UNIVERSITÀ DEGLI STUDI DI CATANIA
FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
CORSO DI LAUREA IN FISICA

ALESSIO VINCENZO CARDILLO

TOPOLOGICAL ANALYSIS OF SCIENTIFIC COAUTHORSHIP NETWORKS

ELABORATO FINALE

RELATORI:
CHIAR.MO PROF. V. LATORA

ANNO ACCADEMICO 2005/2006

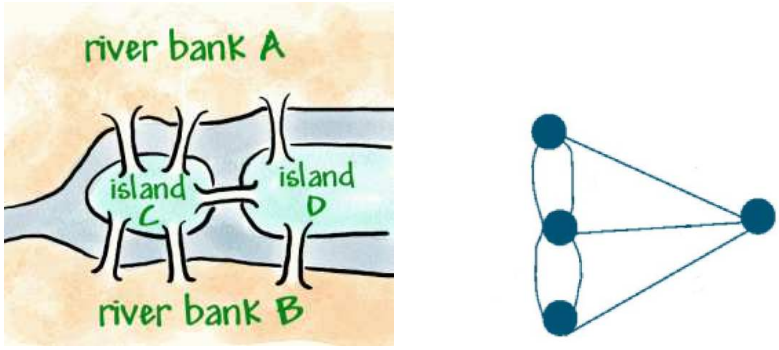
A Sacha

Contents

1 Introduction

Networks are all around us, and we are ourselves, as individuals, the unit of a network of social relationships of different kinds and, as biological systems, the delicate result of a network of biochemical reactions. Networks can be tangibile objects in the Euclidean space, such as electric power grids, the Internet, highway or subway systems, and neural networks. Or they can be entities defined in an abstract space, such as networks of acquaintances or collaborations among individuals.

Historically, the study of networks has been mainly the domain of a branch of discrete mathematics since its birth in 1736, when Leonard Euler published the solution to the Könisberg bridge problem (it consisted in finding a round trip that traversed each of the Könisberg's bridges exactly once; cf. Fig 1(a)), this theory is known as *graph theory*. Graph theory has developed and has provided answers to a series of practical questions such as: what is the maximum flow per unit time from source to sink in a network of pipes? How to assing n jobs to n people with maximum total utility? In addition, the study of networks has seen important achievements in some specialized contexts such as social sciences.



(a) (b)

Figure 1: A schematic view of the Könisberg bridge: (a) map view and (b) corrisponding graph.

The last decade has known the birth of a new movement of interest and research in the study of *complex networks*, i.e. networks whose structure is irregular, complex and dynamically evolving in time, with a renewed attention to the properties of networks of dynamical units. This flurry of activity has been triggered by two seminar papers: one by Watts and Strogatz on small-world network appeared in 1998 in "Nature" [1], and one by Barabàsi and Albert on scale-free networks appeared in 1999 in "Science" [2]. The main character of this activity was the physicists' community. They were induced by the possibility to study the properties of a plenty of large databases of real networks. These include transportation networks, phone call networks, the Internet and the World Wide Web, the actors' collaboration network in movie databases, neural or genetic networks, metabolic and protein networks, scientific coauthorship and citation networks from the Science Citation Index.

This thesis shows some results concerning the *topological* features of networks based on the collaboration among scientists, physicists in particular. The study of some basic characteristics, such as the *degree* and the *clustering coefficient* joined with the *centrality indexes* such as *betweenness* and *closeness*, allows a deeper comprehension of the inner stucture of these networks and the individuation of key role played by people using an objective criterion. Moreover, another goal of this thesis is to illustrate some techniques used to analyze social networks and discuss their applications. Following previous works by Newman [3, 4, 5, 6], and Barabási *et al.* [7], our networks are constructed by considering two scientists connected if they have coauthored one or more preprints together in the same year. In particular, we focus on the Los Alamos preprint database `http://xxx.lanl.gov/` in the period from 2000 to 2005, in order to study how the pattern of collaborations have changed over time in recent years. This thesis is based on the paper of Cardillo et al. [8] that is going to appear in Physica A (in press at the time of writing).

1.1 Social Networks

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 bounds [9, 10]. Social network analysis has emerged as a key technique in modern sociology, anthropology, social psychology and organizational studies.

Research in a number of academic fields has demonstrated that social networks operate on several 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 Fig. 2.

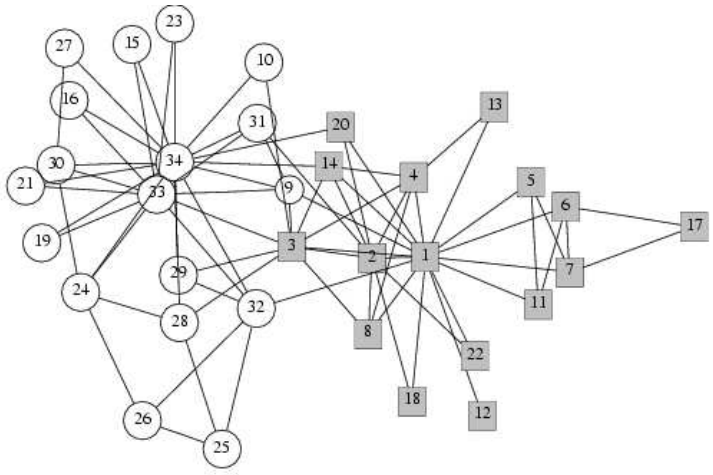


Figure 2: An example of social network, the Zachary *Karate Club* social network [11].

It is interesting to apply these considerations to *Scientific Collaboration Networks* (referred below as SCNs from now on), a particular kind of social networks whose actors are scientists and the investigated relationships are 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. [3, 4, 5, 6], 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 spread knowledge, 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 SCN 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 `http://www.oakland.edu/enp/`), 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.

