networks in the Period**

The average path length for the three cumulative networks is show in Figure 8. We observe that the growth of the Br and Fr-Sw-UK path lengths happens during the same period (1998 to 2002), which coincides with the accelerated growth of the giant component (Figure 7). When the graph components start to merge together, they are still weakly connected, and thus the average path length grows. As time passes, the two groups become more knit together, causing the reduction of the giant component and the average path length. For example, the Fr-Sw-UK giant component is still growing, as well as its path length. This is also seen in [5]. Despite the growth of the Br and Fr-Sw-UK path lengths, we see that the curve of Ca-US decreases steadily. This phenomenon, called shrinking diameter, reflects the fact that the effective shortest path decreases as the network grows, and has been observed in [9] in a range of real networks.

Figure 9 shows the average clustering coefficient over time for the three cumulative networks. From 1996 to 2006 the Br clustering coefficient has been continuously reduced (0.59 to 0.26), which may be related to the inclusion of new authors into the network and to the increased collaboration between authors of different communities. The Fr-Sw-UK clustering coefficient has remained relatively steady over the period, while the Ca-US clustering coefficient has slightly reduced. The Fr-Sw-UK network has the highest clustering coefficient, which reflects its smaller giant component, meaning that there are isolated communities collaborating internally with intensity.

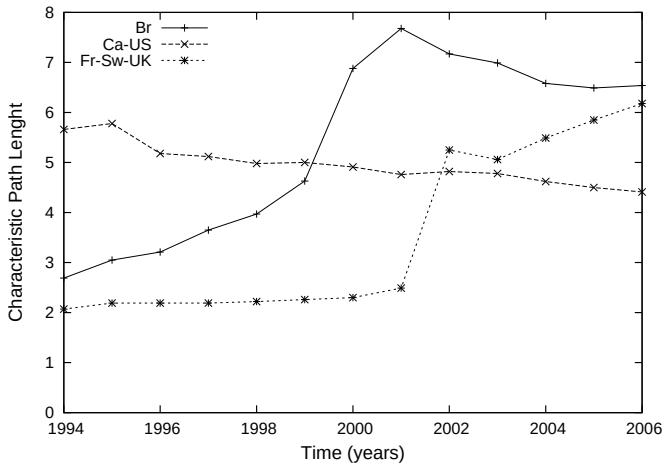


**Figure 8: Characteristic Path Lenght for Br, Ca-US and Fr-Sw-UK Cumulative Networks in the Period**

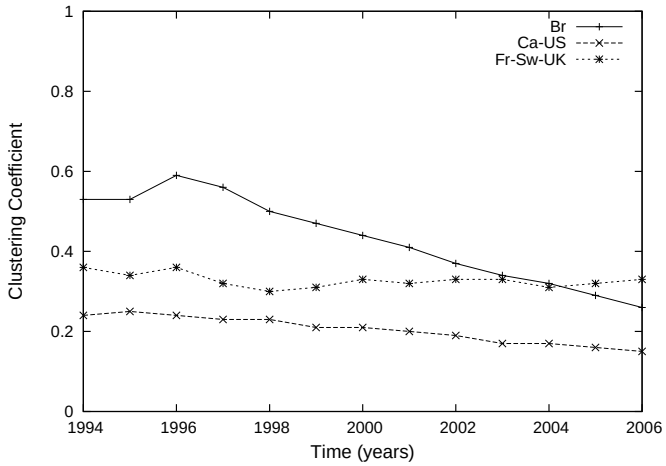


**Figure 9: Clustering Coefficient for Br, Ca-US and Fr-Sw-UK Cumulative Networks in the Period**

The average number of authors per paper for the three networks is shown in Figure 10. The increase in the average number of authors collaborating on papers indicates that in recent years researchers tend to participate in teams of increasing size. In [19] the authors observed an increase in the research team sizes of the science and engineering field, the social sciences field, and the arts and humanities field. They also observed an increase in the number of inventors per patent during the last decades. A possible reason is the reduction of communication costs with the rise and growth of the Internet.

