Available online at www.sciencedirect.com
ScienceDirect
Physica A 372 (2006) 333–339
PHYSICA A
www.elsevier.com/locate/physa

A topological analysis of scientific coauthorship networks

Alessio Cardillo[a], Salvatore Scellato[b], Vito Latora[a,*]

[a]*Dipartimento di Fisica e Astronomia, Università di Catania, and INFN sezione di Catania, Via S.Sofia 64, 95125 Catania, Italy*
[b]*Scuola Superiore di Catania, Via S. Paolo 73, 95123 Catania, Italy*

Available online 18 September 2006

Abstract

We study coauthorship networks based on the preprints submitted to the Los Alamos cond-mat database during the period 2000–2005. In our approach two scientists are considered connected if they have coauthored one or more cond-mat preprints together in the same year. We focus on the characterization of the structural properties of the derived graphs and on the time evolution of such properties. The results show that the cond-mat community has grown over the last six years. This is witnessed by an improvement in the connectivity properties of coauthorship graphs over the years, as confirmed by an increasing size of the largest connected component, of the global efficiency and of the clustering coefficient. We have also found that the graphs are characterized by long-tailed degree and betweenness distributions, assortative degree–degree correlations, and a power-law dependence of the clustering coefficient on the node degree.

Keywords: Structure of complex networks; Social systems

A social network is defined by a set of *actors*, mostly individuals or organizations, and a set of *ties* between couples of actors. It describes how the actors are connected through various social relationships ranging from casual acquaintance to close family bonds [1,2]. Social network analysis has emerged as a key technique in modern sociology, anthropology, social psychology and organizational studies, as well as a popular topic of speculation and study. Research in a number of academic fields has demonstrated that social networks operate on many levels, from families up to nations, and play a critical role in determining the way problems are solved, organizations are run, and the degree to which individuals succeed in achieving their goals. The shape of the social network helps determining a network's usefulness to its individuals. Networks with many weak ties [3] are more likely to introduce new ideas and opportunities to their members than closed networks with many redundant ties. That is to say that tight groups of friends share the same knowledge and opportunities, while a group of individuals with connections to other social worlds is likely to have access to a wider range of information. It is better for individual success to have few connections to a variety of networks rather than many connections within a single network. Similarly, individuals can exercise influence or act as brokers within their social networks by bridging two networks that are not directly linked [4].

These considerations can be interestingly applied to *scientific collaboration networks* (referred as SCNs from now on), a particular kind of social networks whose actors are scientists and the investigated relationships are

*Corresponding author.
E-mail address: Vito.Latora@ct.infn.it (V. Latora).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.08.059

scientific collaborations. One way to define the existence of a scientific collaboration is through scientific publications: two scientists are considered connected if they have coauthored one or more publications together. As indicated in Refs. [5–8], this appears to be a useful and reasonable definition of scientific acquaintance, for people who have been working together will know each other quite well and are more likely to set up a continuative collaboration and therefore contribute to a knowledge spread, particularly if two related scientists belong to different fields (e.g. physics and computer science). Furthermore, data related to coauthorships can be easily found on the huge publication records that are now accessible on the Internet, and offer one of the largest and most precise database to date on social networks. Focusing on SCNs by using data extracted from the publication records is not a new topic: one of the most famous result of this interest is the Erdös number (see, for instance, the Erdös Number Project [9]), which is a number assigned to each mathematician indicating the number of steps in the shortest path to the incredibly prolific Hungarian mathematician Paul Erdös on the relative SCN.

Here, we present a study of a SCN constructed by using data drawn from the Los Alamos e-Print cond-mat Archive at the website `http://xxx.lanl.gov/archive/cond-mat`. Following some previous works by Newman [5–8], and Barabási et al. [10], we construct the network by considering two scientists connected if they have coauthored one or more cond-mat preprints together in the same year. In particular, we focus on the cond-mat database in the period from 2000 to 2005, inclusive, in order to study how the pattern of collaborations have changed over time in the most recent years.

