Collaboration in Computer Science: A Network Science Approach

Massimo Franceschet
Department of Mathematics and Computer Science, University of Udine, Via delle Scienze 206, 33100 Udine, Italy, E-mail: massimo.franceschet@uniud.it

Co-authorship in publications within a discipline uncovers interesting properties of the analyzed field. We represent collaboration in academic papers of computer science in terms of differently grained networks, namely affiliation and collaboration networks. We also build those sub-networks that emerge from either conference or journal co-authorship only. We take advantage of the network science paraphernalia to take a picture of computer science collaboration including all papers published in the field since 1936. Furthermore, we observe how collaboration in computer science evolved over time since 1960. We investigate bibliometric properties such as size of the discipline, productivity of scholars, and collaboration level in papers, as well as global network properties such as reachability and average separation distance among scholars, distribution of the number of scholar collaborators, network resilience and dependence on star collaborators, network clustering, and network assortativity by number of collaborators.

Introduction

Collaboration is a fundamental and increasingly common feature in scientific research. Collaborative research has been associated with higher productivity, with higher impact, and, ultimately, with higher quality: from an economic perspective, collaboration allows the division of labor leading to reduced costs and time saving, facilitate access to scientific funding, to expensive (possibly large-scale) equipment, and to unique scientific data. From a bibliometric perspective, collaborative works are generally more visible and more cited by other scholars; moreover, they are rated higher by peer reviewers with respect to papers written in isolation, although notable exceptions exist (Franceschet & Costantini, 2010).

In this article, we study collaboration in computer science using a network science approach. The field of *network science*—the holistic analysis of complex systems through the study of the structure of networks that wire their components—exploded in the last decade, boosted by the availability of large databases on the topology of various real networks, mainly the web and biological networks (Newman, 2010). The network science approach has been successfully applied to analyze disparate types of networks, including technological, information, social, and biological networks. Here, we use co-authorship in publications as a proxy for scientific collaboration and build two differently grained network representations of collaboration in computer science: an author–paper *affiliation network*, which is a bipartite graph with two types of nodes for authors and papers and links running from authors to papers that they wrote. We use affiliation networks to investigate bibliometric properties such as the number of published papers, the number of active scholars, the distribution of scientific productivity of scholars, and the distribution of collaboration level in papers.

A coarser and highly informative alternative representation is the *collaboration network*, in which the nodes represent authors and the links are collaborations between authors in publications. A collaboration network is a type of social network, since co-authorship in publication can be interpreted as a social relationship between authors: in most cases, two authors who have written a paper together do know each other quite well, at least from a scientific perspective. This is particularly true in disciplines like computer science where the typical paper has few co-authors and the share of single-authored papers is not large.[1] We study the large-scale structure of the collaboration network for computer science, investigating properties such as reachability and average separation distance among scholars, distribution of the number of scholar collaborators, network resilience and dependence on star collaborators, network clustering, and network assortativity by number of collaborators.

In the computer science publication culture, conferences are important publication sources, and journals often publish deeper versions of papers already presented at conferences. This is a peculiarity of computer science that makes it an

Received May 3, 2011; revised June 23, 2011; accepted June 23, 2011

 DOI: 10.1002/asi.21614

original research discipline: in all other sciences, indeed, journals are the primary publication source, whereas monographs are the standard publication type in many social sciences, arts, and humanities. This singularity motivated us to analyze separately two sub-networks of the whole collaboration graph, namely the *conference* and the *journal* collaboration networks. It is worth observing that the role of conferences in computer science is currently heartily discussed in the computer science literature. Conferences have the undeniable advantages of providing fast and regular publication of papers and of bringing researchers together by offering the opportunity to present and discuss the paper with peers. These peculiar features of conferences are particularly important because computer science is a relatively young and fast evolving discipline (Choppy, van Leeuwen, Meyer, & Staunstrup, 2009). Nevertheless, *Communications of the ACM*, the flagship magazine of the Association for Computing Machinery, recently published a series of thought-provoking Viewpoint columns and letters that swim against the tide (Birman & Schneider, 2009; Crowcroft, Keshav, & McKeown, 2009; Fortnow, 2009; Franceschet, 2010; Vardi, 2009).

The paper is divided into two parts. The first part of the study is atemporal: we investigate bibliometric and network properties of the cumulative affiliation and collaboration networks considering all papers published in computer science since 1936. The second part of our contribution is a longitudinal (time-resolved) study of collaboration in computer science: we observe how bibliometric and network properties evolved in time since 1960. While the first investigation provides a *static* picture, the second analysis gives a *dynamic* perspective on collaboration in computer science.

Related Literature

Academic collaboration has been extensively studied in *bibliometrics*, the branch of information and library science that quantitatively investigates the process of publication of research achievements. Bibliometricians observed that collaboration intensity neatly varies across disciplines (Franceschet & Costantini, 2010; Larivière, Gingras, & Archambault, 2006; Moody, 2004). The intensity of research collaboration is negligible in arts and humanities, while social scientists often work in team, but collaborations are smaller in scale and formality compared with science disciplines. By contrast, collaborative work is heavily exploited in science, in particular, physics, medicine, and biology. Collaboration is, however, moderate in mathematics, computer science, and engineering. Most studies exploring the connection between research collaboration (taking co-authorship as a unit of measurement) and citational impact have pointed out a positive correlation between the two variables (Asknes, 2003; Franceschet & Costantini, 2010; Goldfinch, Dale, & Rouen, 2003; Persson, Glänzel, & Danell, 2004). Furthermore, collaborative works are generally valued higher by peer experts (Franceschet & Costantini, 2010; Lawani, 1986; Presser, 1980). Both impact and quality of papers are further enhanced when the affiliations of authors are heterogeneous (Goldfinch et al., 2003; Katz & Hicks, 1997). Interestingly, in computer science a little collaboration, but not more than that, seems fruitful to produce more valuable papers (Franceschet & Costantini, 2010).

Collaboration has been also investigated under the network analysis umbrella. Sociologists have the longest tradition of quantitative study of social networks (Davis, Gardner, & Gardner, 1941; Moreno, 1934; Scott, 2000; Wasserman & Faust, 1994). However, the notion of an academic collaboration network, a particular type of social network, first appeared in 1969 in a brief note by mathematician Goffman (1969). Goffman defined the *Erdős number* for a given mathematician as the length of the shortest path on the mathematics collaboration network connecting the mathematician with Paul Erdős.[2] The idea of Erdős number and hence of collaboration network, however, was informally already present in the mathematics community before 1969, since in Goffman's note we can read:

> *I was told several years ago that my Erdős number was 7. It has recently been lowered to 3. Last year I saw Erdős in London and was surprised to learn that he did not know that the function v(Erdős; .) was being considered. When I told him the good news that my Erdős number had just been lowered, he expressed regret that he had to leave London the same day. Otherwise, an ultimate lowering might have been accomplished.*

The note of Goffman is followed by a brief series of (occasionally sarcastic) by colleagues of his, including one written by Paul Erdős himself, speculating on some theoretical properties of the collaboration graph in mathematics (Erdős, 1972; Harary, 1971; Odda, 1979).

Newman was the first to experimentally study large-scale collaboration networks with the aid of a modern network analysis tool-kit. He analyzed the structural properties of collaboration networks for biomedicine, physics (Newman, 2001b,c), and mathematics (Newman, 2004), as well as the temporal evolution of collaboration networks in physics and biomedicine (Newman, 2001a). Barabási et al. (2002) studied the evolution in time of collaboration networks in neuroscience and mathematics. The temporal dynamics of mathematics collaboration networks is also investigated by Grossman (2002). Moody (2004) studied the structure and the temporal evolution of a social science collaboration network.

As for studies concerning the computer science collaboration network, Huang, Zhuang, Li, and Giles (2008) considered publications from 1980 to 2005 extracted from the CiteSeer digital library. The dataset consists of 451,305 papers authored by 283,174 distinct researches. The authors studied properties at both the network level and the community level and how they evolve in time. Bird et al. (2009) focused on the structure and dynamics of collaboration in research communities within computer science. They isolated 14 computing areas, selected the top tier conferences for each area, and extracted publication data for the chosen conferences from DBLP 2008. The dataset contains 83,587 papers, 76,598 authors, and 194,243 collaboration pairs. They used

DOI: 10.1002/asi

TABLE 1. Structural properties of discipline collaboration networks.

Disc	Source	Nodes	Edges	Deg	Com	Dis	Dia	Tra	Clu	Mix
MAT	Mat. Rev.	253,339	496,489	3.92	0.82	7.57	27	0.15	0.34	0.12
PHY	arXiv	52,090	245,300	9.27	0.84	6.19	20	0.45	0.56	0.36
BIO	Medline	1,520,251	11,803,064	15.53	0.92	4.92	24	0.09	0.60	0.13
NEU	–	209,293	–	11.54	0.91	6.00	–	–	0.76	–
SOC	Soc. Abs.	128,151	–	–	0.53	9.81	–	0.19	–	–
CS	CiteSeer	283,174	–	5.56	0.66	7.10	26	–	0.63	0.28
CS	DBLP	688,642	2,283,764	6.63	0.85	6.41	23	0.24	0.75	0.17
CS-C	DBLP	503,595	1,584,108	6.29	0.85	6.54	23	0.24	0.75	0.16
CS-J	DBLP	356,822	987,059	5.53	0.77	7.26	25	0.37	0.77	0.30

Column names are abbreviated as follows: disc (discipline: MAT (mathematics (Grossman, 2002; Newman, 2004, 2010)), PHY (physics (Newman, 2001b, 2010)), BIO (biomedicine (Newman, 2001b, 2010)), NEU (neuroscience (Barabási et al., 2002)), SOC (social science (Moody, 2004)), CS (computer science (Huang et al., 2008) and this paper), CS-C (computer science conferences; this paper), CS-J (computer science journals; this paper)), source (bibliographic source), nodes (number of nodes), edges (number of edges), deg (average node degree), com (percentage of the largest connected component), dis (average geodesic distance), dia (largest geodesic distance), tra (transitivity coefficient), clu (clustering coefficient), mix (assortative mixing). A dash sign indicates that the data are not available.

network analysis metrics to find differences in the research styles of the areas and how these areas interrelate in terms of author overlap and migration. Menezes, Ziviani, Laender, and Almeida (2009) made a geographical analysis of collaboration patterns using network analysis. They considered publications from 1954 to 2007 for members of 30 research institutions (8 from Brazil, 16 from North America, and 6 from Europe) and focused on the differences in collaboration habits among these geographical areas. The dataset, extracted from DBLP, contains 352,766 papers and 176,537 authors. Newman (2001b) studied the collaboration graph for computer science as well, using the NCSTRL library, a database of preprints published in computer science during 1991–2001 and submitted by 160 participating institutions. Unfortunately, as acknowledged by Newman himself, the coverage of the used dataset (13,169 papers and 11,994 authors) is rather limited and hence the sample is not representative of the set of computing publications. Finally, the temporal evolution of the collaboration graph for the database community is studied by Elmacioglu and Lee (2005). The dataset is extracted from DBLP and contains 38,773 publications written by 32,689 authors from 1968 to 2003 covering 19 journals and 81 conferences closely related to the database community. Table 1 contains a summary of network statistics for different disciplines including the results found in this paper.

Our investigation differs from the previously mentioned studies on computer science for the following reasons:

- we use the most complete dataset and build the largest affiliation and collaboration networks ever investigated for computer science;
- with the support of the affiliation network representation of collaboration, we study bibliometric properties for computer science, such as size of the discipline in terms of papers and scholars, author productivity, and collaboration level in papers, and we observe how these properties evolved in time;
- with the aid of the collaboration network, we study meaningful network properties and their temporal evolution; in particular, the size of biconnected components, the concentration of collaboration using the Lorenz curve and the Gini coefficient, the collaboration network resilience and dependence on star collaborators have never been examined before for computer science;
- we investigate separately the networks emerging from scholar collaborations in conference and journal papers.