The Br network had a sudden increase in the average number of authors per paper from 1998 to 1999, which is due to a shift in the Brazilian governmental policy towards research support at the time, favoring team financing instead of individual financing. Furthermore, the incresing preasure to publish more in national and international high-quality vehicles is likely to have intensified collaboration in Brazil.

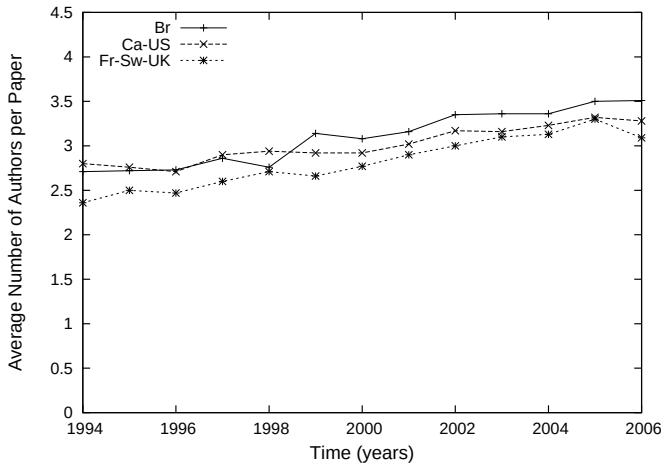


**Figure 10: Average Number of Authors Per Paper for Br, Ca-US and Fr-Sw-UK Networks for Each Year**

## 5.2 Evolution of Computer Science Subfields

In the next step, we studied the evolution of Computer Science subfields over time (see Table 1), observing two well-established subfields, namely, Computer Architecture and Databases, and two emerging ones, namely, Bioinformatics and Geoinformatics. Figure 11 displays the evolution in the number of vertices per field over time. All the four subfields have continuously grown from 1994 until 2006. In special, Bioinformatics had a great relative growth in the period, from 147 vertices in 1994 to 1,394 vertices in 2006. Geoinformatics has also grown (50 to 177 vertices), however, its reduced proportion makes this difficult to visualize in the graph. The same is true for the evolution in the number of edges for the four subfields, shown in Figure 12. Notice that Bioinformatics had a fast increase in the number of edges from 1999 onwards, indicating the establishment of many new connections between authors in this period.

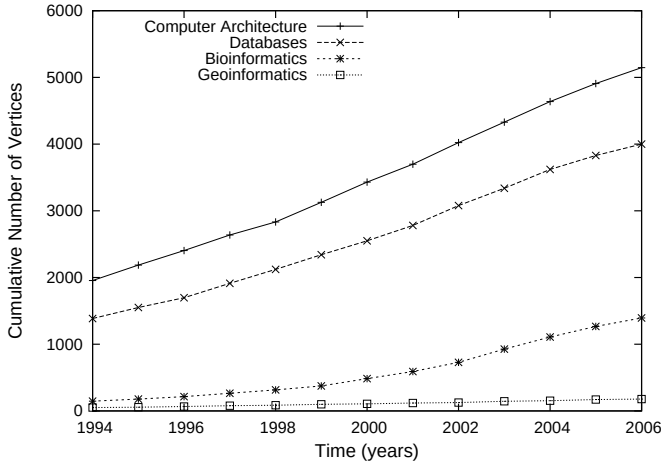


**Figure 11: Cumulative Number of Vertices per Field in the Period**

The clustering coefficient evolution for Computer Architecture, Databases, Bioinformatics and Geoinformatics is