2 Graphs

For a greater comprehension of the results obtained, it is necessary to introduce some definitions and notations. Definitions are related with the characterization and modelling of the structural properties of a network. Graph theory [12, 13, 14] is the natural framework for the exact mathematical treatment of complex networks and, formally, a complex network can be represented as a graph. An *undirected* (*directed*) *graph* $G = (\mathcal{N}, \mathcal{L})$ consist of two sets $\mathcal{N}$ and $\mathcal{L}$, such that $\mathcal{N} \neq \emptyset$ and $\mathcal{L}$ is a set of unordered (ordered) pairs of elements of $\mathcal{N}$. The elements of $\mathcal{N} \equiv \{n_1, n_2, \ldots, n_N\}$ are the ***nodes*** (or *vertices*, or *points*) of the graph G, while the elements of $\mathcal{L} \equiv \{l_1, l_2, \ldots, l_k\}$ are its ***links*** (or *edges*, or *lines*). The number of elements in $\mathcal{N}$ and $\mathcal{L}$ are denoted by N and K, respectively. Then, later, a graph will be indicated as $G(N, K) = (\mathcal{N}, \mathcal{L})$, or simply $G(N, K)$ or G_N, whenever it is necessary to emphasize the number of nodes and links in the graph. A powerful way to represent

a graph is the *adjacency* (or *connectivity*) matrix $\mathcal{A}$, a $N \times N$ square matrix whose entry $a_{ij}\,(i,j=1,\ldots,N)$ is equal to 1 when the link l_{ij} exists, and zero otherwise. However, many real networks display a large heterogeneity in the capacity and intensity of connections. Examples are the diversity of predator-prey in food webs, passengers in airline networks, and the existence of weak ties between individuals in social networks [15, 16, 17, 18]. These systems can be better described in terms of *weighted networks*, i.e. networks in which each link carries a numerical value measuring the strength of the connection. In this sense a new set $\mathcal{W} \equiv \{w_1, w_2, \ldots, w_K\}$ must be considered. With this new position a graph G is defined as: $G^W = \{\mathcal{N}, \mathcal{L}, \mathcal{W}\}$.

2.1 Topological characteristics

When weighted networks are considered, it is useful to represent them using another matrix over the adjacency one. In these cases it is useful to consider a *weight matrix* $\mathcal{W}$, a $N \times N$ square matrix whose entry $w_{ij}\,(i,j=1,\ldots,N)$ is equal to the link weight. This study focuses on the *topological characteristics* of network based on SCN data. This means that our networks are all *unweighted* and the weight of each link $w_{i,j}$ is described as:

$$w_{i,j} = \begin{cases} 1, & \text{if the edge between nodes } i \text{ and } j \text{ exists,} \\ 0, & \text{otherwise.} \end{cases}$$

In this context, it is important to define some quantities to help the estimation of these characteristics.

2.1.1 Node degree, degree distribution, and correlation

The ***degree*** (or *connectivity*) k_i of a node i is the number of edges incident with it. It is defined in terms of the adjacency matrix $\mathcal{A}$ as:

$$k_i = \sum_{j \in \mathcal{N}} a_{ij}. \tag{1}$$

The most basic topological characterization of a graph G can be obtained in terms of the ***degree distribution*** $P(k)$, defined as the probability that a node uniformly chosen at random has degree k. The degree distribution completely determines the statistical properties of uncorrelated networks. A large number of real networks are *correlated* in the sense that the probability that a node of degree k is connected to another node of degree, say k', depends on k. In these cases, it is necessary to introduce the ***conditional probability*** $P(k'|k)$, defined as the probability that a link from a node of degree k points to a node of degree k'. $P(k'|k)$ satisfies the degree detailed balance condition $kP(k'|k)P(k) = k'P(k'|k)P(k')$ [19, 20]. For uncorrelated graphs, the balance condition gives $P(k'|k)P(k) = k'P(k')/\langle k\rangle$. Instead of using conditional probability, due to finite size problems, it is possible to define the ***average nearest neighbours degree*** of a node i as:

$$k_{nn,i} = \frac{1}{k_i}\sum_{j\in\mathcal{N}_i} k_j = \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}k_j \tag{2}$$

where the sum runs on the nodes belonging to $\mathcal{N}_i$, the set of nearest neighbours of i. The average nearest neighbours degree may be expressed in terms of conditional probability using definition (2) as:

$$k_{nn}(k) = \sum_{k'} k'P(k'|k). \tag{3}$$

Correlated graphs are classified as ***assortative*** if k_{nn} is an increasing function of k, whereas they are referred to as ***disassortative*** when k_{nn} is a decreasing function of k [21]. In other words, in assortative networks the nodes tend to connect to their connectivity peers, while in disassortative networks nodes with a lower degree are more likely connected with highly connected ones.

2.1.2 Clustering

Clustering, also known as transitivity, is a typical property of acquaintance networks, where two individuals with a common friend are likely

to know each other [9]. This can be quantified by defining the *transitivity* T of the graph as the relative number of triples, i.e. the fraction of connected nodes which also form triangles. An alternative possibility is to use the graph's ***clustering coefficient*** C, a measure introduced by Watts and Strogatz [1] defined as follows. A quantity c_i (local clustering of node i) is introduced, expressing how likely $a_{jm} = 1$ for two neighbors j and m of node i. Its value is obtained counting the number of edges (denoted by e_i) in the subgraph G_i of node i neighbours. The local clustering coefficient is defined as the ratio between e_i and $k_i(k_i - 1)/2$, the maximum possible number of edges in G_i:

$$c_i = \frac{2e_i}{k_i(k_i - 1)} = \frac{\sum_{j,m} a_{ij}a_{jm}a_{mi}}{k_i(k_i - 1)}. \tag{4}$$

The clustering coefficient of the graph C is the average of c_i over all nodes in G:

$$C \equiv \langle c \rangle = \frac{1}{N} \sum_{j \in \mathcal{N}} c_i. \tag{5}$$

By definition, $0 \leqslant c_i \leqslant 1$, and $0 \leqslant C \leqslant 1$. It is also useful to consider $c(k)$, the clustering coefficient of a connectivity class k, which is defined as the average of c_i over all nodes with degree k.