Methodology

Data were collected from *The DBLP Computer Science Bibliography* (DBLP, for short) (Ley, 2010). The DBLP literature reference database was developed within the last 15 years by Dr. Michael Ley at Trier University, Germany. DBLP is internationally respected by informatics researchers for the accuracy of its data. As of today, DBLP contains more than 1.6 million entries. The scopes of DBLP are scientific journals and proceedings from the field of computer science in a broad sense, including journals of information science such as Journal of the American Society for Information Science and Technology, Scientometrics, and Journal of Informetrics. Editors and publishers may enter in the bibliography the table of contents of new proceedings and journal volumes as soon as the content is related to the field of computer science.

Each publication record in DBLP has a key that uniquely identifies the publication and a property that represents the publication type, such as journal article, conference article, book, book chapter, and thesis. Moreover, it contains a semi-structured list of bibliographic attributes describing the publication, such as authors, title, and year of publication. This list varies according to the publication type (Ley, 2009).

DBLP is particularly careful with respect to the quality of its data, and is especially sensible to the *name problem*, which includes the cases of a scholar with several names (synonyms) and that of several scholars with the same name (homonyms) (Reuther, Walter, Ley, Weber, & Klink, 2006). DBLP uses full names and avoids initials as much as possible. This reduces, but does not eliminate, the name problem. Furthermore, it uses effective heuristics on the collaboration graph to identify possible cases of synonyms or homonyms. For instance, if two lexicographically similar names are assigned to authors that have a distance of two in the collaboration graph, that is,

DOI: 10.1002/asi

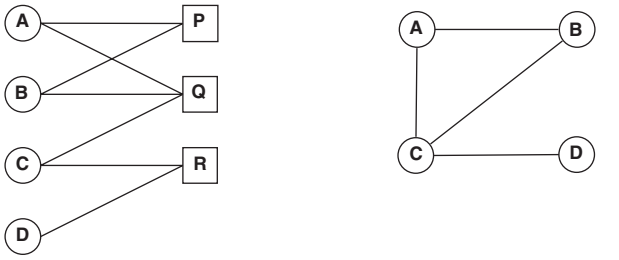

FIG. 1. A simple example of an author–paper affiliation network (left graph). Authors (circle nodes) match the papers (square nodes) that they wrote. We also show the corresponding collaboration network (right graph). In this case, two authors are connected if they wrote at least one paper together.

these authors never directly collaborated in a paper but they have a common collaborator, then these names are identified as possible synonyms and they are further manually investigated. Furthermore, if the list of co-authors of an author splits into two or more clusters of highly interconnected authors, but with no collaborations among authors of different clusters, then we might have a case of homonym, and an additional manual check is performed.

DBLP can be used free of charge. Data can be accessed using a web interface or through automatic HTTP requests, and the entire dataset can be downloaded in XML format to run experiments on top of it.

We downloaded the XML version of DBLP bibliographic dataset in early 2010 (637.9 MB) and filtered all publications from 1936 to 2008 inclusive.[3] On top of this database we built the following networks:

- *Author–paper affiliation network*. This is a bipartite graph with two types of nodes: authors and papers. There is an edge from an author to a paper if the author has written the paper. See an example given in Figure 1. Affiliation networks are the most complete representations for the study of collaboration (Newman, 2010); in particular, on top of such bipartite representations, one can investigate both author-oriented and paper-oriented properties.

  We study the cumulative author–paper affiliation network for computer science, including bibliographic data for all papers published from 1936 to 2008. The resulting network contains 731,333 author nodes, 1,216,526 paper nodes, and 3,112,192 crossing edges. Moreover, for each year Y from 1960 to 2008, we investigate the author–paper affiliation network $AN(Y)$ for year Y, which contains bibliographic data for all papers published in year Y.[4]
- *Collaboration network*. A collaboration network is an undirected graph obtained from the projection of the author–paper affiliation network on the author set of nodes. Nodes of the collaboration network represent authors and there is an edge between two authors if they have collaborated in at least one paper. An example is given in Figure 1. Clearly, the collaboration network is a coarser representation with respect to the affiliation network; for instance, if three authors are mutually linked in the collaboration network, then it is not clear, from the analysis of the collaboration network alone, whether they have collaborated in a single paper or in three different ones. Nevertheless, the collaboration network is highly informative since many collaboration patterns can be captured by analyzing this form of representation. Furthermore, the collaboration network is the main (mostly unique) representation of collaboration that has been studied in the network science literature.

  We analyze the cumulative collaboration network for computer science, including bibliographic data for all papers published from 1936 to 2008. The resulting network has 688,642 nodes (authors) and 2,283,764 edges (collaborations).[5] This is, to our knowledge, the largest computer science collaboration network and the second largest discipline collaboration network ever studied, second only to the Medline collaboration network for biomedicine investigated by Newman (2001b). Furthermore, for each year Y from 1960 to 2008, we studied the collaboration network $CN(Y)$ up to year Y, which contains collaboration data for all papers published until year Y.
- *Conference collaboration network*. In the conference collaboration network, two scholars are linked if they have ever collaborated in at least one conference paper. We analyze the cumulative conference collaboration network for computer science including all conference papers published from 1936 to 2008; the resulting network contains 503,595 nodes and 1,584,108 edges.
- *Journal collaboration network*. In the journal collaboration network, two scholars are connected if they have ever co-authored at least one journal paper. We study the cumulative journal collaboration network for computer science including all journal papers published from 1936 to 2008; the resulting graph contains 356,822 nodes and 987,059 edges.

We saved the collaboration networks in GraphML format (an XML syntax for graphs). We loaded them in the R environment for statistical computing (R Development Core Team, 2008) and analyzed the structure of the networks using the R package *igraph* developed by Gábor Csárdi and Tamás Nepusz. On the other hand, we never materialized the (much larger) affiliation networks. Instead, we used XQuery, the standard XML query language, and BaseX (DBIS Research Group, 2011), a light-speed native XML database, to extract the relevant properties from the XML version of the DBLP database.

Atemporal Analysis

In this part, we study the properties of the cumulative affiliation and collaboration networks considering all papers published in computer science since 1936.

Scientific Productivity and Collaboration Level

In this section, we investigate two typical bibliometric distributions for the set of papers of a discipline: the distribution of the number of papers per author (scientific productivity) and the distribution of the number of authors per paper (collaboration level). These distributions are extracted from the author–paper affiliation network. We recall that this network is a bipartite graph with two node types representing authors and papers; edges match authors with papers they wrote.

DOI: 10.1002/asi

TABLE 2. The scientific productivity of computer scientists.

# of papers	1	2	3	4	5	6	7	8	9	10
% of authors	53.2%	15.8%	7.7%	4.7%	3.2%	2.3%	1.8%	1.4%	1.1%	0.9%

The table shows the relative frequency of authors (second row) who wrote a given number of papers (first row, from 1 to 10 papers).

TABLE 3. The collaboration level of computer science papers.

# of authors	1	2	3	4	5	6	7	8	9	10
% of papers	23.3%	32.8%	23.5%	11.6%	4.6%	1.9%	0.8%	0.4%	0.2%	0.1%
% of conferences	19.2%	32.4%	25.4%	13.3%	5.4%	2.2%	0.9%	0.4%	0.2%	0.1%
% of journals	29.7%	33.4%	20.5%	9.0%	3.4%	1.4%	0.7%	0.4%	0.2%	0.1%

The table shows the relative frequency of papers (second row for all papers, third row for conference papers, and fourth row for journal papers) having a given number of authors (first row, from 1 to 10 authors).

The distribution of the number of papers per author corresponds to the distribution of the node degree for nodes of type author in the author–paper affiliation network. Indeed, the degree (number of adjacent nodes) of a node of type author on the affiliation network is precisely the number of papers published by the author. In substantial agreement with one of the oldest bibliometric laws—Lotka's law of scientific productivity (Lotka, 1926)—the distribution of the number of papers per author is highly skewed, with most of the authors who produced a small number of contributions and few prolific ones who published a large volume of papers. In Table 2 we show the relative frequency of authors who wrote a given number of papers. The table shows only the first 10 number of papers, but the distribution has a long tail ending at 528, the number of papers of the most prolific author. This asymmetry in scientific productivity is not a characteristic of computer science but it has been noticed in many fields; for instance, Moody (2004) found a similar pattern in the productivity of social scientists, with 65.8% of them with 1 paper, 15.1% with 2 papers, 6.5% with 3 papers, 3.7% with 4 papers, 2.2% with 5 papers, and the remaining 6.7% with 6 or more contributions.

As for the distribution of the number of authors per paper, it corresponds to the distribution of the node degree of nodes of type paper in the author–paper affiliation network. We found that the average computer science paper has 2.56 authors. This figure is significantly lower than the average collaboration level in other scientific fields such as physics, chemistry, biology, and medicine, but it is higher than the average collaboration level in social sciences and humanities (see Franceschet & Costantini (2010) for the collaboration level of different disciplines). Hence, computer science stays in a peculiar intermediate position where a little collaboration (2 or at most 3 scholars), but not more than that, is the standard.

Table 3 shows the relative frequency of papers having a given number of authors, distinguishing between conference and journal papers. We observe that conference papers are more collaborative (2.69 authors on average) than journal papers (2.35 authors on average). In particular, notice that 19% of the conference papers are single-author works, while this share is significantly higher, 30%, for the papers in journals.

Connected Components

In this section, we study the property of reachability of scholars on the collaboration network. A *connected component* of an undirected graph is a maximal subset of nodes such that any node in the set is reachable from any other node in the set by traversing a path of intermediate nodes. A connected component of a collaboration graph is hence a maximal set of authors that are mutually reachable through chains of collaborators.

It is reasonable to assume that scientific information flows through paths of the collaboration networks; we expect, indeed, that two authors who collaborated on some papers are willing to exchange scientific information with a higher probability than two scholars who never collaborated. Having a large connected component in the collaboration graph, of the order of the number of scholars, is a desirable property for a discipline that signals its maturity: theories and experimental results can reach, via collaboration chains, the great majority of the scholars working in the field, and thus scholars are scientifically well informed and can incrementally build new theories and discover new results on top of the established knowledge. Furthermore, the connectedness of a discipline is welcome in the view of proofs of theorems and validation of algorithms and experimental results as a social human process and a community project (Millo, Lipton, & Perlis, 1979). Of course, collaboration represents only one way to spread scientific information; the processes of journal publishing and conference attendance make also notable contributions in this direction.

On the other hand, a high level of discipline connectedness might also have negative effects, since it favours the globalization and the standardization of results, and hence the publication of mainstream contributions at the expense

DOI: 10.1002/asi

of more innovative papers that explore research directions outside the established core subjects. Moreover, the independent discovery of the same theories and results by different groups of scientists, which is more likely when the discipline community is disconnected, increases the confidence of the whole community in the validity of these theories and results.

The computer science collaboration network is widely connected. The largest component counts 583,264 scholars, that is 85% of the entire network. It is a *giant component*, since it collects the great majority of nodes. There are two second largest components, the size of which, only 40 nodes, is negligible compared with that of the giant component. The third largest component has 30 nodes, and there are components for each size smaller than 30. In total, we have 34,691 connected components, most of which have small sizes: 18,244 of them have size 2, whereas 8,354 have size 3, and 3,854 have size 4; hence, 88% of the components have size at most 4. The distribution of the size of the connected components that are different from the giant component has a long tail in which most components have small size and a few of them have large size.

Interestingly, the relative size of the giant component of the collaboration network for computer science matches quite well that for physics, and it is a bit higher than that for mathematics. With respect to computer science, the networks for biomedicine and neuroscience are better connected, while the cohesion of social science is lower (see Table 1). This means that research collaboration is more effective in medical disciplines than in social as well as hard sciences.

A *biconnected component* of an undirected graph is a maximal subset of nodes such that for each pair of nodes there are two independent (disjoint) paths connecting them. It follows that the removal of a single node from a biconnected component does not destroy the connectivity of the component. A biconnected component is hence more tightly connected than a connected component. Information flowing on a biconnected component has more chance to reach a target node of the component since there exist two independent paths from any component node to the target node.