In Table 1 we report, year by year, the number of papers submitted to the archive and the average number of authors per paper. An important thing to be noticed is that the databases include also the *cross listings* papers (and authors). Such papers do not belong directly to the cond-mat archive, but they are listed there and so they have been included in the analysis. The basic properties of the six graphs under study are reported in Table 2. We notice that both the number of nodes N (the number of different authors per year), and the number of links K in the graph, increase over the years. The number of scientists who submitted at least one paper to the cond-mat archive has almost doubled in the period considered. This number is, in fact, equal to $N = 9077$ in year 2000, and equal to $N = 15\,964$ in 2005. A similar monotonic increase over the years has been observed in the number of edges, and also in the average degree, i.e., the number of different collaborators per author. This might be due to an effective growth of the cond-mat community in the past six years (as denoted for instance by the increasing number of authors per paper shown in Table 1), but also to the fact that an increasing number of scientists are getting used to submit their manuscripts to the e-Print archives. The six graphs under study show a rather high value of efficiency E [11,12] and clustering coefficient C [4], and a rather small value of the characteristic path length $\langle l \rangle$ [13]. The values of C we have found are about two times larger than the values reported by Newman in Ref. [6]. This is probably due to the fact we are considering smaller graphs. The improvement in the connectivity properties of coauthorship networks over the years is confirmed by the increasing size of the largest component [14], and by the increasing value of global efficiency. It is worth to notice that the value of the efficiency increases of about 90% in the period considered (from $E = 0.043$ in 2000 to $E = 0.071$ in 2005). Conversely, a measure of the local properties of the graph, such as the clustering coefficient C, exhibits only a slight increase by less than 5%. The behavior of $C(k)$ is shown in Fig. 1 for year 2000 and year 2005, and indicates that authors with few collaborators are more likely to work within groups in which all the scientists collaborate together (high clustering) than authors with a large degree, usually collaborating with a large number of scientists (eventually belonging to different universities and research groups) not having direct scientific collaborations one with the other. Moreover, as indicated in Fig. 1, the clustering coefficient averaged over nodes of the same degree decays approximatively as

Table 1
Some fundamental information on the Los Alamos cond-mat Archive year by year in the period 2000–2005

	2000	2001	2002	2003	2004	2005
Total number of papers	6581	7616	8395	9096	9882	10 220
(Cross listings)	(556)	(600)	(627)	(728)	(862)	(985)
Mean authors per paper	2.94	3.20	3.11	3.23	3.32	3.37

Table 2
Basic properties of coauthorship graphs in the period 2000-2005

	2000	2001	2002	2003	2004	2005
N	9077	11 013	12 125	13 377	14 732	15 964
K	21 971	31 539	32 643	38 399	44 141	48 443
$\langle k \rangle$	4.79	5.73	5.38	5.72	5.96	6.07
k_{max}	92	84	84	89	101	86
$S(\%)$	58.5	66.3	61.5	66.7	69.6	69.5
D	35	23	31	27	23	22
$\langle l \rangle$	3.18	3.54	3.28	3.54	3.66	3.62
E	0.043	0.062	0.051	0.063	0.071	0.071
C	0.69	0.71	0.71	0.72	0.72	0.73

We report the number of nodes N, the number of links K, the average degree (average number of links per node) $\langle k \rangle$, the maximum degree k_{max}, the size S of the largest connected component (in percentage of N), the diameter D, the characteristic path length $\langle l \rangle$, the global efficiency E, and clustering coefficient C.