2.2 Centrality indexes

In order to identify the key role playing subject in a network, many different *centrality indexes* and *measures* have been defined. These quantities allow to estabilish some criteria for comparison analysis of network, and thus making their study easier for the researchers. Some of these quantities were initially introduced to quantify the importance of an individual in a social network [9].

2.2.1 Shortest path lengths, diameter and betweenness

The shortest path plays an important role in transport and communication within a network. Suppose one needs to send a data packet from

one computer to another through the Internet: the *geodesic* or ***shortest path*** is the shortest path connecting two nodes (in an unweighted network the minimum number of hops to reach node j from i). Geodesics provide optimal path way, since one would achieve a fast transfer and save of system resources. For such a reason, shortest paths have also played an important role in the characterization of the internal structure of a graph [9]. It is useful to represent all the shortest path lengths of a graph G as a matrix $\mathcal{D}$ in which the entry d_{ij} is the length of the geodesic from node i to node j. The maximum value of d_{ij} is called the ***diameter*** of the graph. A measure of the typical separation between two nodes in the graph is given by the ***average shortest path length***, also known as the ***characteristic path length***, defined as the mean of geodesic lengths over all couples of nodes [22, 9].

$$L = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{N}, i \neq j} d_{ij}. \tag{6}$$

A problem arising from the above definition is that L diverges if there are disconnected components in the graph. One possibility to avoid such a divergence is to limit the summation in Eq. (6) only to pairs of nodes belonging to the largest connected component, or ***giant component***. An alternative and in many cases, useful approach is to consider the harmonic mean of geodesic lengths, and to define the so-called ***efficiency*** of G as [17, 23]:

$$E = \frac{1}{N(N-1)} \sum_{i,j \in \mathcal{N}, i \neq j} \frac{1}{d_{ij}}. \tag{7}$$

Such quantity is an indicator of the traffic capacity of a network, and avoids the divergence in Eq. (6), since any pairs of nodes belonging to disconnected component of the graph yields a contribution equal to zero to the summation in Eq. (7).

The communication of two non-adjacent nodes, say j and k, depends on the nodes belonging to the paths connecting j and k. Consequently, a measure of the relevance of a given node can be obtained by counting the number of geodesics going through it, and defining the so called

node betweenness. Together with degree and ***closeness*** of a node (defined as the inverse of the average distance from all other nodes), the betweenness is one of the standard *measure of node centrality*, originally introduced to quantify the importance of an individual in a social network [9]. More precisely, the ***betweenness*** b_i of a node i is defined as [9, 24, 25]:

$$b_i = \sum_{j,k \in \mathcal{N}, j \neq k} \frac{n_{jk}(i)}{n_{jk}}. \tag{8}$$

where n_{jk} is the number of shortest paths connecting i and k, while $n_{jk}(i)$ is the number of shortest paths connecting i and k and passing through i. The concept of betweenness can be extended also to edges. The ***edge betweenness*** is defined as the number of shortest paths between pairs of nodes which run through that edge [15].

3 Scientific Coauthorship Network

A large number of papers have been written about social networks [26, 27, 28, 29], and in particular about scientific collaboration networks. A social network is defined by a set of *actors*, and a set of *ties*. It describes how actors are connected through various social relationships ranging from casual acquaintance to close family bounds [9]. Research in a number of academic fields has demostrated that social networks operate on many levels 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 networks helps determinig a network's usefulness to its individuals. Networks with many weak ties [18] 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 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 [3, 30].

3.1 Introduction

These considerations can be interestingly applied to *Scientific Coauthorship Networks* or SCN, a particular kind of social networks whose actors are scientists and the investigated relationships are 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. This appear to be a useful and reasonable definition of scientific aquaintance, because 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 spread knowledge, particularly if two related scientist belong to different fields (i.e. physics and computer science). Furthermore, data related to coauthorship can be easily found on the huge publication records that are now accesible on the Internet, and offer one of the largest and most precise database to date on social networks. Focusing on SCN by using data extracted from the publication records is not a new topic: one of the most famous results of this interest is the Erdös number on web at `http://www.oakland.edu/enp/`, 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.

3.2 SCN Based on Los Alamos Archive

Here a study of a SCN constructed by using data drawn from the Los Alamos e-Print Archive at the website `http://xxx.lanl.gov/` is presented. In particular, the sub archives considered are:

- Condensed Matter `[ cond-mat ] ;`
- General Relativity and Quantum Cosmology `[ gr-qc ] ;`

- High Energy Physics - Experiment [hep-ex] ;
- High Energy Physics - Lattice [hep-lat] ;
- High Energy Physics - Phenomenology [hep-ph] ;
- High Energy Physics - Theory [hep-th] ;
- Mathematical Physics [math-ph] ;
- Nuclear Experiment [nucl-ex] ;
- Nuclear Theory [nucl-th] ;
- Physics [physics] ;
- Quantum Physics [quant-ph] ;

Following previous works by Newman and Barabàsi *et al.* [3, 4, 7], these networks are constructed by considering two scientists connected if they have coauthored one or more preprints together in the same year. In particular, two different kinds of analysis are made: time-series topological analysis during the period 2000 - 2005, and one year (2005) topological analysis. The former has been made on cond-mat , hep-ex and hep-ph ; the latter has been made on the others archives.

One could ask "Why use a database of preprints instead of regular articles ones?" The reason is that in preprint archives, articles published (or unpublished) in different magazines are stored. Such systems allow a better view of the scientific collaboration background, thus making results more affordable and general.