The largest biconnected component of the collaboration graph for computer science counts 418,001 nodes, or 61% of the entire network, and it covers a share of 72% of the largest connected component. The second largest biconnected component has only 32 nodes. As a comparison, the physics collaboration has a biconnected component of 59% of the entire network (Newman & Ghoshal, 2008), and for social science the biconnected component occupies a share of 23% of the network space.

The conference collaboration network is also well connected, with 429,193, that is 85% of the authors belonging to the giant component. The distribution of the size of secondary components has a long tail, with the second largest component counting 44 nodes. The journal collaboration network is somewhat less connected: 273,861, that is 77% of the authors lie in the giant component. Again, smaller components distribute with a long tail in terms of size, with a second largest component of 37 elements. Hence, the conference collaboration network is more connected than the journal counterpart, indicating that information has a broader reach when flowing via conference collaboration links.

Geodesic Distances

A high level of connectedness in the collaboration network means that scientific information—theorems, algorithms, and experimental results—can reach almost the whole community via collaboration paths. Connectedness, however, does not tell us the whole story, since it says nothing about how fast the information flows. Information flows faster along shorter paths. In this respect, there exists a substantial difference if the average path connecting two scholars has length, say, six edges, or one hundred links.

We may assume that information preferentially flows along *geodesics*, which are the shortest paths in terms of the number of edges on a graph.[6] A *geodesic distance* between two nodes is defined as the length (number of edges) of any geodesic (shortest path) connecting the nodes—notice that a geodesic is not necessarily unique. The average geodesic distance is the mean geodesic distance among all pairs of nodes of a graph. If the graph is not connected, then there are node pairs that are not reachable. In this case the mean is typically computed on the subset of connected pairs only. The largest geodesic distance in the graph is called the *diameter* of the graph. It tells us how far are two connected nodes in the worst case.

We computed the geodesic distances for all pairs of nodes in the computer science collaboration network and took the average over the subset of connected pairs (the pairs with a defined distance). Since the graph has 688,642 nodes, the number of node pairs is 237,113,557,761, of which 72% are connected by a path.[7] This figure matches the relative size of the giant component, which was found to be about 0.85 (see the section Connected Components). Indeed, if we randomly pick two nodes in the graph, the probability that they fall in the giant component is $0.85^2 \simeq 0.72$. Since the sizes of the other components are negligible compared with that of the giant component, this probability is a close approximation of the probability that two nodes are connected by a path.

Figure 2 shows the share of geodesics having a given length. Notice that geodesics have typically very short lengths compared with the number of nodes: 19% of geodesics have length 5, 33% have length 6, and 26% have length 7. The average geodesic distance is 6.41, and, interestingly, distances normally distribute around this peak. The largest distance, the diameter of the computer science collaboration graph, is also remarkably small: 23 (there are eight different geodesics with this length). Hence, computer scientists are separated on average by six collaboration links, a figure that matches well the legendary six degrees of separation found by the experimental psychologist Stanley Milgram in the 1960s with his popular small-world experiment (Milgram, 1967). This is additional good news for the computing community: not only the collaboration network is mostly connected, but also the average distance is short, and the longest one is not that

DOI: 10.1002/asi

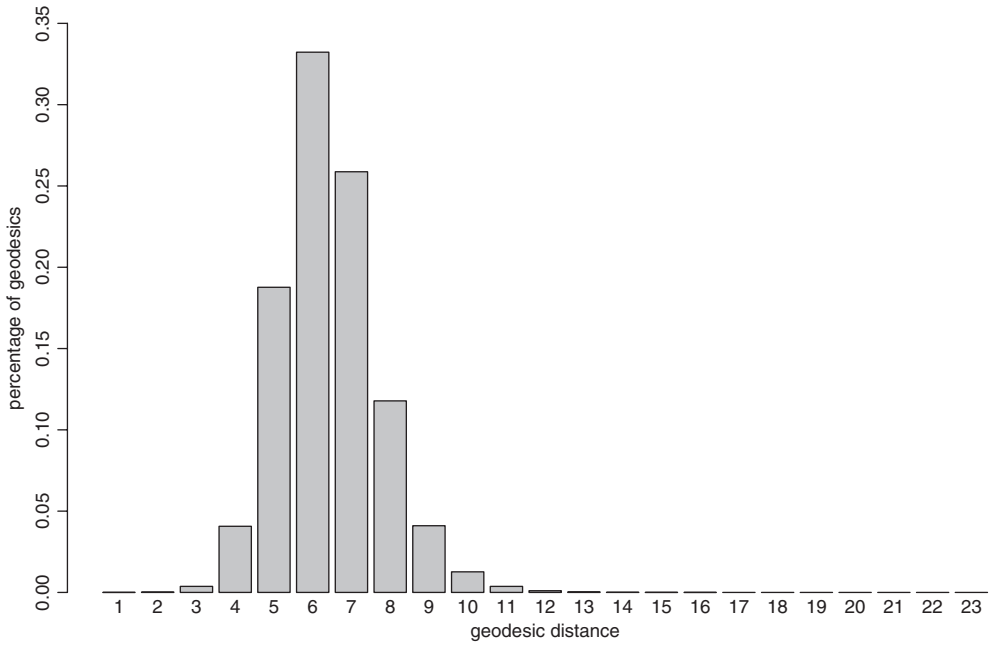

FIG. 2. The distribution of geodesic distances. The distances are shown in the x-axis (from 1 to the network diameter 23) and the height of the corresponding bar is the percentage of geodesics with that distance.

longer. This means that scientific information can spread *quickly* on the great majority of the computing community through its collaboration network.

The fact that distances distribute normally is interesting because it means that the average distance of six links represents a typical value of all distances in the network. Furthermore, since the distribution of distances drops off rapidly around the mean, the time-consuming computation of the exact average distance can be approximated by computing the average distance on a relatively small random sample of node pairs. To demonstrate this, we estimated the average distance on a sample of 10,000 node pairs belonging to the giant component of the network. The outcome is extremely close to the real distance of 6.415: the approximated distance is 6.427, with a 95% confidence interval of [6.401, 6.453].

Is the computer science collaboration network a *small world*? Watts and Strogatz (1998) define a social network as a small world if typical distances grow roughly logarithmically in the number of nodes of the network. More precisely, a network of n nodes and m edges is a small world if the average geodesic distance is roughly $d = \log n / \log k$, where $k = 2m/n$ is the average node degree. Plugging into the formula the corresponding values for our collaboration network we have $k = 6.63$ and $d = 7.10$. Recalling that we measured an average distance of $6.41 < 7.10$, we conclude that the collaboration network of computer science is indeed a small world.

Comparing the observed mean geodesic distance for computer science collaborations with that of other disciplines (see Table 1), we notice that the separation distance for computer science is comparable with that for physics and neuroscience. Moreover, biomedicine has a lower collaboration distance, indicating that collaborations in this field are more densely intertwined. On the other hand, mathematics and, in particular, social science collaboration distances are higher, meaning that collaborations in these disciplines are less frequent and less effective.

The conference collaboration network matches well the whole network in terms of geodesic distances. A share of 73% of node pairs are connected with an average geodesic distance of 6.54 and a largest geodesic distance of 23. In particular, 31% of all shortest paths have length exactly 6. On the other hand, separation distances on the journal collaboration network are larger: the typical distance is now 7.26 and the largest is 25. The largest share of paths, 27%, have length 7. Hence, scholars on the journal collaboration network have on average seven degrees of separation instead of six. Furthermore, only 59% of the node pairs are connected. Summing up, the conference collaboration network is not only more widely, but also more densely connected than the journal counterpart.

Kautz, Selman, and Shah (1997) proposed to use paths on social networks among scientists as *referral chains* to establish contacts with domain experts. In the simplest case, suppose I am searching for a piece of information and I am aware that you are a domain expert who most likely can answer my query. If I do not know you personally, it might be useful to know that we have a common collaborator that can arrange an introduction. In general, as we have seen, there exists a referral chain of intermediate collaborators connecting almost any scholars in computer science. Furthermore, this chain is short in the average case. We might use the people in this chain in order to smoothly get in contact with the target scientist. To be sure, the chance of success depends not only on the path length, but also on the strength of the intermediate path links. If, for instance, I can reach

DOI: 10.1002/asi

you through either collaborator A or collaborator B, and I have published with A 100 papers and with B just one paper, common sense suggests to use A as a broker. In other terms, one might reasonably argue that the intensity of the scientific relationship between two scholars is proportional to the number of papers they have written together. We can implement such an intuition by labelling each edge (x,y) of the collaboration graph with a positive weight $1/k$, where k is the number of papers that x and y have written together.[8] The edge label can be interpreted as a scientific distance among scholars: the more papers two authors have written together, the closer they are scientifically.

The weighted collaboration graph naturally induced the notion of *weighted geodesic*: a shortest path in terms of path weight, defined as the sum of the weights of the path edges. The *weighted geodesic distance* is hence the weight of a weighted geodesic. Notice that a weighted geodesic is not necessarily unique. Moreover, we expect that it differs from its unweighted counterpart. This opens an interesting question connected to the above-mentioned referral chain issue: is preferable a short and weighty path, or a longer and lighter one in order to reach a given domain expert? Notice that a short path has the obvious advantage of having few intermediate scholars to bother, but a light path is desirable since the intermediate links are stronger and more reliable.

In order to investigate whether weighted and unweighted shortest paths are significantly different in the computer science collaboration network, we conducted the following experiment. We extracted a random sample of 10,000 node pairs belonging to the giant component of the computer science collaboration network and computed, for each pair of nodes, the weighted geodesic distance as well as the length of the weighted geodesic. The average weighted geodesic distance is 3.15, and the average length of the weighted geodesics is 11.27. Hence, light paths are much longer, almost twice longer, than the typical geodesic path, which is about six edges long. We furthermore generated 10,000 random node pairs belonging to the giant component and computed, for each of them, the (unweighted) geodesic distance as well as the weight of the geodesic. The average geodesic distance is 6.43, and the average weight of the geodesics is 5.07. Thus, short paths are significantly heavier than the typical weighted geodesic path, which weights about 3. As conjectured, for the computer science collaboration network, weighted shortest paths and unweighted shortest paths are different referral chains; the information seeker is hence faced with the dilemma of following short and weighty (unreliable) chains or long and light (reliable) paths in order to get in contact with the coveted expert.

Node Degree Distribution

A property of the full-scale structure of a network that is typically investigated is the distribution of the network node degrees. Recall that the *degree* of a node is the number of neighbors of the node. In a collaboration network, the degree is the number of unique collaborators of a scholar. For any natural number k, the quantity p_k is the fraction of nodes having degree k. This is also the probability that a randomly chosen node in the network has degree k. The quantities p_k represent the *degree distribution* of the network.

In most real networks, the degree distribution is highly *right skewed*: most of the nodes (*the trivial many*) have low degrees while a small but significant fraction of nodes (*the vital few*) have an extraordinarily high degree. A highly connected node, a node with remarkably high degree, is called *hub*. Since the probability of hubs, although low, is significant, the degree distribution, when plotted, shows a *long tail*, which is much fatter than the tail of a Gaussian or exponential model.

This asymmetric shape of the degree distribution has important consequences for the processes taking place on networks. The highly connected nodes, the hubs of the networks, are generally responsible for keeping the network connected. In other words, the network falls apart if the hubs are removed from the network. On the other hand, since hubs are rare, a randomly chosen node is most likely not a hub, and hence the removal of random nodes from the network has a negligible effect on the network cohesion. Substantially, networks with long tail degree distributions are resilient to random removal of nodes (failure) but vulnerable to removal of the the hub nodes (attack). In the section Network Resilience, we will investigate the resilience of the collaboration network under removal of nodes.

Hubs are also important for the spread of information or of any other quantity flowing on the network. In fact, hubs play a dual role in information diffusion over the network: on the one hand, since they are highly connected, they quickly harvest information, on the other hand, and for the same reason, they effectively spread it. In a network with hub nodes, the probability that each node spreads the information to its neighbors need not be large for the information to reach the whole community.