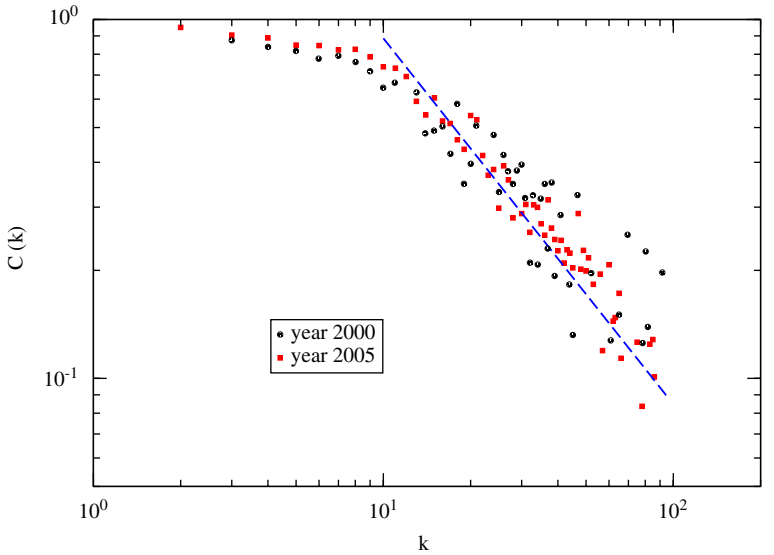

Fig. 1. Average clustering coefficient $C(k)$ of nodes with degree k. Only the results for two years, namely 2000 and 2005, are reported. Other years show the same behavior. The dashed line indicate a curve $C(k) \sim k^{-1}$.

$C(k) \sim k^{-\alpha}$, with an exponent $\alpha = 1$ (this is the same exponent found in other systems [15]). This behavior is similar to that observed in hierarchical network models [16], and indicates that there exists hierarchical structures in the network, such that several submodules combine into larger modules.

Now we consider how the degree and the betweenness [17] are distributed among the nodes of the graph. The cumulative distribution $P_c(k)$ and $P_c(b)$ are respectively defined as

$$P_c(k) = \sum_{k'=k}^{+\infty} \frac{N(k')}{N}, \quad P_c(b) = \int_b^{\infty} \frac{N(b')}{N}\, \mathrm{d}b', \tag{1}$$

where $N(k)$ is the number of nodes with degree equal to k, and $N(b)\,\mathrm{d}b$ is the number of nodes with betweenness in the range $b, b + \mathrm{d}b$. In Fig. 2 we report the results for the year 2000. Plots relative to the other years show similar shapes. Instead of $P_c(k)$ and $P_c(b)$ we prefer to plot $N_c(k) = NP_c(k)$ and $N_c(b) = NP_c(b)$. The curve reported in the left panel indicates that the typical $N_c(k)$ is long tailed due to the presence of

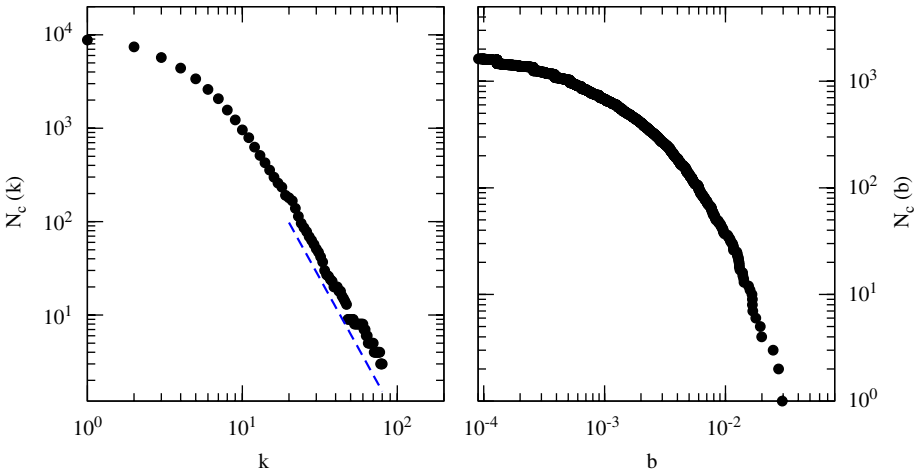

Fig. 2. Left panel. Cumulative degree distribution (year 2000): the number of nodes $N_c(k)$ with degree larger or equal to k is reported as a function of k. The line is a power law $N_c(k) \sim k^{-(\gamma-1)}$ with an exponent $\gamma - 1 = 3$. Right panel. Cumulative node betweenness distribution (year 2000): the number of nodes $N_c(b)$ with betweenness larger or equal to b is reported as a function of b.