In order to analyze the structure of SCN networks, it is necessary to find a database where one can get data from preprint databases. There are several databases available on the net where to find such data. An example is the Stanford Public Information Retrieval System (SPIRES), a database of preprints and published papers in high-energy physics, both theoretical and experimental. Another example is the Networked

Computer Science Technical Reference Library (NCSTRL), a database of preprints in computer science, submitted by participating institutions and stretching back about fifteen years.

The Los Alamos e-Print Archive is a database containing unrefereed preprints in physics, self-submitted by their authors, running from 1992 to the present. This database is subdivided into specialities within physics, such as condensed matter and high energy physics. Los Alamos has been chosen for data collection because it gives the widest view upon the physicist communities since 1992. To be exact, Los Alamos provides open access to 379,778 e-prints in Physics, Mathematics, Computer Science and Quantitative Biology so it could be used to analyze communities different from the physicists' one.

4 Data

The study of the SCN based networks is based on the data collected from Los Alamos archive. Two steps have been done in order to obtain results discussed in Section 5. First, data were collected and then they have been analyzed. In the following these two steps are discussed.

4.1 Data collecting

The data collected for this study is taken from the website of Los Alamos archive. Once selected, the chosen (for example cond-mat), a JAVA™ program is used to make a parsing of the HTML web pages code of each year in order to extract information about authors and then build an adjacency list, used to generate all the results in this work, and a match list used to associate at each node ID a key on the database. The match list and the adjacency list are made using the database analysis technique of *associative array* [31]. An example of the data structure provided by this programme is shown below:

1 2 ⟶ adjacency list

1 Ko ⟶ (Japanese for this)
2 Otsu ⟶ (Japanese for that).

As discussed above, two kinds of data analysis have been performed. For this reason two "different" kinds of data have been collected. The reason for making a graph for each year is both conceptual and computational. In fact in order to study the evolution of the topological characteristics it is necessary to have more graphs at different times. The longer the time period between two graphs, the more accurate the is study. However, a too large graph (a graph with many nodes and links) makes computations very slow. Therefore a good compromise to balance these two is to get a single year data graph. Once data have been collected, various algorithms have been run on them. For example, Djikstra alghorithm [32] to calculate the shortest path, or the associative array to create the corrispondency list. Tables from 1 – 4 contain all the topological data for that archive/year.

An important thing to note is that the databases include also the *cross listings* papers (and authors). Such papers do not belong directly to the archive considered, but they are listed there and so they have been included in the analysis. As we can see, each archive has a proper structure and a characteristic that may be found by considering the whole results or only a few of them. For example, the small size of the giant component in the `gr-qc` or `math-ph` archives indicates that such communities are divided in a lot of sub communities. Another example is given by the high mean degree of NUCL-EX archive. This is due to the fact that a lot of people are involved in nuclear physics experiments, so papers have a large number of authors.

4.2 Multiple author problem

The first thing that has to be said is that a background problem (error) affects results. This error is due to the fact that in the database each

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Tot. papers (*cross listings*)	6581 (556)	7616 (600)	8395 (627)	9096 (728)	9882 (862)	10220 (985)
Mean authors per paper	2.94	3.20	3.11	3.23	3.32	3.37
N	9077	11013	12125	13377	14732	15964
K	21971	31539	32643	38399	44141	48443
$\langle k \rangle$	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

Table 1: Basic properties of coauthorship graphs of `cond-mat` archive in the period 2000-2005. Here, it is reported the number of papers, the average number of authors per paper, 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 the clustering coefficient C.

author is related to an ID. This ID is mainly composed by the surname and the first initial of the name of the author. For example:

Alessio Cardillo $\longrightarrow$ Cardillo_A

Alessio Vincenzo Cardillo $\longrightarrow$ Cardillo_A

This cause a multiple author assignment to a single ID with a consequently misleading of results. To avoid this problem, another rule for ID assignment has been created. It is a modification of the previous one in the following sense: each ID is composed by the surname of the author followed by all the capital letters of all its names. For example:

Alessio Cardillo $\longrightarrow$ Cardillo_A

Alessio Vincenzo Cardillo $\longrightarrow$ Cardillo_AV

	gr-qc	hep-lat	hep-th	math-ph	nucl-ex	nucl-th	physics	quant-ph
Tot. Papers (*cross listings*)	2670 (1006)	861 (197)	4659 (1419)	1881 (928)	717 (256)	1596 (531)	3326 (566)	3376 (515)
Mean auth. per paper	2.099	3.581	2.127	1.896	6.261	2.932	2.974	2.632
N	2047	798	3537	1223	2054	1528	5616	3855
K	4169	2928	4382	929	23786	3431	18285	8199
$\langle k \rangle$	4.033	7.338	2.409	1.519	23.161	4.499	6.433	4.256
k_{max}	60	51	15	9	107	27	84	54
$S(\%)$	22.03	73.81	31.83	2.94	70.69	47.51	3.79	43.53
D	26	12	27	12	15	22	11	18
$\langle\, l \,\rangle$	0.488	2.752	1.035	0.008	3.023	1.950	0.012	1.300
E (10^{-2})	0.866	12.8	1.21	0.185	10.77	3.466	0.232	3.201
C	0.503	0.672	0.488	0.341	0.864	0.669	0.730	0.617

Table 2: Basic properties of coauthorship graphs. Here is reported the number of paper, the average number of authors per paper, number of nodes N, number of links K, the average degree $\langle k \rangle$, 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 the clustering coefficient C.