The degree distribution for the collaboration network in computer science is depicted in Figure 3. The distribution has in fact a long tail: roughly half of the scholars have one, two, or three unique collaborators. The other half of the scholars distribute over a slow decreasing long tail. There are, for instance, 350 scholars with 50 collaborators and 40 scholars with 100 collaborators. The tail is in fact longer than that shown in the figure, with 28 authors with more than 300 collaborators and the most collaborative computer scientist with 595 unique co-authors. The degree distributions for conference and journal articles show a similar pattern (see insets of Figure 3), but the tail for the journal degree distribution is shorter (maximum degree is 260) than the conference counterpart (maximum degree is 481).

To quantitatively study the asymmetry of the degree distribution, we investigate the skewness and the concentration of the distribution. *Skewness* measures the symmetry of a distribution. A distribution is symmetric if the values are equally distributed around its mean, it is right skewed if it contains many low values and a relatively few high values, and it is left skewed if it comprises many high values and a relatively

DOI: 10.1002/asi

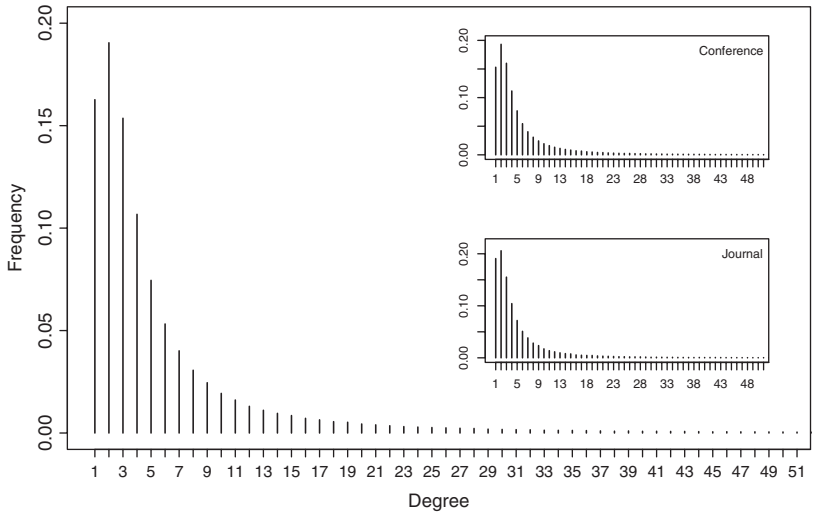

FIG. 3. The degree distribution for computer science collaboration network. Insets: the degree distributions for conference (top) and journal (bottom) collaboration networks.

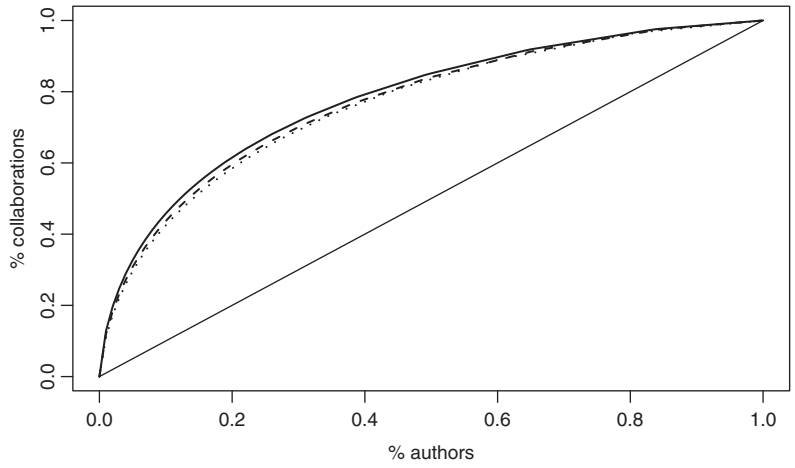

FIG. 4. The Lorenz collaboration concentration curves for the whole (solid curve), conference (dashed curve), and journal (dotted curve) collaboration networks. The share of most collaborative scholars collecting a given percentage of collaboration is plotted.

few low values. As a rule of thumb, when the mean is larger than the median the distribution is right skewed and when the median dominates the mean the distribution is left skewed. The mean degree for the whole collaboration network is 6.63, it is 6.29 for the conference collaboration network, and it is 5.53 in the journal case. The median degree is always 3, and the third quartile is 7 for the whole and conference networks, and 6 for the journal network. A numerical indicator of skewness is the third standardized central moment of the distribution: positive values for the skewness indicator correspond to right skewness, negative values correspond to left skewness, and values close to 0 mean symmetry. The skewness indicator is 8.04 for the whole collaboration network, 7.64 for the conference network, and 6.02 for the journal network. It follows that the analyzed degree distributions are right skewed, and the conference distribution is more asymmetric than the journal counterpart.

Concentration measures how the character (in our context, the collaborations) is equally distributed among the statistical units (the scholars). The two extreme situations are equidistribution, in which each statistical unit receives the same amount of character (each scholar has the same number of collaborators) and maximum concentration, in which the total amount of the character is attributed to a single statistical unit (there exists a super-star collaborator that co-authored with all other scholars, and each other scholar collaborated only with this super-star). We analyze the concentration of collaborations among computer scientists, that is, the concentration of edges attached to nodes in the collaboration graph. Figure 4 depicts the Lorenz concentration curves representing the concentration of collaboration in the whole, conference, and journal networks. Each concentration curve is obtained by sorting scholars in decreasing order with respect to the number of collaborators. Then, the share of most collaborative scholars (or network nodes) collecting a given percentage of collaboration (or network edges) is plotted. It is clear that the concentration of collaboration is far from the equidistribution situation, which is illustrated by the straight line

DOI: 10.1002/asi

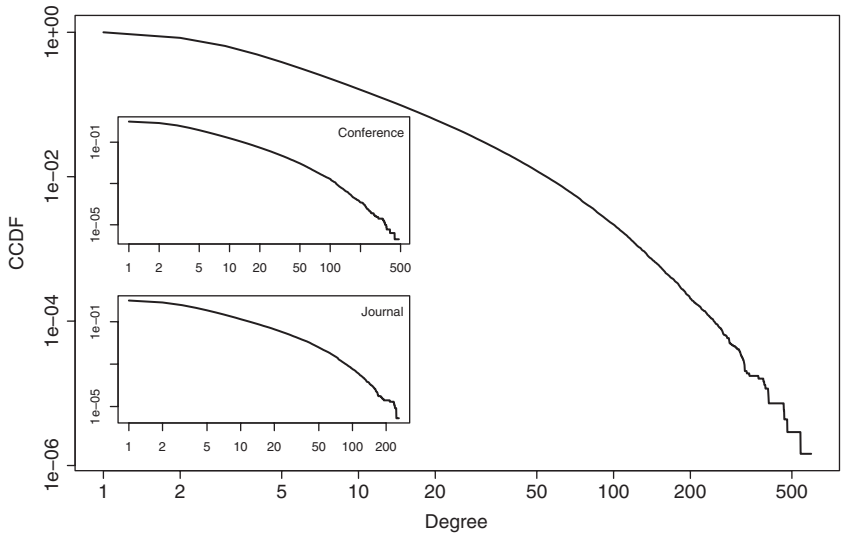

FIG. 5. The complementary cumulative distribution function for the degree distribution of the whole collaboration network. Insets: the same for the conference collaboration network (top) and for the journal collaboration network (bottom).

with slope 1. For instance, the most collaborative 1% of the scholars harvest 13% of the collaborations, 5% of them collect one-third (33%) of the collaborations, and 10% of them attract almost half (46%) of the collaborations. A numerical indicator of concentration is the *Gini coefficient*, which is the ratio between the area contained between the concentration curve and the equidistribution line and the area representing maximum concentration. The index ranges between 0 and 1 with 0 representing equidistribution and 1 representing maximum concentration. The Gini coefficient is 0.56 for the whole network, 0.54 for the conference network, and 0.53 for the journal network. Notice that journal collaborations are slightly less concentrated than conference collaborations, but the three concentration curves are very close.

To be sure, the most popular long tail probability distribution is the *power law*. For a degree distribution, it states that the probability p_k of having a node with k neighbors is $Ck^{-\alpha}$, where C is a normalization constant and α is an exponent parameter. A convenient method to visualize and detect a power-law behavior is to plot the complementary cumulative distribution function (CCDF) on log-log scales (both axes are on logarithmic scales). If a distribution follows a power law, then so does the CCDF of the distribution, but with exponent one less than the original exponent (see Newman, 2010, p. 252). When plotted on log–log scales, a power law appears as a straight line. Figure 5 plots the CCDFs for the networks at hand; all show a clear upward curvature, a sign that they do not match the power-law model on the entire domain.

In practice, few empirical phenomena obey power laws on the entire domain. More often the power law applies only for values greater than or equal to some minimum location. In such case, we say that the *tail* of the distribution follows a power law. Clauset, Shalizi, and Newman (2009) developed a principled statistical framework for discerning and quantifying power-law behavior and analyzed 24 real-world datasets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution in the previous studies. Only 17 of them passed the test with a p-value of at least 0.1, and all of them show the best adherence to the model when a (limited) suffix of the distribution is considered. We applied the techniques developed by Clauset et al. (2009) to detect a power-law behavior in the degree distribution of the computer science collaboration network. As expected, the degree distributions do not follow a power law on the entire regime. Nevertheless, the degree distribution for the whole collaboration network has a power-law tail starting from degree 111 ($\alpha = 4.4$, p-value $= 0.11$). The tail contains, however, only 1098 highly collaborative scholars, which correspond to 0.16% of all authors. The degree distribution for the conference network does not follow a power law in any significant portion of its tail. Finally, the degree distribution for the journal network matches a power law from degree 105 ($\alpha = 5.79$, p-value $= 0.67$); the tail is 178 scholars long, or 0.05% of the entire distribution. There is, however, a longer (551 scholars, 0.15%) but statistically less significant power law distributed tail starting from degree 77 ($\alpha = 4.91$, p-value $= 0.09$). All in all, two of the analyzed networks (the whole and the journal one) have a power law distributed tail, but the relative size of the tail is in both cases rather limited. We conclude that the process of *preferential attachment*—the attitude of scholars to collaborate preferentially with highly collaborative peers, which is one of the possible causes for the power-law behavior (Barabási & Albert, 1999; de Solla Price, 1976), is not a valid explanation for the generation of the computer science collaboration network.

Network Resilience

Percolation[9] is one of the simplest *processes* taking place on networks. The process progressively removes nodes, as long as the edges connected to these nodes, from the

DOI: 10.1002/asi

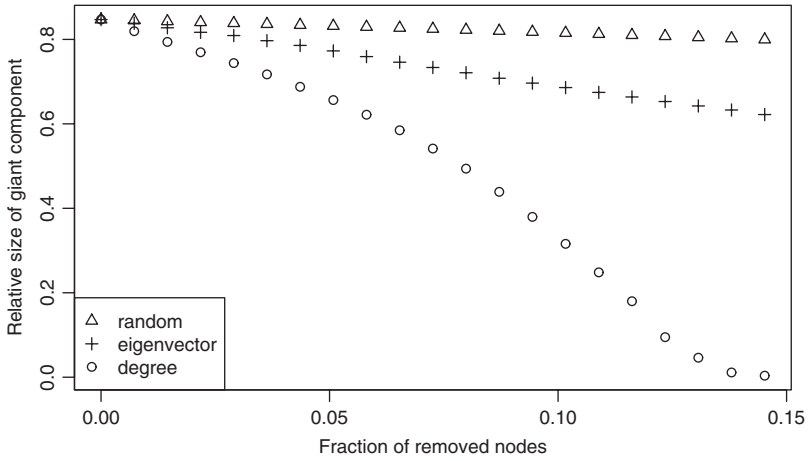

FIG. 6. Effect of percolation on the connectivity of the computer science collaboration network. The relative size of the giant component is shown as an increasing fraction of nodes that are removed according to three different percolation strategies.

network, and studies how the connectivity of the network changes. In particular, one wants to find the fraction of nodes to remove from the network in order to disintegrate its giant component into small disconnected pieces. If such a fraction is relatively large, then the network is said to be resilient (or robust) to the process of percolation. Our study of percolation on collaboration networks is shaped by the following questions:

- What is the best removal strategy to destroy the overall connectivity of the collaboration network?
- What is the tipping point in the percolation process after which the network consists of only small disconnected clusters?
- Are the most collaborative scholars responsible for keeping the network connected?