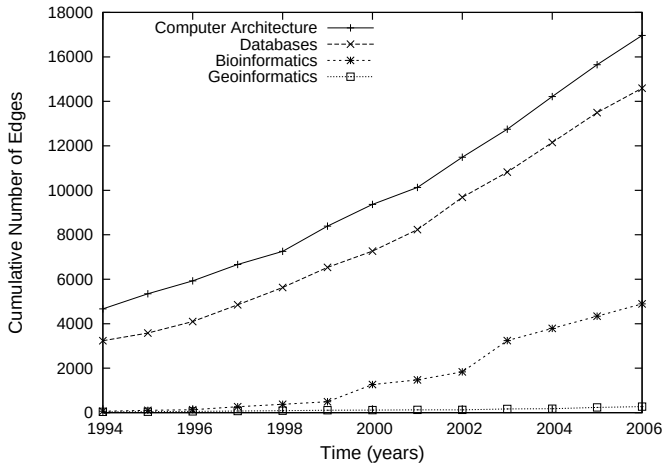


**Figure 12: Cumulative Number of Edges per Sub-Area in the Period**

shown in Figure 13. The clustering coefficient measure reflects the division between well-established and emerging subfields: the Computer Architecture and Databases subfields show a slight increase in the clustering coefficient measure with time, while the Bioinformatics and Geoinformatics subfields have a much faster increase. Moreover, the increase in the measure indicates a process of densification of these subfields, in which more edges are being inserted into their networks. This increase contrasts with the decrease observed in the clustering coefficient of networks which involve all subfields (see Figure 9).

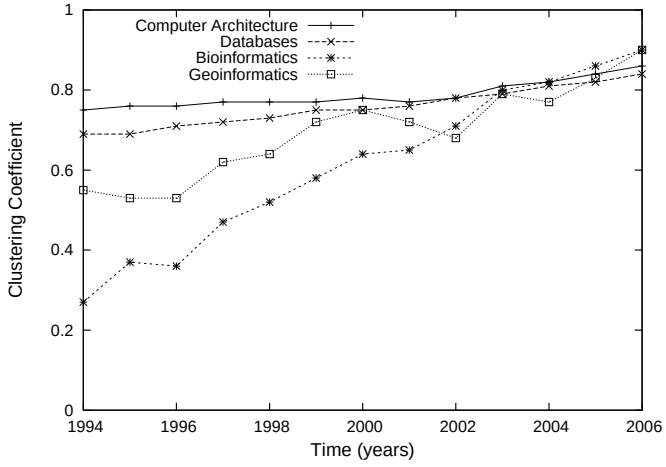


**Figure 13: Clustering Coefficient per Field in the Period**

Figure 14 displays the average number of subfields in which each author publishes in a given single year. These values are *not* cumulative over time. The graph shows clearly a trend of diversification, i.e., an increase in the number of subfields per author over time. Furthermore, in general, the number of subfields per author is larger for the Ca-US network and smaller for the Br network.

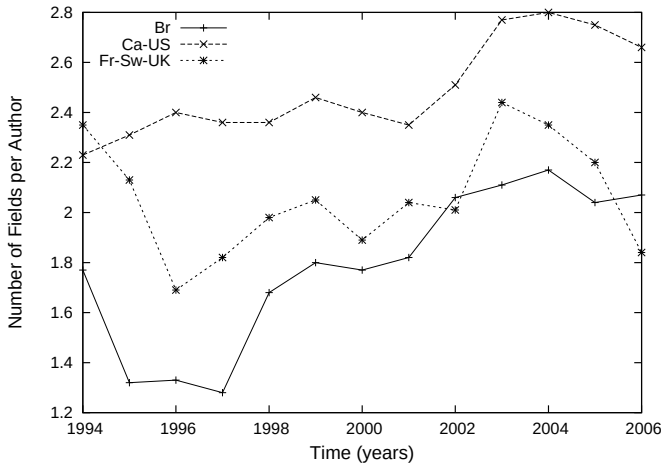


**Figure 14: Average Number of Fields per Author for Br, Ca-US and Fr-Sw-UK Networks for Each Year**

# 6. MAPPING COMPUTER SCIENCE SUBFIELDS

In the final part of our study we analyzed the interrelationship between Computer Science subfields. If we consider a subfield as a set of authors who publish in it, the proximity between subfield $A$ and subfield $B$ can be computed as follows:

$$P_{AB} = \frac{|R_{AB}|}{\sum_K |R_{AK}|} + \frac{|R_{AB}|}{\sum_K |R_{BK}|}$$

Lets define the union of the edges incident to elements of $A$ as $I(A)$. Then, $R_{AB}$ is the intersection between $I(A)$ and $I(B)$. We excluded all edges with $P_{AB} < 0.08$. We used the algorithm described in [7] and implemented in Pajek[10] for the generation of a graph for visualization of the interrelationships. The vertices represent Computer Science subfields, edges represent relationships between them, and the edge weights are computed using the proximity function between fields, $P$.