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Tot. papers (*cross listings*)	1079 (343)	1203 (385)	1211 (413)	1175 (404)	1333 (448)	1261 (406)
Mean authors per paper	3.28	3.02	3.04	3.33	2.75	3.48
N	1470	1495	1604	1643	1405	1797
K	13268	12491	13865	14378	13444	20037
$\langle k \rangle$	18.04	16.70	17.27	17.50	19.13	22.30
k_{max}	143	69	126	81	106	135
$S(\%)$	13.67	9.03	44.45	10.77	13.59	39.23
D	9	12	12	7	6	8
$\langle l \rangle\ 10^{-2}$	9.067	6.505	94.458	7.018	8.265	57.195
$E\ 10^{-2}$	2.142	1.643	5.714	1.769	2.341	5.779
C	0.722	0.691	0.670	0.735	0.649	0.744

Table 3: Basic properties of `hep-ex` archive coauthorship graphs in the period 2000-2005. Here is reported the number of paper, the average number of authors per paper, 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 the clustering coefficient C.

However this method cannot solve two further problems. The first is that even with this method two authors like Alessio Cardillo and Alessandro Cardillo will have the same ID, Cardillo_A . The second problem is that some authors have more than a name but they do not use always both in papers. For example Alessio Cardillo and Alessio Vincenzo Cardillo are the same person (author of this thesis) but they will be identified as Cardillo_AV and Cardillo_A, respectively. This fact is due to a wrong policy of database mantainers, who do not provide a unique identifier per authors (like MathSciNet does), and a lack of authors submission rules in this sense.

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
Tot. papers (*cross listings*)	5204 (1080)	5357 (1131)	5443 (1198)	5179 (1215)	5354 (1216)	5148 (1229)
Mean authors per paper	2.40	2.30	2.38	2.34	2.45	2.46
N	4178	4193	4277	4152	4322	4324
K	20599	10505	15983	10076	21591	15597
$\langle k \rangle$	9.86	5.00	7.47	4.85	9.98	7.21
k_{max}	195	94	136	91	228	149
$S(\%)$	58.47	57.19	57.59	53.23	60.46	57.42
D	21	24	22	25	17	22
$\langle\, l \,\rangle$	2.212	2.569	2.189	2.319	2.213	2.266
E	0.064	0.049	0.060	0.041	0.072	0.058
C	0.608	0.571	0.594	0.596	0.608	0.620

Table 4: Basic properties of `hep-ph` archive coauthorship graphs in the period 2000-2005. Here is reported the number of paper, the average number of authors per paper, 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 the clustering coefficient C.

4.3 Data analysis

The goal of the first type of data analysis is to calculate the topological characteristics of all the graphs. Total number of authors, mean number of authors per paper, number of nodes and links, mean and maximum degree, clustering coefficient and so on are listed in tables 1–4. In order to find assortative properties of these networks, degree–degree, clustering–degree and degree–betweenness correlation have been studied. Results of this analysis are explained in Section 5 and in Figs. 4 and 5. In addition, even the cumulative distribution of degree , betweenness and closenes have been studied to compare their characteristics with the ones found by Newman [3, 4] and with the results of graph theory. The use the analysis of multiple indexes allow a better identification of both structural and

sociological characteristics. Indeed, the connectivity peers in assortative networks may be also related with high betweeneess peers. Instead, the relation between degree and clustering help to determine if peers have their own sub communities or not. All these relations, combined with a study of the centrality index, are very useful in social sciences.

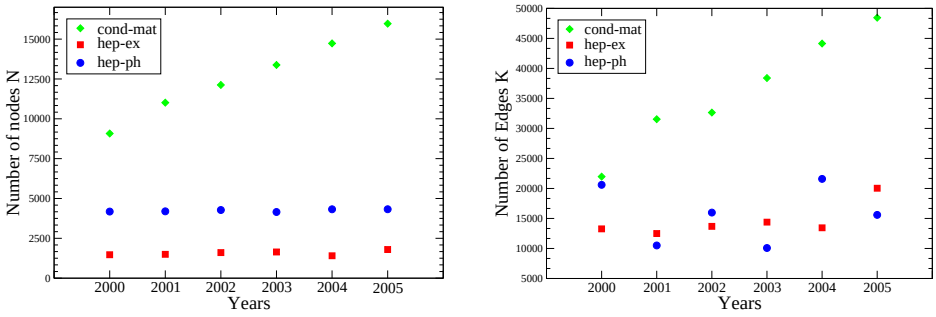
Finally, the last kind of data analysis is focused on answering the question of how to find the best connected authors in the SCN. In order to quantify the importance of a node in the graph, different centrality indices [9, 33] have been considered, namely the degree, the betweenness (both node and edge) and the closeness. Using this method, the discovery of the best connected authors is easier, and the result demostrates the effectiveness of a centrality–based method. Indeed, in the ranked lists it is easy to find names of important people in the scientific community. In addition, this result could be made even more reliable if we consider weighted network. However, there is no unique procedure to assign a weight to links. In fact, there are many ways to define a weight for coauthorship networks. The ranked lists of authors over all these indices can be found on: `http://www.ct.infn.it/~cardillo/`.

5 Results

Since two different kinds of data have been collected, two different types of analysis have been done on them. This leads to two different kinds of results. One kind is related with time–series analysis, and one with one–year data analysis. Here are presented all the results found for this two kinds of data/analysis. However, it is worth to note that the best kind of data/analysis is the time series one. Indeed, this method allows a better comprehension of the characteristics of our networks and their evolution during time.

5.1 Time series results

The analysis of the data collected and the results found allows to infer that many different kind of social structures and characteristics may be found in the community of physicists. Indeed the time series data analysis shows a community like `cond-mat` one that grows during the pediod 2000–2005 as indicated by Figs. 3(a,b), showing 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 ha



(a) (b)

Figure 3: (a) Number of nodes and (b) links per year for the three time series archives `cond-mat` (diamonds), `hep-ex` (squares), `hep-ph` (circles)

.

equal to $N = 9077$ in year 2000, and equal to $N = 15964$ in 2005. Such a behaviour is missing in the other two archives `hep-ex` (nodes) and `hep-ph`. The reason of such a behaviour could be found in the different time scale evolution of these two communities. In fact, high energy physics experiments need a lot of time to be built and then started (the construction of LHC will take more than 6 years !). On the other hand, condensed matter physics needs more "smaller" equipments and less time to build them. It would be interesting to look at the `hep-ex` number of papers and authors after the LCH operations start to see if these conclusions are correct or not.