To address the above questions, we performed the following computer simulation. We implemented a computer procedure that progressively removes nodes from the collaboration network. At each step the procedure removes an increasing number of nodes from the original network and, after each removal, it computes the relative size of the giant (largest) component of the resulting sub-network. More precisely, the procedure initially generates, according to a given strategy that will be discussed below, a number of $5000 \times 20 = 100{,}000$ nodes (about 15% of the total number of nodes) to remove from the original graph. At each step i from 1 to 20, the procedure removes the first $n_i = 5000 \times i$ nodes of the generated ones from the original network as well as the incident edges and computes the share of the giant component of the resulting network. The procedure can choose the removal nodes according to three strategies: (i) *random-driven percolation*, in which randomly chosen nodes are removed, (ii) *degree-driven percolation*, in which the nodes are removed in decreasing order of node degrees, and (iii) *eigenvector-driven percolation*, in which the nodes are removed in decreasing order of node eigenvector centrality scores.[10] To avoid possible biases in random-driven percolation due to the non-deterministic output of the random generator, we repeated the experiment a large number of times and took the average of the results.[11]

Figure 6 clearly shows that the most effective strategy (among the surveyed ones) to destroy the connectivity of the collaboration network is to percolate the highly collaborative scholars: after the removal of about 12% of the most collaborative scholars, the giant component of the collaboration network, which initially contains 85% of the nodes, falls below 10%, and it soon vanishes when the removal fraction is a little higher (15%). The effect of removing scholars with high eigenvector centrality is much less substantial: the largest component is still considerable, about 62% of the network, when a fraction of 15% of the most central nodes are removed. Finally, the random removal of nodes has a negligible effect on the giant component of the network: after the percolation of a share of 15% of randomly chosen nodes the network is still highly connected, with 80% of the nodes belonging to the giant component. For a comparison, Newman (2010) found that to destroy the connectivity in the physics collaboration network it is necessary to remove between 20 and 30% of highly collaborative physicists.

The shape of the degree-driven percolation curve shows clear curvatures. In particular, we can distinguish three main phases in the percolation process. An initial phase, up to a removal fraction of 7%, in which the decrease in connectivity is limited (3–4% points at each step). In this phase, although we severely attack the network by removing its most important hubs, the effect is somewhat reduced since the collaboration frame is still densely intertwined. That is, each node pair in the giant component is connected by more independent paths, and the removal of some of them do not prevent reachability.[12] In a following phase, which extends up to a removal share of 12%, the reduction of connectivity is more notable (5–9% points). In this phase, the size of the largest component is below 50% and its collaboration frame is weaker and more vulnerable. Hence, the effect of the attack is more devastating. In the last segment, up to a removal

DOI: 10.1002/asi

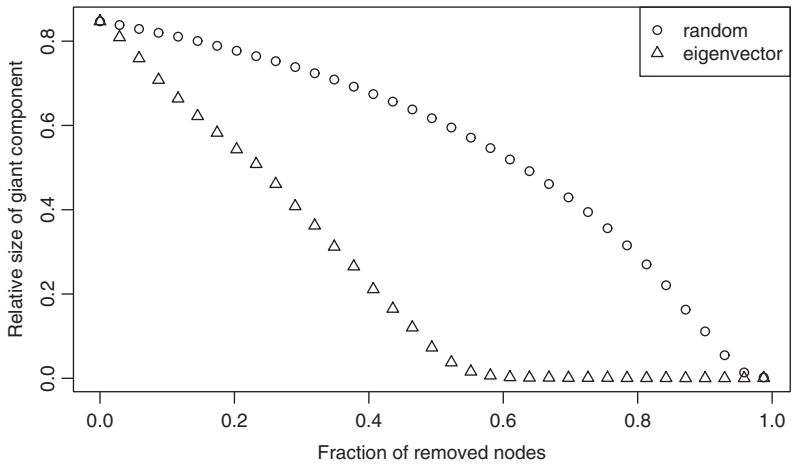

FIG. 7. Effect of random-driven and eigenfactor-driven percolation on the connectivity of the computer science collaboration network.

percentage of 15%, the relative size of the giant component is below 10% and it goes rapidly to 0; in this phase the network consists of small disconnected clusters, none of which strongly dominates the others. Hence, for degree-driven percolation the tipping point after which only small disconnected clusters exist is around 15%.

The overall effect of random-driven and eigenvector-driven percolations of network connectivity is shown in Figure 7. The shape of the random-driven percolation curve shows a clear upward curvature, meaning that the effect of random removal has a higher impact on connectivity when a significant fraction of the nodes have been already removed, as in the degree-driven case. Most of the nodes must be removed before connectivity is lost, and the tipping point where the network falls apart into small pieces is around 90%, far beyond the percolation threshold for the degree-driven case. The eigenvector-driven percolation curve is instead linear up to its tipping point around 50%, which lies between the percolation thresholds of the degree-driven and random-driven processes.

A similar analysis for the second largest component shows that its size during the percolation process is never significant (always below 1%). This means that when a giant component exists, relatively small pieces belonging to the periphery of the giant component separate during the percolation process, and it never happens that the giant component splits into two fragments of similar size. Only when the network is divided into small disconnected clusters, the size of the giant and sub-giant components are comparable.

We are left to the last question posed at the beginning of the present section: are the star collaborators, the hubs of the collaboration network, responsible for the connectivity of the overall network? Our answer, maybe surprisingly, in negative. The hubs are nodes with an *extraordinary* number of neighbors. Let us define hubs as those nodes with a degree higher than or equal to the 99th percentile in the degree distribution, that is the 1% of nodes with highest degree. These are the 7036 scholars of the network with at least 54 collaborators. Recall that the average scholar has between 6 and 7 collaborators, hence hub scholars have a number of collaborators eight times higher than the average scholar. The removal of these super-star collaborators has, in fact, a negligible effect on the connectivity of the network: the size of the giant component decreases form 85 to 81%. On the other hand, as we have seen, to dismantle the network we need to remove at least 15% of top collaborative scholars, that is all authors with a degree higher than 11. In our assessment, scholars with 11 unique collaborators are not collaboration hubs (I have 15 unique collaborators, and I really do not feel I am a collaboration star in computer science). This conclusion matches the findings of Moody (2004) for the social science collaboration network. Hence, the computer science collaboration network is not glued together by star collaborators and, while such actors are likely very influential within their local communities, they do not control information diffusion on the *whole* computer science collaboration network.

As for conference and journal collaboration networks, the results are similar but the tipping points are lower. In particular, the tipping points for the journal network are below those for the conference network. For instance, the journal network falls into pieces after 7% of the most collaborative scholars are inhibited, whereas the conference network crumbles when 9% of the most collaborative scholars are removed. This means that the journal network is more fragile and its connectivity is more dependent on star collaborators.

Clustering and Mixing

In common parlance, clustering, known also as transitivity, measures the average probability that "the neighbor of my neighbor is also my neighbor". The definition of clustering can be formalized as follows. Let a *connected triple* be a triple of nodes x, y, and z such that x is linked to both y and z. That is, y and z have a common neighbor, x. A *triangle* is a triple of nodes such that all pairs of nodes are connected by an edge. Notice that a triangle codes for three connected triples, each one centred at one vertex of the

DOI: 10.1002/asi

triangle. The *global clustering coefficient* of a network can be defined as

$$T = 3\frac{N_\Delta}{N_\wedge} \quad (1)$$

where N_Δ is the number of triangles and $N_\wedge$ is the number of connected triples in the network. The factor 3 constrains the coefficient to lie in the range between 0 and 1. Thus, $T = 1$ implies perfect transitivity, that is, a network whose components are *complete* graphs,[13] while $T = 0$ implies no triangles, which happens, for instance, in a network whose components are *tree* graphs.[14]

Local clustering refers to a single node. For a vertex i, its *local clustering coefficient* C_i is the fraction of neighbors of i that are connected, that is the number of pairs of neighbors of i that are connected divided by the number of pairs of neighbors of i. By taking the average local clustering over all nodes of a network, we have an alternative (but different) definition of global clustering coefficient (Watts & Strogatz, 1998):

$$C = \frac{1}{n}\sum_i C_i \quad (2)$$

It is worth noticing that these two definitions of clustering—T and C expressed, respectively, by Equations. (1) and (2)—are not equivalent and can give substantially different results for a given network. We prefer definition T because it has an intuitive interpretation—the average probability that two neighbors of a node are themselves neighbors. To distinguish between them, in the following we will refer to T as transitivity coefficient and to C as clustering coefficient.

The transitivity coefficient of the computer science collaboration network is 0.24. This means that, on average, the chance that two scholars who share a common collaborator wrote a paper together is almost one-fourth. This is rather a high probability, indeed. As a comparison, the transitivity coefficient for a random network of the same size is 9.6×10^{-6}, and that for a network with the same degree distribution of our collaboration network but otherwise random is 1.4×10^{-4}; both values are orders of magnitude lower that what we computed on the real collaboration network.[15] This large discrepancy is a clear sign of real social effect at work in the context of academic collaboration: authors with a common collaborator have several good reasons to write a paper together, for instance they are probably working on very close topics or they scientifically know each other through the common collaborator.

The transitivity coefficient for the conference and journal collaboration networks are 0.24 and 0.37, respectively. These values might indicate that journal collaboration establishes a stronger relationship among authors, so that authors having a common journal collaborator are professionally and maybe socially closer than authors sharing a common conference collaborator, and hence more inclined to collaborate themselves. In computer science publishing, a paper in a journal is, generally, harder than writing a paper for a conference (Franceschet, 2010), and this might explain the difference in strength between conference and journal co-authorship. Comparing with other disciplines, the magnitude of transitivity in computer science is comparable with that in sociology, larger than that in mathematics and biomedicine, and lower than that in physics (Table 1).

The clustering coefficient for the computer science collaboration network is 0.75 (0.75 for conference collaboration and 0.77 for journal collaboration), much larger than the transitivity coefficient. This confirms, once again, that the two clustering measures are very different. Further, we computed the local clustering coefficient for each node and noticed that almost half of the nodes (48%) have local clustering coefficient equal to 1, and this explains the large value of the clustering coefficient. Recall that a node i has local clustering $C_i = 1$ if its immediate neighbors form a complete graph (a clique). Notice that the neighborhood clique can be extended by adding node i itself. We noticed that most of these special cliques have small size. The most frequent pattern (37%) is a clique of size 3, that is a triple i, j, k of scholars such that j and k are the only collaborators of i and they are themselves collaborators (not necessarily on a paper with i). Moreover, 26% of these special cliques have size 4, 15% of them have size 5, 8% of them have size 6, and the distribution decreases slowly with a long tail ending with size 114. This large maximum value, however, is easily explained by the existence of a paper with 114 co-authors.[16]

Social networks differ from most other types of networks, including technological and biological networks, at least in two aspects (Newman & Park, 2003). First, the transitivity coefficient is higher for social networks. Second, they show positive correlation between the degrees of adjacent nodes, while other networks have negative correlation. *Assortative mixing* is the tendency of nodes to connect to other nodes that are similar to them in some way. In particular, assortative mixing by degree is the tendency of nodes to connect to other nodes with a similar degree. In our context, we have assortative mixing by degree if scholars collaborate preferentially with other scholars with similar number of collaborators. We have disassortative mixing by degree if collaborative scholars co-author with hermits and vice versa. We have no mixing at all if none of these patterns is clearly visible.