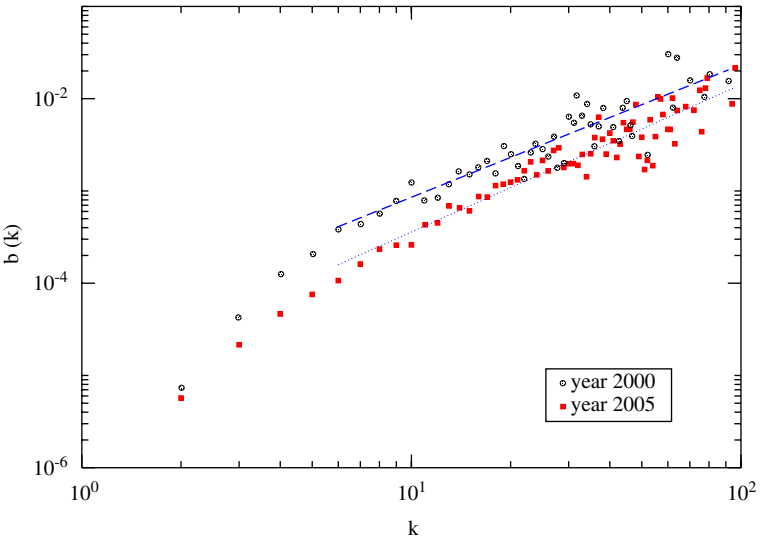

Fig. 3. Average betweenness $b(k)$ of nodes with k links, as a function of k for the years 2000 and 2005. The straight lines are power-law fits $b(k) \sim k^{\eta}$ with an exponent η, respectively, equal to 1.4 (dashed line) and 1.6 (dotted line).

individuals with a large number of collaborators, up to 100 different collaborators in one year. To notice that such numbers are even larger in other coauthorship networks, as for instance in the Stanford Public Information Retrieval System (SPIRES), a database of both theoretical and experimental papers in high-energy physics, due to presence of extremely large experimental collaborations in high-energy physics [6]. The straight line reported in the left panel shows how a power law $N_c(k) \sim k^{-(\gamma-1)}$, with an exponent $\gamma - 1$ equal to 3, would look like. By fitting the tails in $N_c(k)$ by power laws we have found a value of the exponent $\gamma - 1$ ranging from 2.5 to 3. These are steeper slopes than those observed by Newman for a single network including the preprints submitted to cond-mat in the 10-years period 1992–2001 [6,7]. For what concerns the betweenness distributions, we observe long tails, although not perfect power laws. In Fig. 3 we have studied the correlations degree–betweenness by reporting the average betweenness $b(k)$ of nodes with degree k as a function of k. We have found that the relation between the distributions of b and k is of the form $b \sim k^{\eta}$, similar

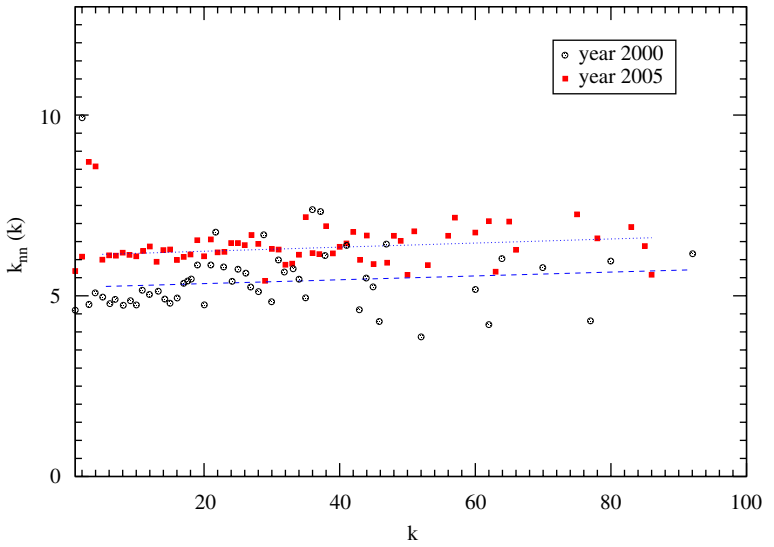

Fig. 4. Average degree $k_{nn}(k)$ of nearest neighbors of nodes with degree k, as a function of k. Results refer to year 2000 and year 2005. The two dashed lines are fitting to the data, with slopes v respectively equal to 0.005 and 0.006

to that observed in other networks [20]. The values of the exponent η range, in the years considered, from a minimum value of 1.4 to a maximum equal to 1.8. In the figure, we report the cases for the years 2000 and 2005. Unfortunately, it was not possible to check the relation among γ, η and δ (the exponent of the power law betweenness distribution) predicted in Ref. [20], since it was not possible to extract δ (see Fig. 2).