In this sense, the presence of the same authors in the centrality indexes hit-lists for more than one year shows the same behaviour. For instance, in the `cond-mat` 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 ten, 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 ten nodes with the largest degree: two examples are A. R. Bishop and S. D. Sarma. On the other hand, in `hep-ex` hit-lists it is not possible to find the same author in more than one year except for A. Bodek (two years). A similar behaviour can be noted even in hit-lists containing different names every year. For `hep-ph` the situation is a mixture of the previous two. Is not possible to find the same name in more than one year, but S. Heinemeyer (four years) appears several times in the hit lists. However the case of Heinemeyer is a stand-alone case in a structure that looks like more to `hep-ex` than `cond-mat`. It is worth to note that the value of the efficiency, clustering, mean and max degree increases in the period considered for all of the three time series. Even assortative tendencies are different for the different time series. In `cond-mat`, the degree correlation can be investigated by plotting the average degree of the nearest neighbours 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 ν [34]. In Fig. 5, 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 ν extracted is respectively equal to 0.005 in 2000, and to 0.006 in 2005. In general, ν shows a slow tendence to increase with time in the years considered. The same behaviour is not evident for the other two archives. In fact, even if for `hep-ex` a slightly assortativity tendence is visible in the nearest neighbours average degree $k_{nn}(k)$ given by values

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
1	Eisaki H.	Lee S.	Lee S.	Lee S.	Lee S.	Wang Y.
2	Revcolevscki A.	Kim H.	Kim H.	Sasaki T.	Yamada K.	Lee S.
3	Uchida S.	Thompson J.	Sarrao J.	Canfield P.	Uchida S.	Sarrao J.
4	Ueda Y.	Cava R.	Kim K.	Tajima S.	Ando Y.	Tokura Y.
5	Shen Z.	Revcolevscki A.	Thompson J.	Kim K.	Vedeshwar A:	Lee J.
6	Cheong S.	Tokura Y.	Kim J.	Liu X.	Chen C.	Berger H.
7	Tokura Y.	Wang Y.	Takagi H.	Kim H.	Pfeiffer L.	Gossard A.
8	Fisk Z.	Sarrao J.	Choi E.	Furdyna J.	West K.	Nakatsuji S.
9	Kim Y.	Pagliuso P.	Maeno Y.	Uchida S.	Tokura Y.	Maeno Y.
10	Kim C. & Sarrao J.	Canfield P.	Wang Y.	Takagi H.	Eisaki H.	Pfeiffer L.

Table 5: The ten authors with the highest degree are listed for `cond-mat` archive, in order of rank, for each of the six years considered. The number reported in each cell is the corresponding value of node degree.

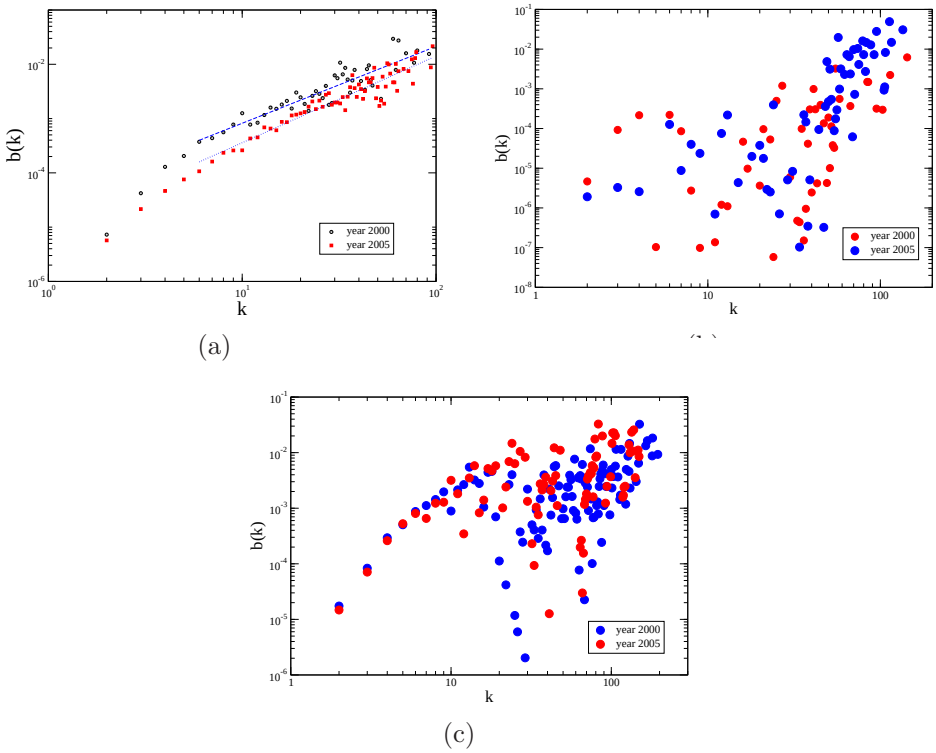


Figure 4: (a) Plots of betweenness degree correlation for `cond-mat`, (b) `hep-ex`, (c) `hep-ph`.

of ν between 0.005 and 0.016, for `hep-ph` there is even a disassortative tendence given by negative values of the slope ν -0.057 found for in 2004 Fig. 5. In both cases for the betweenness correlation it is not possible to find a function that fits data as found in `cond-mat`. This is due to the different structure of the degree distribution in these cases. However plots confirm the presence of a correlation (in particular for `hep-ph`) as is possible to see in Fig. 4.