A quantitative measure of the magnitude of mixing by degree can be computed using the *Pearson correlation coefficient* applied to the degree sequences of nodes connected by an edge. The coefficient ranges between -1 and 1, where negative values indicate disassortative mixing, positive values indicate assortative mixing, and values close to 0 indicate no mixing. The coefficient is 0.17 for the whole network, 0.16 for the conference one, and 0.30 for the journal one. The values are statistically significant. Hence, collaboration networks in computer science confirm assortative mixing by degree, as other social networks: collaborative computer scientists tend to collaborate with other collaborative computer scientists, and solitary authors match preferentially with other solitary authors. Notice the higher value for collaboration in journal papers, indicating that this collaboration pattern is stronger in this case. These findings are useful to picture the structure of the collaboration network. A network that

DOI: 10.1002/asi

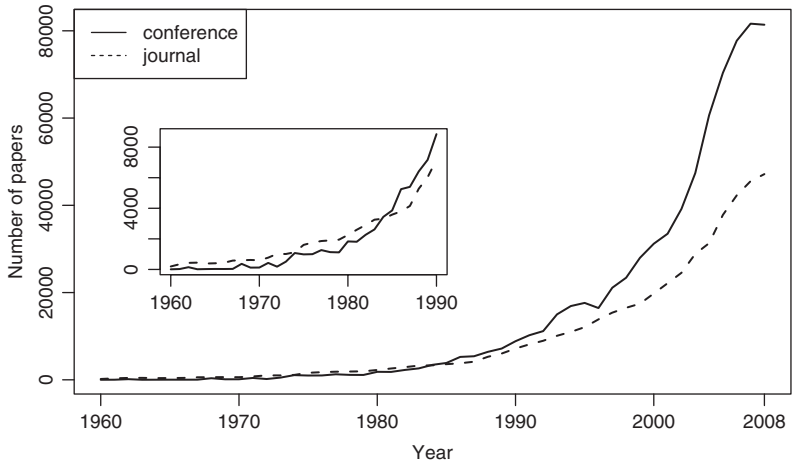

FIG. 8. The size of the computer science discipline in terms of number of conference papers (solid line) and number of journal papers (dashed line) that are published each year. The inset refers to the shorter period from 1960 to 1990.

is assortative by degree has a *core–periphery structure*: a dense core of high-degree nodes is surrounded by peripheral low-degree ones.

In fact, the correlation is even stronger. Pearson correlation coefficient is an appropriate measure of correlation when the data samples roughly follow a normal distribution. In this case, the mean of the samples, which is used in the computation of the coefficient, represents a characteristic scale for the network. However, we have seen in the section Node Degree Distribution that our collaboration networks are *scale-free*: their degree distribution is highly right skewed and the distribution mean does not represent a typical value of the number of collaborators of a scholar. To correct for the bias introduced in the Pearson correlation coefficient by the use of asymmetric degree distributions, we can either make a logarithmic transformation of the degree sequences before using the Pearson coefficient formula, or use a non-parametric correlation method, like the Spearman one. Both methods give the same results: the correlations increase to 0.25, 0.21, and 0.36 in the whole, conference, and journal collaboration network, respectively.

Temporal Analysis

In this part we study how bibliometric and network properties evolved in time since 1960.

Affiliation Network Properties

In this section, we investigate the temporal evolution of the properties of the author–paper affiliation network of computer science. We recall that the computer science author–paper affiliation network is a bipartite graph with two node types representing authors and papers; edges match authors with papers they wrote. For each year from 1960 to 2008, we study the yearly affiliation network build from all papers published in that year. In particular, with the aid of the affiliation network, we investigate the temporal evolution of the following bibliometric properties for computer science: number of published papers, number of active authors, average number of papers per author, and average number of authors per paper.

Figure 8 depicts the size of the computer science discipline in terms of number of papers that are published for each year since 1960. With respect to the author–paper affiliation network, this is the number of paper type nodes in the network. The computer science field is steadily expanding in terms of number of published papers. However, the proportion of conference and journal papers changed over years. Until 1983, the volume of journal papers dominates that of conference papers. However, since 1984 conferences are the most popular publication venue in computer science, and in the recent years computer science published almost two conference papers for each journal paper (Franceschet, 2010).

What are the reasons for the growth of the computer science field? We investigated, for each year, the number of active authors, who published at least one paper in the year, and the author productivity, defined as the average number of papers published by an active author in the year. With respect to the author–paper affiliation network, the number of active authors is the number of nodes of type author in the network, and the author productivity is the average degree of nodes of type author. Indeed, the degree of a node of type author on the author–paper bipartite network is precisely the number of papers published by the author.

Figure 9 shows the temporal evolution of both variables (number of active authors and their productivity). Both variables are growing over time, although the author productivity shows some oscillations during the 1960s. Hence, the expansion over time of computer science in terms of the number of papers is justified by an increase in the number of active authors and by a rise of the author productivity. Substantially, the discipline grows since there are more active scholars and the typical active scholar writes more papers.

Finally, we observed, for each year, the average collaboration level in papers, defined as the average number of authors

DOI: 10.1002/asi

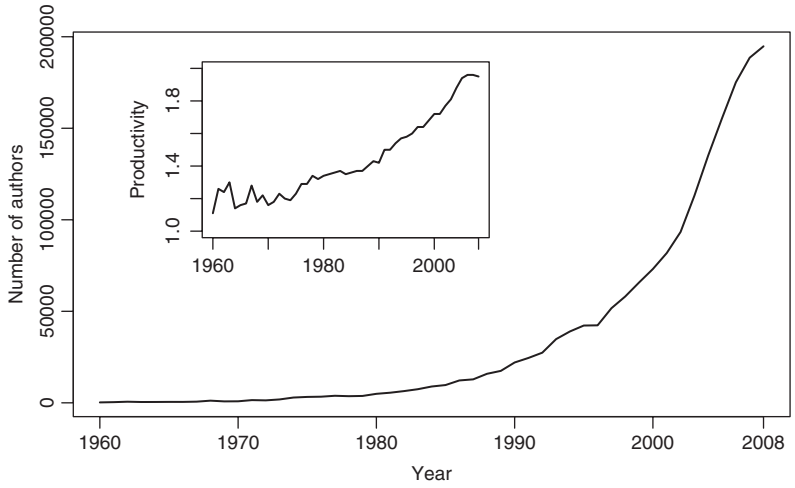

FIG. 9. The size of the computer science discipline in terms of number of active authors for each year. The inset shows the author productivity in terms of the average number of papers published by an author.

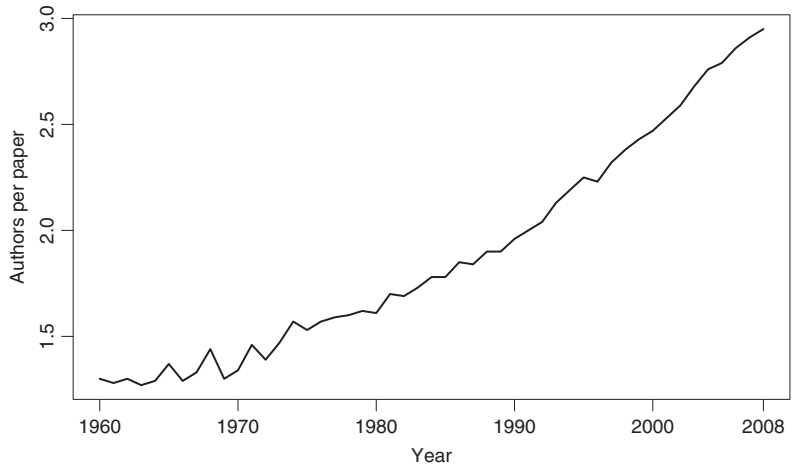

FIG. 10. The average collaboration level in papers for each year.

per paper published in the year. With respect to the author–paper affiliation network, the average collaboration level is the average degree of nodes of type paper. Indeed, the degree of a paper node is the number of authors who signed the paper. Figure 10 shows that collaboration in papers increased over time. The average paper in 2008 has 2.95 authors, whereas the average one in 1960 had 1.30 authors.[17]

Collaboration Network Properties

In this section, we study the temporal evolution of global properties of the collaboration network of computer science. Recall that the computer science collaboration network is a graph with nodes representing the computer science scholars and edges representing the collaborations between scholars in research papers. For each year from 1960 to 2008, we study the cumulative collaboration network build from all papers published until that year. Specifically, we investigate the temporal evolution of the following properties of the computer science collaboration network: connectivity of the network, average separation distance among scholars, distribution of the number of scholar collaborators, network clustering, and network assortativity by number of collaborators.

Figure 11 plots the temporal evolution of the share of the largest connected component and of the largest biconnected component of the collaboration network. The fraction of the collaboration network that is taken by the largest connected component grows steadily since 1970. This fraction is below half of the network until 1990, and it covers a share of 80% in 2004. The figure for the last year, 2008, is 85%. Notice that the slope of the curve decreases in the last few years, that is the increase in the share of the largest connected component in a given year with respect to the previous year is less noticeable in the recent years. This pattern can be interpreted as a convergence of the collaboration network toward a steady state, where the largest part of the network, although not the entire graph, is connected by paths of collaborations. The line for the largest biconnected component follows a trend similar to the curve for the largest connected component. Not surprisingly, the bicomponent line is below the connected component curve, since any bicomponent is embedded into a connected component. The share of the largest biconnected

DOI: 10.1002/asi

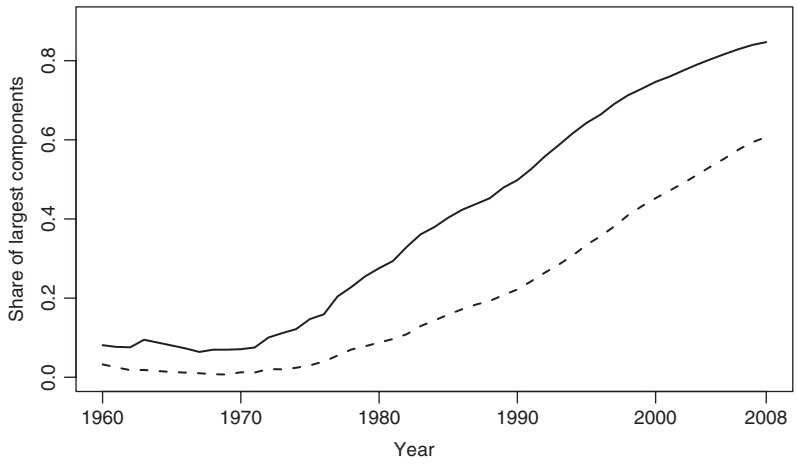

FIG. 11. The temporal evolution of the share of the largest connected component (solid line) and of the largest biconnected component (dashed line) of the collaboration network.

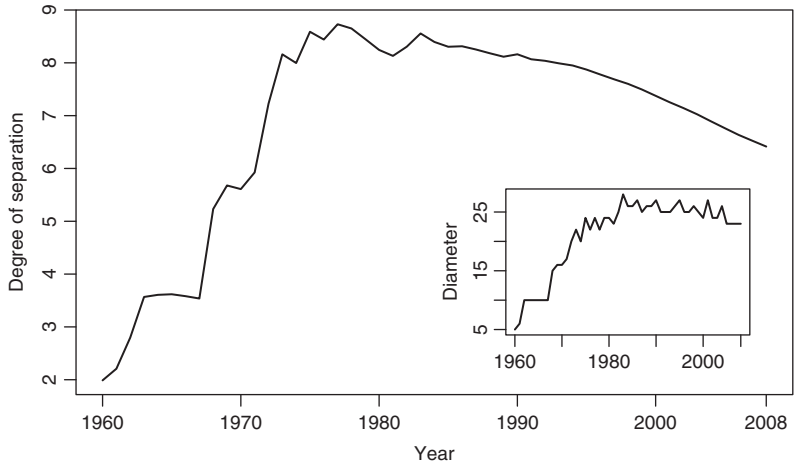

FIG. 12. The temporal evolution of the average degree of separation between scholars (main plot) and of the maximum degree of separation (inset plot) in the collaboration network.

component continuously grows from the 1970s and reaches half of the collaboration network in 2003, and 61% of the whole graph in 2008.