The SCN databases we have studied reveal the presence of degree–degree correlations. Degree correlations can be investigated by plotting the average degree of the nearest neighbors of nodes with degree k, $k_{nn}(k)$, as a function of k, and by measuring the numerical value of the slope, denoted in the following as v [21]. In Fig. 4 we show the cases for the year 2000 and the year 2005. The two graphs are slightly assortative, as denoted by the positive slopes of the curves $k_{nn}(k)$. This means that the nodes tend to connect to their connectivity peers, i.e., authors with a high number of collaborators tend to collaborate with other highly connected authors. The value of v we have extracted is, respectively, equal to 0.005 in 2000 and 0.006 in 2005, and in general v shows a slow tendency to increase with time in the years considered.

Finally, we addressed the question of how to find who are the best connected authors in the SCN. In order to quantify the importance of a node in the graph, we have considered two different centrality indices [1,22], namely the degree and the betweenness. The ranked lists of authors can be found at `http://www.ct.infn.it/~latora`. We notice that some of the authors are in the top 10 in more than a single year: for instance, in the degree-based list, this is the case of Y. Tokura (present in each of the six years), J. Sarrao (five years), H. Eisaki (four) and A. Revcolevschi (three). Some of the authors in the degree-based top 10, as Y. Tokura and A. Revcolevschi, have also a very high value of the betweenness. Conversely, there are authors in the top rank by betweenness that do not appear among the 10 nodes with the largest degree: two examples are A.R. Bishop and S.D. Sarma. There is an important note about the betweenness. Since nodes with a large value of b have usually also a large value of k (see correlations in Fig. 3), it can be interesting to study which of the collaborators of authors with a large value of b contribute the most to connecting couples of scientists. For such a reason, we have also computed the edge betweenness of the six graphs. In Table 3 we report the 10 edges with the highest value of betweenness found for each of the years, i.e., the 10 collaborations that connect the largest couples of scientists. One could expect that highest betweenness edges are those connecting two nodes both having an extremely large centrality value. Instead, we notice that one of the two nodes is an author also appearing in the node top 10s, while the second node it has been checked to be an author with an intermediate degree value. The argument we propose to explain this result is that many of the highest betweenness edges are coauthorships between a well known scientist and a younger collaborator who afterwards changed team and from there on continued working with other groups. A statistical investigation

Table 3
The 10 links (collaborations) with the highest betweenness are listed, in order of rank, for each of the six years considered

	2000	2001	2002	2003	2004	2005
1	Y. Tokura	M.P. Maley	S.W. Cheong	S. Tajima	A.G. Biblioni	Y. Ando
	T.Y. Koo	C. Helm	S.V. Diaz	E.A. Yelland	K.O. Rasmussen	L. Li
2	R.L. Greene	A.R. Bishop	S.D. Sarma	X. Liu	Y. Ando	W. Wegscheider
	B.G. Kim	A. Smerzi	E. Demler	W.A. Atkinson	T. Suzuki	K.L. Kampman
3	Y. Tokura	A. Smerzi	S.I. Lee	J.K. Furdyna	B. Keimer	X. Wang
	C. Kim	W.H. Zurek	P. Samuely	T. Wojtowkz	S. Ono	W. Zhuo
4	Y. Tokura	C. Geibel	S.D. Sarma	L.W. Molenkamp	A.G. Biblioni	A.C. Gossard
	N. Motoyama	A.A. Menovsky	B. Koiller	S.K. Bose	E. Kaneshita	F.M. Mendoza
5	G. Aeppli	A.R. Bishop	S.I. Lee	S. Tajima	S. Uchida	Q. Niu
	T.E. Mason	W.H. Zurek	K. Yoshida	A.Q.R. Baron	A.N. Lavrov	S. Wang
6	R. Coldea	A.R. Bishop	E. Kaminska	M. Matsumoto	S.D. Sarma	X. Zhou
	T.E. Mason	C.M. Chang	T. Dieti	S. Lee	M.P. Lilly	F. Yang
7	A. Revcolevschi	C. Geibel	A.J. Millis	X. Liu	Y. Ando	X. Zhou
	N. Motoyama	O.E. Kvitnitskaya	E.H. Hwang	M.B. Stone	N. Mannella	X. Chen
8	Y. Tokura	A. Saxena	S.I. Lee	Y. Tanaka	I. Martin	L. Li
	H. Eisaki	F.X. Bronold	J.H. Kim	D. Ishikawa	Z.G. Yu	W. Willinger
9	E.W. Plummer	C.C. Homes	D.D. Awschalom	T. Sasaki	K.W. West	N. Samarth
	G.S. Tian	E.M. Choi	A. Mascarenhas	C.T. Lin	I.B. Spielman	X. Li
10	S.W. Cheong	Y.G. Naidyuk	S.I. Lee	T. Sasaki	A. Fujimori	Y. Tokura
	T.E. Mason	W.N. Kang	M.W. Kim	K. Kindo	A.N. Lavrov	A. Damascelli