5.2 One year data results

The results found for one–year data (Tab. 2) show a multitude of different structures and behaviours. Indeed it is possible to find small well–

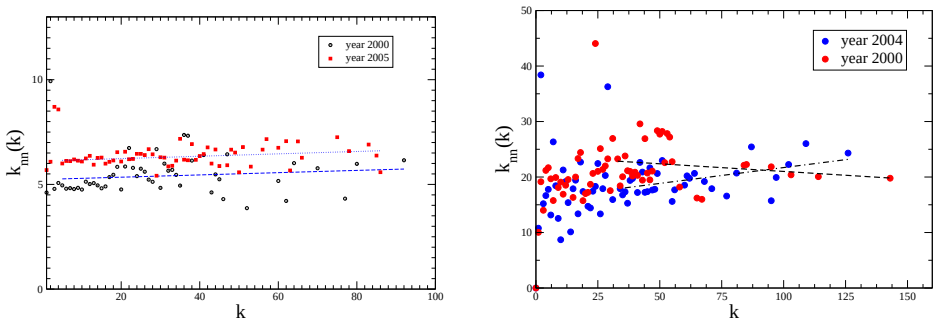


Figure 5: Plots of nearest average degree–degree correlation for (a) `cond-mat`, (b) `hep-ex`, (c) `hep-ph`.

connected sub–communities in archives such as `math-ph` and `gr-qc` (low giant component, high clustering coefficient, low efficiency and characteristc path length). Even large structured communities are found (`quant-ph`, `nucl-ex`) with large giant component, mean and maximum degree and clustering coefficient. These results denote the presence of small well–connected groups that work alone in the first case, and large well–connected and wide–collaborating structures in the second case.

A particular case is physics which shows "strange" characteristics. In fact, the physics archive is one of the biggest communities. Nevertheless, it shows a very fragmented structure as one may note looking at the giant component size, diameter and characteristic path length. Such indexes confirm the presence of a great number of small communities. However,

clustering coefficient, efficiency and average degree indicate that sub–communities are well self-connected. These results can be explained by the fact that the `physics` archive is divided in twenty-two sub–archives ranging from History of Physics, to Space Physics, thus making very hard that such different communities could be well connected between them.

Looking at the hit–lists, it is possible to find key role–playing actors. However, it is not possible to say whether these people play a crucial role in the community only for a year or more, because only one year has been analyzed, so it is not possible to say if the archive's structure looks like either `cond-mat` or `hep-ex` or `hep-ph`. In addition, only in some of these archives it is possible to identify key–role playing actors like C. Vafa in `hep-th`, F. Zimmermann in `physics`, U.G. Meissner in `nucl-th` and V. Vedral in `quant-ph`. Similar results could not be obtained for `math-ph` `gr-qc` and `nucl-ex`. A particular case is that of U. Lombardo (Catania's physics department) who does not have a high degree but appears in all the other centrality hit lists. This fact could be explained in the following way. Professor Lombardo does not have a great number of different collaborators but they belong to different "communities". In fact, Professor Lombardo's collaborators are mainly Chinese and European people, so his function is to join these two different communities. In this sense Professor Lombardo represents a "bridge" between Europe and Asia like the ones over the Bosforo Straits.

	gr-qc	hep-lat	hep-th	math-ph
1	Damour T.	Orginos K.	Vafa C.	Joye A.
2	Buonanno A.	Wenger U.	Cvetic M.	Sigal I.M.
3	Berti E.	Heller U.M.	Jejjala V.	Merkli M.
4	Kokkotas K.D.	Papinutto M.	He Y.H.	Froehlich J.
5	Stergioulas N.	Schierholz G.	Simon J.	Stolz G.
6	Corichi A.	Yamada N.	Balasubramanian V.	Stollmann P.
7	Macias A.	Edwards R.G.	Vandoren S.	Muller P.
8	Quevedo H.	McNeile C.N.	Ahn C.	Schlein B.
9	Cortez J.	Schroers W.	Vazquez Poritz J.F.P.	Klein A.
10	Schnetter E.	Zanotti J.M.	Ross S.F.	Naboko S.

	nucl-ex	nucl-th	physics	quant-ph
1	Brown B.A.	Li B.A.	Zimmermann F.	Vedral V.
2	Berg A.M.	Schwenk A.	Chevallier M.	Eisert J.
3	Steiner M.	Lombardo U.	Raimondi P.	Dowling J.P.
4	Cortina D.	Toki H.	Kim D.	Ralph T.C.
5	Carstoiu F.	Oset E.	Rohe T.	Milburn G.J.
6	Coll. STAR	Meissner U.G.	Cremaldi L.	Pan J.W.
7	Roberts D.A.	Lynch W.G.	Regenfus C.	Zoller P.
8	Magestro D.	Nunes F.M.	Rozet J.P.	Zhang J.
9	Nakamura T.	Meng J.	Smith S.	Zanardi P.
10	Samanta C.	Nogga A.	Rubbia A.	Paunkovic N.

Table 6: First ten authors sorted by their node betweenness for all the one year sub archives.