Figure 15 presents the interrelationship among the 30 subfields considered. The size of the vertices is an indication of the number of authors in the community representing each subfield. The smallest community is Geoinformatics, with 177 authors, and the largest is Computer Architecture, with 5,239 authors. The distance between vertices denotes the intensity of the inter-relationship among subfields. Notice that there is a clear division of the graph into two sets of subfields. For example, the upper part of the figure displays subfields related to Computer Systems, such as Computer Networks, Computer Architecture, and Operating Systems. Software Engineering and Algorithms and Theory also belong to this group. In turn, the bottom part of the figure shows subfields which are closely related to Databases, such as Web, Hypermedia and Multimedia, Information Retrieval, Data Mining, and Information Systems. It also includes subfields related to Artificial Intelligence, Computer Graphics and Computer Vision.

[10]Pajek – Program for Large Network Analysis.

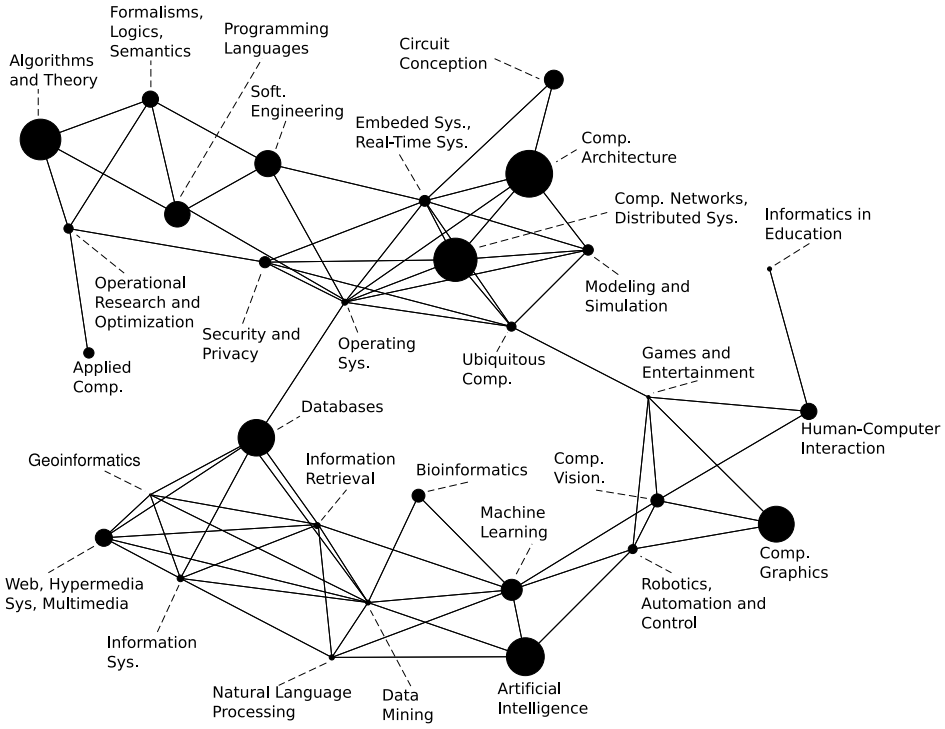


**Figure 15: Interrelationship Between Fields**

## 7. CONCLUSIONS

In this paper, we used collaboration networks to analyze the scientific production in Computer Science. We formed networks for 30 Computer Science graduate programs in three regions and studied their characteristics, using several metrics. Furthermore, we defined 30 different subfields and studied the characteristics of their collaboration networks. A temporal study was also performed, using a period of 12 years, from 1994 to 2006. Next, we visually studied the inter-relationship between different Computer Science subfields.