Figure 12 shows the temporal evolution of the average degree of separation between scholars (main plot) and of the maximum degree of separation (inset plot) in the collaboration network. We can neatly distinguish two phases. The first phase goes from 1960 to 1983. It is characterized by the alternation of periods in which the geodesic distance expands followed by shorter periods in which it shrinks, drawing a curve with a series of ridges. We will see next in this section that the collaboration density among computer scientists, measured as the mean number of collaborators per scholar, continuously increases since 1960. In this initial phase from 1960 to 1983, the collaboration graph is split up in many relatively small connected components, one of them, the largest, collect an increasing share of nodes, from 8% in 1960 to 36% in 1983. As the collaboration density increases, it may happen that some of these components glue together to form bigger components, with the effect of raising the average geodesic distance between nodes. These expansion periods are followed by shorter periods in which the collaboration density increases but no significant merging of components occurs, and hence the average geodesic distance between nodes remains stable or more frequently it declines. Since the expansion periods are longer than the contraction ones, in this initial phase the overall average geodesic distance increases, with a maximum of 8.73 reached in year 1977. Also, in this phase, the largest geodesic distance (the diameter) increases, reaching its peak of 28 in year 1983.

The first phase from 1960 to 1983 is followed by a second phase, from 1983 to 2008, in which the average geodesic distance almost continuously decreases, drawing a valley that gently slopes, as opposed to the sharp ridges of the previous phase. Moreover, the diameter oscillates in a short range (from 23 to 28). In this second phase, there exists a large connected component that contains an important share of the network; this share contains the majority of the nodes starting from 1991. The pairs of nodes in this giant component dominate the computation of the average geodesic

DOI: 10.1002/asi

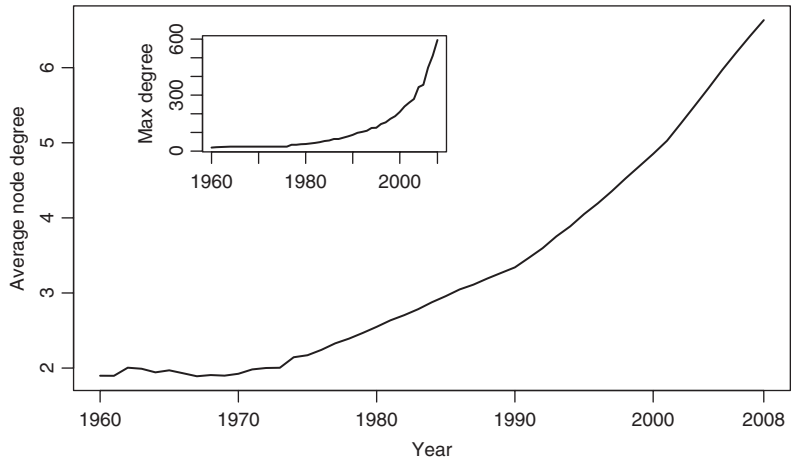

FIG. 13. The temporal evolution of the average number of collaborators of a scholar (main plot) and of the maximum number of collaborators of a scholar (inset plot).

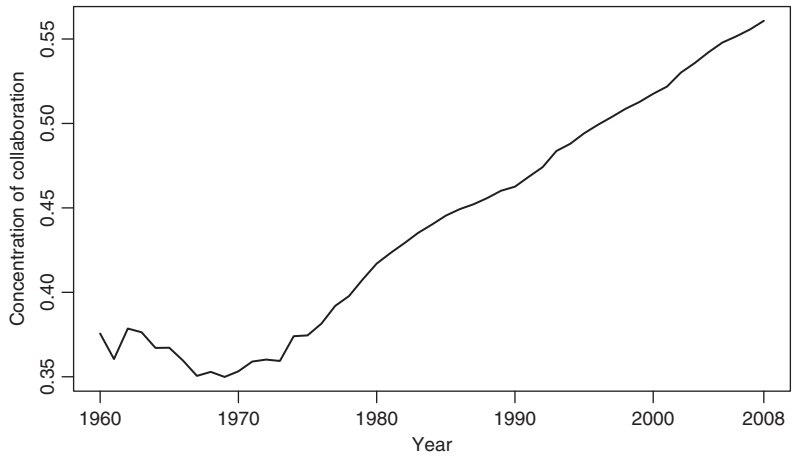

FIG. 14. The temporal evolution of the collaboration concentration measured with the Gini concentration coefficient.

distance. Since the collaboration density inside this component becomes higher and higher, the collaboration distances get lower and lower. We conjecture that the computer science collaboration network is slowly converging toward a steady state, where the majority of pairs of scholars are connected by short paths (of six or less edges).

In the following we investigate how the density and the distribution of the collaboration among computer scientists evolved over time. A measure of the density of collaboration is the average number of collaborators per scholar. This is the average degree of a node in the collaboration graph. Figure 13 depicts the temporal evolution of the average number of collaborators of a scholar (main plot) and of the maximum number of collaborators of a scholar (inset plot). The mean number of collaborators is stable in the 1960s, with an average of 1.9 co-authors per scholar. From 1970 it steadily increases: the average computer scientist has 2.2 collaborators in the 1970s, 2.9 collaborators in the 1980s, 4 collaborators in the 1990s, and 5.7 collaborators in the 2000s. The mean number of collaborators in the last year 2008 is 6.6, more than 3 times the number of collaborators a computer scientists had in the 1970s. The maximum number of collaborators of a scholar, although a less significant figure, is also increasing with time, reaching the impressive maximum amount of 595 collaborators for a single scholar in 2008.[18]

We have just noticed that the number of collaborations increased over time. However, did the distribution of collaborations among scientists change over time? Figure 14 shows how the Gini concentration coefficient varied with time in the computer science collaboration network. From 1970 the concentration of collaboration among our peers moved away from the equidistribution toward a more concentrated scenario.

Finally, we focus on the topological structure of the computer science collaboration network. To this end, we use two apparently unrelated measures: the transitivity coefficient and the assortativity coefficient. Figure 15 illustrates the transitivity coefficient (main plot) and the assortativity coefficient (inset plot) of the collaboration network. Substantially, the transitivity coefficient is decreasing in time, ranging from a value of 0.76 in 1960 to a value of 0.24 in 2008. In 1960

DOI: 10.1002/asi

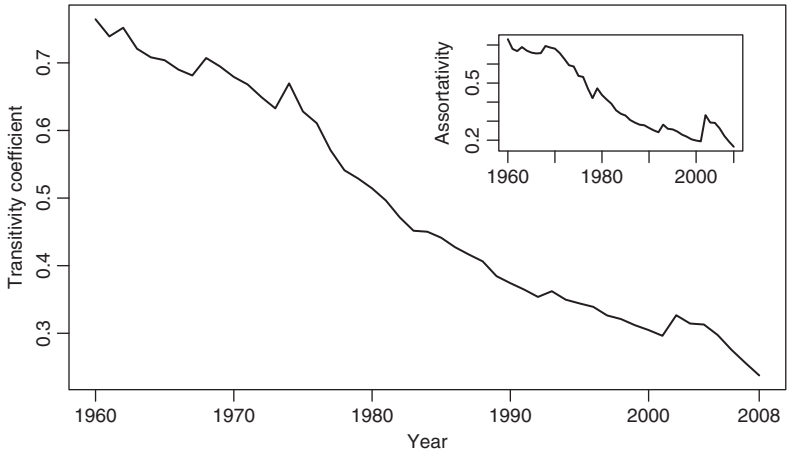

FIG. 15. The temporal evolution of the transitivity coefficient (main plot) and of the assortativity coefficient (inset plot) of the collaboration network.

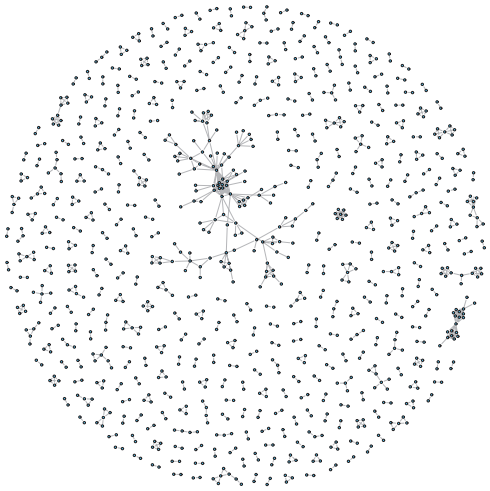

FIG. 16. The computer science collaboration networks in 1964.

it holds that 3 times out of 4 two scholars sharing a common collaborator are themselves co-authors. After 50 years of computer science, the same phenomenon occurs only one time out of 4.

As to assortativity, in the early years, the collaboration network is strongly assortative by degree, with an assortativity coefficient of 0.73. The collaboration network remains assortative by degree in the following years, but the strength of assortativity decreases, and reaches its minimum in 2008 with an assortativity coefficient of 0.17.

Interestingly, the assortativity coefficient and the transitivity coefficient follow very similar trends and seem to be highly correlated. Indeed, the Pearson correlation coefficient between the two variables is 0.98, which means almost perfect positive correlation.[19] The synchronous behavior of the transitivity and of the assortativity coefficients reveals an interesting change in the topology of the collaboration network over time. In the early times of computer science, the collaboration network had a clear *core–periphery structure*, with a core of interconnected high-degree nodes surrounded by a less dense periphery of nodes with lower degree. In particular, the periphery of the network is composed of small clusters of nodes, which are internally highly connected, but disconnected from the other small groups and from the core of the network. See Figure 16, depicting the collaboration network as it was in 1964, for an example. This kind of network has both a high clustering coefficient and a high assortativity coefficient. Indeed, the core of the network is highly clustered, and so are also the peripheral small clusters, and hence the clustering coefficient is high. Moreover, the central high-degree nodes of the core are interconnected, and the peripheral low-degree nodes match with other similar nodes, making the network highly assortative by degree.

The computer science collaboration network progressively lost this peculiar core–periphery structure over time, or at least this structure is not so clear in the years of modern computer science. The modern structure of the collaboration network is dominated by the largest connected component of the network, which covers the majority of the nodes. The disconnected periphery is still composed of small clusters of nodes, but, unlike in the early years, this part of the graph contains an insignificant share of nodes. The nodes in the largest component are not as highly linked as were the nodes in the core component of early networks. Moreover, highly collaborative scholars in the modern collaboration network are more willing to collaborate with those with few collaborators (as in the typical relationship supervisor–student). Figure 17 shows the largest component for the 1980 collaboration network. Notice that, to a certain degree, the core–periphery structure is still noticeable, although it is less clear than in the core component of Figure 16.

Conclusions

We have analyzed collaboration in computer science using a network science approach. Substantially, we have found that the scientific productivity of computer scientists is highly asymmetric, in agreement with Lotka's law of scientific productivity. The collaboration level in computer science papers is rather moderate with respect to other scientific fields, and

DOI: 10.1002/asi

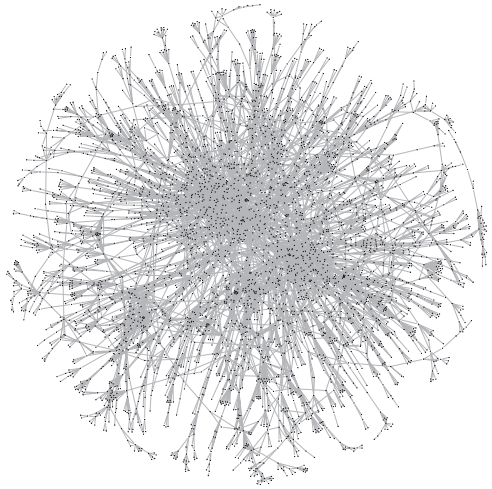

FIG. 17. The computer science collaboration networks in 1980 (only the largest component is shown).

a little collaboration of two or at most three authors is typical in computer science. However, conference papers are more collaborative than journal ones. This suggests that collaboration is more important when there are stringent deadlines for the production of a paper, like those imposed by computer science conferences.

The computer science collaboration network is a widely connected small world, hence scientific information flows along collaboration links very quickly and it potentially reaches almost all scholars in the discipline. This signals the scientific maturity of the relatively young field of computer science. The distribution of collaboration among computer science scholars is highly skewed and concentrated, with few star collaborators responsible for a relatively high share of collaborations. The collaboration network is, however, resilient to the removal of these star collaborators, meaning that the connectivity of the network does not crucially depend on them. This is good news for the computer science community, since it means that this restricted circle of influential scholars with many contacts do not control the diffusion of information about the whole discipline, although they are probably very influential within their local communities.