	2000	**2001**	**2002**	**2003**	**2004**	**2005**
1	Frixione S.	Boos E.	Weiglein G.	Anchordoqui L.	Heinemeyer S.	Heinemeyer S.
2	Seymour M.H.	Roeck A.D.	Heinemeyer S.	Lokhtin I.P.	Roeck A.D.	Kalinowski J.
3	Ilyin V.	Zerwas P.M.	Hinchliffe I.	Takai H.	Balazs C.	Hesselbach S.
4	Mangano M.L.	Bigi I.	Laenen E.	Vogt R.	Djouadi A.	Hurth T.
5	Heinemeyer S.	Battaglia M.	Nason P.	Petreczky P.	Boos E.	Ellis J.
6	Berger E.L.	Miller D.J.	Oleari C.	Niemi H.	Kramer M.	Ali A.
7	Beneke M.	Blumlein J.	Logan H.E.	Nikitenko A.	Godbole R.	Freitas A.
8	Richter Was E.W.	Choi S.Y.	Duca V.D.	Arleo F.	Barklow T.	Abe T.
9	Brock R.	Dittmaier S.	Zeppenfeld D.	Hashimoto S.	Guasch J.	Eberl H.
10	Baur U.	Heuer R.D.	Catani S.	Eskola K.J.	Moretti S.	Kraml S.

Table 7: First ten authors sorted by their degree for the archive `hep-ph` in the period 2000–2005.

6 Conclusions

In conclusion, the characterization of SCN networks using topological properties and centrality indexes allows to identify the inner structure of different physicist communities, thus delineating the main properties of relative networks. The centrality indexes study permits to find the key role–playing actors using an objective criterion. Results allow to say that this is a good method. However, several problems were encountered during data collecting and analysis. In particular, multiple author assignment cause errors in the networks construction and analysis. Maybe in the future the policy in terms of authors identification will improve making it possible to uniquely identify the authors.

In addition, a study of SCN networks in terms of weighted networks may lead to results more similar with the real structure of scientific community. However as discussed above, assigning a weight is not an easy task. Indeed, many different issue have to be taken into account, such as the number of authors for each paper (writing a paper with 3 people is different than writing a paper with other 20 people or more). Even the number of papers written with the same person has to be taken into account, because writing many papers with the same people denote the presence of a strong relationship. On the other hand, a single paper with a person indicate an occasional collaboration.

At any rate, social network analysis is in continuous development and maybe these problems could be solved in future works.

Ringraziamenti

La lista dei ringraziamenti sarebbe davvero lunga (praticamente infinita) ma vorrei ringraziare in modo particolare alcune persone:

- I miei genitori ed in particolare mia madre per quanto hanno fatto per me in tutti questi anni;
- Il Dott. (San) Giuseppe Angilella senza il quale oggi probabilmente starei facendo chissà cosa ...
- Il Dott. Vito Latora per avermi instradato al mondo delle reti e per tutto il supporto che mi ha fornito;
- Salvo Scellato che con me condivide gioie e dolori della ricerca scientifica e senza il quale non avrei potuto scrivere questo lavoro;
- I miei amici per essermi sempre stati vicino anche nei momenti peggiori. In particolare: Angela, Gabriele e Stefano.

A tutti gli altri dico $\mathcal{GRAZIE}$ qualunque cosa abbiate fatto per me!!!!

References

[1] D. J. Watts S. H. Strogatz, Nature (1998).

[2] A.L. Barabási R. Albert, Science **286**, 509 512 (1999).

[3] M. E. J. Newman, Phys. Rev. E **64**, (2001).

[4] M. E. J. Newman, Phys. Rev. E **64**, (2001).

[5] M. E. J. Newman, Proc. Natl. Acad. Soc. **98**, 404 (2001).

[6] M.E.J. Newman, Proc. Natl. Acad. Soc. **101**, 5200 (2004).

[7] Z. Néda E. Ravasz A. Schubert T. Vicsek A.L. Barabási, H. Jeong, Physica A **311**, 590 (2002).

[8] A. Cardillo S. Scellato and V. Latora, Physica A (2006), in press.

[9] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).

[10] James Moody Peter S. Bearman and Katherine Stovel, American Journal of Sociology **110**, (2004).

[11] W. Zachary, Journal of Anthropological Research **33**, 452 473 (1977).

[12] B. Bollobàs, *Random Graphs* (Academic Press, London, 1985).

[13] B. Bollobàs, *Modern Graph Theory*, *Graduate Text in Mathematics* (Springer, New York, 1998).

[14] D. B. West, *Introduction to Graph Theory* (Prentice-Hall, Englewood Cliffs, NJ, 1995).

[15] M. Girvan M. E. J. Newman, Phys. Rev. E **69**, (2004).

[16] V. Latora M. Marchiori, Physica A **285**, (2000).

[17] M. Marchiori V. Latora, Phys. Rev. Lett. **87**, (2001).

[18] M. Granovetter, American J. Sociology **78**, (1973).

[19] R. Pastor-Satorras M. Boguñà, Phys. Rev. E **66**, (2002).

[20] A. Vespignani M. Boguñà, R. Pastor-Satorra, Lect. Notes Phys. **625**, (2003).

[21] M. E. J. Newman, Phys. Rev. Lett. **89**, (2002).

[22] D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, Princeton, NJ, 1999).

[23] M. Marchiori V. Latora, Eur. Phys. J. B **32**, (2003).

[24] C. L. Freeman, Sociometry **40**, (1977).

[25] C. L. Freeman, Social Networks **1**, (1979).

[26] B. Wellman, J. Educ. Res. **14**, (1926).

[27] S. Leinhardt P. W. Holland, Amer. J. Sociol. **76**, (1970).

[28] A. Bavelas, Human Organization **7**, (1948).

[29] P. Bonacich, Amer. J. Sociol. **92**, (1987).

[30] M. Zelen K. Stephenson, Social Networks **11**, (1989).

[31] `http://www.nist.gov/dads/HTML/assocarray.html`.

[32] T. H. Cormen C. E. Leierson R. L. Rivest and C. Stein, *Introduction to Algorithms* (MIT University Press, Cambridge, 2001).

[33] V. Latora P. Crucitti and S.Porta, Phys. Rev. E **73**, (2006).

[34] R. Pastor-Satorras A. Vázquez and A. Vespignani, Phys. Rev. Lett. **87**, (2001).