The Ca-US network has shown a much larger number of papers per author and of collaborators per author. It also has the lowest clustering coefficient, indicating diversified collaboration and a low transitivity. The Fr-Sw-UK network has a small giant component size when compared to the other two networks, which could be related to the existence of programs from different countries in this network, but also could be a characteristic intrinsic to the Fr-Sw-UK programs, as shown in Table 3. The average path length and diameter in the Ca-US network is smaller, which demonstrates it is a denser network, since it is also the largest in number of vertices. Furthermore, the degree distribution suggests a smaller difference from the authors that publish more to the authors that publish less.

The temporal evolution has shown a rapid increase in the scientific production of Br programs from 1997 until 2000, which may be due to Brazilian government agencies' increased efforts to assess the production of the country's researchers in the period. In addition, the Br and Fr-Sw-UK networks have shown a fast increase in the size of their giant component and in the size of their average path length from 1998 onwards, especially in the former. The Br network has also shown a significant reduction in its clustering coefficient in the period.

An analysis of the evolution of two well-established Computer Science subfields (Computer Architecture and Databases) and two emerging subfields (Bioinformatics and Geoinformatics) was also performed. The study has shown a significant increase in the size of the Bioinformatics network in the period. The two emerging subfields have also had an increase in the clustering coefficient measure, indicating they are becoming denser in a fast pace.

Finally, we generated a graph for visualization of the interrelationships between subfields, in which vertices represent subfields, the vertice size represents the size of the subfield in number of authors, and the proximity between vertices represents the intensity of the interrelationship between these vertices. A visual analysis showed a division of Computer Science into separated sets of subfields. For instance, subfields related to Computer Systems, such as Computer Networks, Computer Architecture, and Operating Systems, are displayed in a set, while subfields related to Databases, such as Web, Hypermedia and Multimedia, Information Retrieval, Data Mining, and Information Systems, are displayed in another. However, it should be noticed that this division does not necessarily imply that these subfields are related to each other, but only that their researchers are more likely to work together.

## 8. ACKNOWLEDGEMENTS

This reserach is partially supported by the Brazilian National Institute of Science and Technology for the Web (grant MCT/CNPq 573871/2008-6), by the InfoWeb project (grant MCT /CNPq/CT-INFO 550874/2007-0), and by the authors' individual grants from CNPq.

## 9. REFERENCES


[1] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *PHYSICA A*, 311:3, 2002.

[2] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams: Research articles. *Complex.*, 10(4):57–67, 2005.

[3] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz. Are randomly grown graphs really random? *Physical Review E*, 64:041902, 2001.

[4] C. Chen, I.-Y. Song, X. Yuan, and J. Zhang. The thematic and citation landscape of data and knowledge engineering (1985-2007). *Data Knowl. Eng.*, 67(2):234–259, 2008.

[5] E. Elmacioglu and D. Lee. On six degrees of separation in dblp-db and more. *SIGMOD Rec.*, 34(2):33–40, 2005.

[6] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: characterizing and modeling network evolution. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 107–116, New York, NY, USA, 2008. ACM.

[7] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, 31(1):7–15, 1989.

[8] A. H. F. Laender, C. J. P. Lucena, J. C. Maldonado, E. Souza e Silva, and N. Ziviani. Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs. *ACM SIGCSE Bulletin*, 40:135–145, June 2008.

[9] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187, New York, NY, USA, 2005. ACM.

[10] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Inf. Process. Manage.*, 41(6):1462–1480, December 2005.

[11] S. Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[12] M. A. Nascimento, J. Sander, and J. Pound. Analysis of SIGMOD's co-authorship graph. *SIGMOD Rec.*, 32(3):8–10, 2003.

[13] M. E. Newman. The structure of scientific collaboration networks. *Proc. Nat'l Acad. Sci. of the United States of America*, 98(2):404–409, January 2001.

[14] M. E. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, October 2002.

[15] M. E. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[16] M. E. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proc. Nat'l Acad. Sci. of the United States of America*, pages 5200–5205, 2004.

[17] D. Watts. Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, 105(2):493–527, September 1999.

[18] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, June 1998.

[19] S. Wuchty, B. F. Jones, and B. Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, May 2007.