We also found that the conference collaboration network is more widely and densely connected than the journal counterpart, and journal collaboration establishes a stronger social relationship among authors, also because, as observed above, the typical journal paper has fewer authors than the average conference contribution. The journal network is more dependent on star collaborators, and these highly collaborative authors prefer to collaborate with other star collaborators, leading to a core–periphery structure in the journal graph. These patterns might indicate that conferences are better to widely and quickly communicate scientific results, while journals are optimal to establish stronger and longer scientific relationships with other scholars.

Furthermore, we have analyzed how collaboration in computer science evolved in the last half-century. Computer science has expanded since 1960 both in terms of the number of published papers and number of active authors. Moreover, computer scientists became more productive and collaborative. Productivity and collaboration are mutually reinforcing, since more collaborative authors generally write more papers and more productive scholars typically attract more collaborations. Collaboration changed not only in intensity, but also in concentration: the gap between the richest and the poorest scholars, in terms of collaborations, is increasing.

The temporal evolution of the collaboration network is approaching a steady state, where most of the scholars belong to a giant connected component and are separated by few collaboration links. Finally, the collaboration network moved from a clear core–periphery topology to a more balanced one, with lower transitivity and lower homophily by number of collaborators. In particular, the lower transitivity with respect to the early times of computer science might indicate a higher propensity for collaborations among different research institutes, among different countries, and among different sub-fields of computer science. Triads involving such heterogeneous collaborations, indeed, are less likely to be closed in a collaboration triangle. Recent bibliometric studies have shown that papers involving collaborations over different institutes, countries, or research fields generally enjoy a higher visibility and attract more citations than papers with homogeneous authors, making these types of collaborations more appealing (Franceschet & Costantini, 2010).

What is the contribution of our paper to the bibliometrics reader? We think that the main contribution is the network science approach that we fully exploited in this work. Networks abound in bibliometrics; important examples are collaboration networks, citation networks, co-citation, and co-reference networks. Nevertheless, we are not aware of any study of the structure and evolution of such networks in the context of bibliometrics using the network science approach as we adopted in this paper, the approach that is fully described by Newman (2010). The network science approach opens the possibility for future investigations in the field of bibliometrics. One possible direction is the study of the large-scale structure of collaboration networks in the field of library and information science and the comparison with the relatively close field of computer science. Further interesting future goals might be the analysis of the large-scale structure and evolution of journal citation networks, and the computation of unexplored centrality measures on these networks.

Endnotes

[1]This is true to a less degree for disciplines such as medicine, biology, and experimental physics, where the average number of authors per paper is significantly larger than in computer science. On the other hand, in arts, humanities, and some social sciences, a significant share of contributions are written by a single author and hence are not collaborative works (Franceschet & Costantini, 2010).

[2]Paul Erdős was a notably eccentric Hungarian mathematician who is currently the most prolific and the most collaborative among mathematicians. He wrote more than 1,400 articles cooperating with more than 500 co-authors (Grossman, 1997). Erdős was an itinerant mathematician, living most of his life out of a suitcase visiting those colleagues willing to give him

DOI: 10.1002/asi

hospitality in exchange for collaboration in the writing of articles ("*Another roof, another proof*", he used to say).

[3]We excluded year 2009 since for it the bibliography has not reached the same level of completeness as for the previous years.

[4]We excluded years before 1960 since the number of database records for these early years of computer science is not significant. For instance, DBLP contains 259 records for year 1959, 23 records for year 1949, and 12 records for year 1936.

[5]We excluded isolated nodes from collaboration networks, that is authors who have never collaborated. The share of these authors is about 6% of the total number of authors.

[6]The term geodesic comes from geodesy, the science of measuring the size and shape of Earth; in geodesy a geodesic is the shortest route between two points on the Earth's surface.

[7]The computation of all-pairs shortest paths is computationally intensive. In the unweighted case, it takes $O(nm)$, where n and m are the number of nodes and the number of edges of the graph, respectively. Notice that our collaboration graph is sparse, being $m \simeq 3.32n$, hence the computational complexity is of the order of n^2. Using the igraph R package, the computation took more than 65h, that is 1μs per pair of nodes on average.

[8]Newman (2001c) proposed to consider also the cardinality of the author set of the co-authored papers in order to define the collaboration weight. The plausible intuition is that the intensity of the scientific relationship is higher if two scholars collaborated on a paper in which they are the sole authors than if they wrote the paper with many other collaborators. We do not consider this factor, however, since the typical computer science paper has a small number of authors (typically 2), much smaller than in other experimental sciences in which tens or even hundreds of names can sign a paper.

[9]The name comes from percolation studies in physics.

[10]Eigenvector centrality scores correspond to the values of the dominant eigenvector of the graph adjacency matrix. These scores may be interpreted as the solution of a linear system in which the centrality of a node is proportional to the centralities of the nodes connected to it.

[11]A small bias might be introduced also in the other two cases, whenever one has to partially remove nodes with the same degree or eigenvector centrality. We did not consider this issue in our experiment.

[12]As we have seen in the section Connected Components, the relative size of the largest biconnected component of the network is quite large.

[13]An undirected graph is said *complete* if every pair of different nodes is connected by an edge.

[14]A *tree* is an undirected graph that is both connected and acyclic.

[15]The transitivity coefficient of a random graph with n nodes and m edges is m/m_*, where $m_* = n(n-1)/2$ is the maximum number of edges of the graph. The transitivity coefficient of a random graph with degree sequence k is $1/n(\langle k^2\rangle - \langle k\rangle)^2/\langle k\rangle^3$, where $\langle k\rangle$ is the mean degree and $\langle k^2\rangle$ is the mean-square degree.

[16]A paper with such a large number of authors is quite unnatural in computer science. In fact, this hyper-authored paper is an article in bioinformatics, an area at the intersection of biology and computer science.

[17]Recent bibliometric studies have shown that computer science papers jointly written by two authors are generally of better quality (as judged by peer experts) than single-author papers, and the number of citations received from other papers grows with the number of authors of the paper (Franceschet & Costantini, 2010).

[18]As a comparison, the most prolific and the most collaborative among mathematicians, Paul Erdős, wrote more than 1,400 papers cooperating with more than 500 co-authors (Grossman, 1997).

[19]To the knowledge of the writer, this is the first time that such strong association between transitivity and degree assortativity is observed in the evolution of real networks.

References

Asknes, D.W. (2003). Characteristics of highly cited papers. Research Evaluation, 12(3), 159–170.

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

Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications, 311(3–4), 590–614.

Bird, C., Barr, E.T., Nash, A., Devanbu, P.T., Filkov, V., & Su, Z. (2009). Structure and dynamics of research collaboration in computer science. SIAM International Conference on Data Mining (pp. 826–837).

Birman, K., & Schneider, F.B. (2009). Program committee overload in systems. Communications of the ACM, 52(5), 34–37.

Choppy, C., van Leeuwen, J., Meyer, B., & Staunstrup, J. (2009). Research evaluation for computer science. Communications of the ACM, 52(4), 31–34.

Clauset, A., Shalizi, C.R., & Newman, M.E.J. (2009). Power-law distributions in empirical data. SIAM Review, 51, 661–703.

Crowcroft, J., Keshav, S., & McKeown, N. (2009). Scaling the academic publication process to Internet scale. Communications of the ACM, 52(1), 27–30.

Davis, A., Gardner, B.B., & Gardner, M.R. (1941). Deep South. Chicago: University of Chicago Press.

DBIS Research Group (2011). Basex – Processing and Visualizing XML with a Native XML Database, 2011 from http://www.inf.uni-konstanz.de/dbis/basex/

de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27(5), 292–306.

Elmacioglu, E., & Lee, D. (2005). On six degrees of separation in DBLP-DB and more. SIGMOD Record, 34(2), 33–40.

Erdős, P. (1972). On the fundamental problem of mathematics. The American Mathematical Monthly, 79(2), 149–150.

Fortnow, L. (2009). Time for computer science to grow up. Communications of the ACM, 52(8), 33–35.

Franceschet, M. (2010). The role of conference publications in CS. Communications of the ACM, 53(2), 129–132.

Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. Journal of Informetrics, 4(4), 540–553.

Goffman, C. (1969). And what is your Erdős number? The American Mathematical Monthly, 76(7), 791.

Goldfinch, S., Dale, T., & Rouen, K.D. (2003). Science from the periphery: Publication, collaboration and "periphery effects" in article citation rates of the New Zealand Crown Research Institutes 1995–2000. Scientometrics, 57(3), 321–337.

Grossman, J.W. (1997). Paul Erdős: The master of collaboration. The Mathematics of Paul Erdős (pp. 467–476). Berlin: Springer.

Grossman, J.W. (2002). The evolution of the mathematical research collaboration graph. Congressus Numerantium, 158, 201–212.

Harary, F. (1971). The collaboration graph of mathematicians and a conjecture of Erdős. Journal of Recreational Mathematics, 4, 212–213.

Huang, J., Zhuang, Z., Li, J., & Giles, C.L. (2008). Collaboration over time: Characterizing and modeling network evolution. International Conference on Web Search and Web Data Mining (pp. 107–116).

Katz, J.S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. Scientometrics, 40(3), 541–554.

Kautz, H.A., Selman, B., & Shah, M.A. (1997). Referral Web: Combining social networks and collaborative filtering. Communications of the ACM, 40(3), 63–65.

Larivière, V., Gingras, Y., & Archambault, E. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. Scientometrics, 68(3), 519–533.

Lawani, S.M. (1986). Some bibliometric correlates of quality in scientific research. Scientometrics, 9(1–2), 13–25.

Ley, M. (2009). DBLP: Some lessons learned. Proceedings of the VLDB Endowment, 2(2), 1493–1500.

Ley, M. (2010). The DBLP Computer Science Bibliography, 2010. Retrieved from http://www.informatik.uni-trier.de/ ley/db/

Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16(12), 317–323.

Menezes, G.V., Ziviani, N., Laender, A.H.F., & Almeida, V.A.F. (2009). A geographical analysis of knowledge production in computer science.

DOI: 10.1002/asi

International Conference on World Wide Web (pp. 1041–1050). New York: ACM Press.
Milgram, S. (1967). The small world problem. Physiology Today, 1(1), 61–67.
Millo, R.A.D., Lipton, R.J., & Perlis, A.J. (1979). Social processes and proofs of theorems and programs. Communications of the ACM, 22(5), 271–280.
Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. American Sociological Review, 69(2), 213–238.
Moreno, J. (1934). Who shall survive? New York: Beacon House.
Newman, M.E.J. (2001a). Clustering and preferential attachment in growing networks. Physical Review E, 64(2), 025102.
Newman, M.E.J. (2001b). Scientific collaboration networks: I. Network construction and fundamental results. Physical Review E, 64(1), 016131.
Newman, M.E.J. (2001c). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. Physical Review E, 64(1), 016132.
Newman, M.E.J. (2004). Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America, 101, 5200–5205.
Newman, M.E.J. (2010). Networks: An Introduction. Oxford, United Kingdom: Oxford University Press.
Newman, M.E.J., & Ghoshal, G. (2008). Bicomponents and the robustness of networks to failure. Physical Review Letters, 100(13), 138701.
Newman, M.E.J., & Park, J. (2003). Why social networks are different from other types of networks. Physical Review E, 68(3), 036122.
Odda, T. (1979). On properties of a well-known graph or what is your Ramsey number? Annals of the New York Academy of Sciences, 328, 166–172.
Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. Scientometrics, 60(3), 421–432.
Presser, S. (1980). Collaboration and the quality of research. Social Studies of Science, 10(1), 95–101.
R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. Available from: http://www.R-project.org.
Reuther, P., Walter, B., Ley, M., Weber, A., & Klink, S. (2006). Managing the quality of person names in DBLP. European Conference on Digital Libraries (pp. 508–511).
Scott, J. (2000). Social Network Analysis: A Handbook. Beverly Hills: Sage.
Vardi, M. (2009). Conferences vs. journals in computing research. Communications of the ACM, 52(5), 5.
Wasserman, S., & Faust, K. (1994). Social network analysis. Cambridge, United Kingdom: Cambridge University Press.
Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature, 393, 440–442.

DOI: 10.1002/asi