of how the betweenness of an edge is correlated to the centrality of the two nodes joined by the edge is left for future work.

The study of coauthorship networks allows to derive some interesting conclusions on the properties of SCN and, more in general, on the structural properties of a social system. Our analysis of the Los Alamos cond-mat preprint database during the period 2000–2005 reveals that the cond-mat community has grown over the last six years. This is witnessed by an improvement in the connectivity properties of the coauthorship graph over the years, as confirmed by an increasing size of the largest connected component, of the global efficiency and of the clustering coefficient. Moreover, we have found that the graphs are characterized by assortative degree–degree correlations, and a power-law dependence of the clustering coefficient on the node degree. We have also focused on the centrality distribution of degree and betweenness providing a list of authors and couples of authors in the centrality top ranks. In order to get an overall picture of SCN, longer time windows have to be analyzed, and different kinds of research field should be explored. Moreover, a study in term of weighted relations could help to better investigate the real structure of a scientific collaboration, allowing a more accurate identification of key-role playing actors. However, this aspect generates further discussions on the kind of weights and distances to be used [6,7].

References

[1] S. Wasserman, K. Faust, Social Networks Analysis, Cambridge University Press, Cambridge, 1994.
[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Phys. Rep. 424 (2006) 175.
[3] M. Granovetter, Sociological Theory 1 (1983) 201.
[4] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[5] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 98 (2001) 404.
[6] M.E.J. Newman, Phys. Rev. E 64 (2001) 016131.
[7] M.E.J. Newman, Phys. Rev. E 64 (2001) 016132.
[8] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 101 (2004) 5200.
[9] Erdös Number Project Page: ⟨http://www.oakland.edu/enp/⟩.
[10] A.-L. Barabási, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, A. Schubert, Physica A 311 (2002) 590.
[11] V. Latora, M. Marchiori, Phys. Rev. Lett. 87 (2001) 198701.
[12] V. Latora, M. Marchiori, Eur. Phys. J. B 32 (2003) 249.

[13] The efficiency E is a better measure than the average of the shortest path lengths, $\langle l \rangle$, if there are disconnected components in the graph. In fact, in such a case, the definition of $\langle l \rangle$ should be restricted only to couples of nodes connected by a path [11].
[14] Here we are focusing on the size of the largest connected component only. The investigation of the component size distribution is left for future work.
[15] E. Ravasz, A.-L. Barabási, Phys. Rev. E 67 (2003) 026112.
[16] A.-L. Barabási, E. Ravasz, T. Vicsek, Physica A 299 (2001) 559.
[17] The betweenness of a node (edge) is defined as being proportional to the number of shortest paths between pairs of nodes that run through the node [18] (the edge) [19].
[18] C.L. Freeman, Sociometry 40 (1977) 35.
[19] M.E.J. Newman, M. Girvan, Phys. Rev. E 69 (2004) 026113.
[20] M. Barthélemy, Eur. Phys. J. B 38 (2004) 163–168.
[21] R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. Lett. 87 (2001) 258701.
[22] P. Crucitti, V. Latora, S. Porta, Phys. Rev. E 73 (2006) 036125.