"1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011)"
in Collabaration between EPOKA University and "Aleksandër Moisiu" University of Durrës
on June 2-4 2011, Tirana-Durres, ALBANIA

Real networks; Albanian Scientific Collaboration Network

Eva JANI[1], Eglantina XHAJA[1]

[1]Faculty of Natural Sciences, University of Tirana, *Tirana-ALBANIA*
Email: eva_jani@yahoo.com
Email: eglaxhaja@yahoo.com

Abstract

We have realized that we reside in a world of networks. The Internet and World Wide Web (WWW) are changing our lives. Our physical existence is based on various biological networks. The sophisticated tools of Network Theory have made it possible to quantify human dynamics, the relationships between millions of individuals via the analysis of the Social Networks. "Scale-Free" is the main property which differentiates a lot of networks from the others, real or simulated, meaning that the vertex degree distribution of such networks follows a power low. The Internet is one of the best known scale-free networks.

We study the Albanian Scientific Collaboration Network (ASCNet), which is an undirected graph, where the vertices represent the scientists and each pairs of them are adjacent if the corresponding scientists have coauthored a paper. The data used is taken from the bulletins of the Faculty of Natural Sciences and Aktet of AlbShkenca Institute, in a span 2004-2010 and 2008-2010 respectively. We analyze the data and demonstrate the differences in the patterns of collaboration for various research fields. We argue that the ASCNet is a Scale-Free network with a slope $\gamma = 3.7$. We also find the "small world effect" in our network, and the clustering property.

I. Introduction

As consequence of the rapidly computer sciences development it's made it possible the study and structural organization analysis of various and increasingly large-scale networks. In the last decade are studied numerous real networks, starting from biological (cellular metabolism, genetic regulatory, food webs), social (movies actors, collaboration networks, e-mail, mobile calls, twitter, facebook), industrial, transportation, business, neural networks, information networks (citations, WWW) and arising to the Internet (router level graphs and AS-level graphs) which are the largest real networks with billions nodes and links [1].

Despite of such extremely differences in their function and attributes, mathematicians of networks, based on a deep connection with statistical physics approach, in attempts to build general representative models of real networks, have

provided the emergence of several common features, but also the distinctions on specific details of their individual systems.

The networks' degree distribution serves as an initial differentiation of real networks. The degree (k) of a vertex is the number of edges connected to it. The vertex degree distribution is the probability $P(k)$ that a randomly chosen vertex in the network has degree k. Many real networks have power law degree distribution: $P(k) \approx k^{-\gamma}$, $k \neq 0$, where γ is the exponent of the power law. These networks are called scale-free. The degrees of their vertices are highly right-skewed, meaning that their distribution has a long right tail of values that are far above the mean.

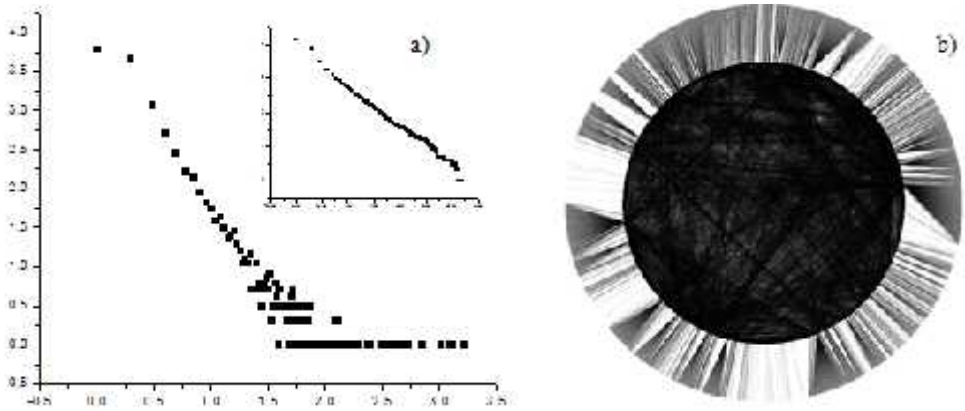


Figure 1 a) Scatter plot in log-log scales of the degree distribution of the Internet' network. Inset: Cumulative distribution in log-log scale. b) Visualization of the Internet's network.

This is the main property of scale-free networks. The Internet is one of the best known networks with such a property. Based on a registration of NETDIMES done in October 2004 (chosen randomly) in [2] is argued that the Internet exhibits a power low degree distribution with a slope $\gamma \approx 2.000$. The visualization of this network is given in Figure 1/b), while a scatter plot on log-log scale of its vertex degree distribution is presented in Figure 1/a). On first sight in the plot, it is easily distinguished much noise on the right tail, which in mathematical terms is translated into a very large level of fluctuations. The slope is founded by selecting by eye a region with a power-law behavior and applying the Least Squares fitting to all points in the region. This technique usually adopted on finding the slope γ still remains empirical; since the fluctuations in the high degree region of the distribution hinder accurate fitting.

An improved way to make the fluctuations less pronounced is the cumulative degree distribution plot [1, 3]. The cumulative distribution function:

$$P_{cum}(k) = \sum_{k'=k}^{\infty} P(k'),$$

is the probability that the vertex degree is greater than or equal to k , and also follows a power low with exponent γ - 1. The cumulative distribution plot on logarithmic scales, besides the avoidance of the noise in the tail, has also the advantage that all the original data is represented. The slope of the power low degree distribution founded for the NETDIMES registration is very accurate, thanks to the large size of the Internet's network, but in practice one rarely has enough measurements to get a good statistics in the tail, so in these cases each data has very much importance on getting the correct results.

The latter considerations have been very useful on the starting point of our work presented by this paper.

The Collaboration Networks have been the object of study in many papers [4, 5]. As members of the researchers in Natural Sciences, we were excited from the idea of investigating the Albanian Scientists Collaborations via the network theory.

Initially, we were faced with some difficulties. As far as we are aware, an electronic database for getting full information about our scientists' publications does not exist. So we took the bulletins of the Faculty of Natural Sciences (9 bulletins, in a span from 2004-2010) and Aktet of AlbShkenca Institute (4 journals in a span from 2008-2010). From Aktet we have selected only the papers from Biology, Chemistry, Environment, Mathematics, Physics, and Computer Science.

We unify both databases, and we construct the network, named Albanian Scientific Collaboration Network (ASCNet). Through this unification, we incorporate the information for each author during the period at issue, presuming a more complete coverage over his publications and collaborators. Such a network would also allow us to inquire how much present is the interdisciplinary collaboration, considering particularly the researchers which develop their activity in the same environment, our faculty, for instance.

II. Basic statistics of the data

A summary of the basic statistics taken from the examined data is given in Table 1. All the papers from both databases are divided in two groups; BCHE (Biology-Chemistry-Environment) and MPHCS (Mathematics-Physics-Computer Science). We must emphasize that this division in two groups is just based on the discipline that each paper covers, despite the fact that the coauthors may belong to different disciplines. For example, in the row "total authors", if one person has published papers on disciplines from different groups, then he/she appears counted in each of them.

Table 1 A summary of statistical results of both Databases

	Total		FNS		AlbShkenca	
	BCHE	MPHCS	BCHE	MPHCS	BCHE	MPHCS
Total papers	218	80	157	51	61	29
total authors	360	119	242	60	164	68
mean papers per author	1.81	1.46	1.81	1.53	1.38	1.31
mean authors per paper	3.06	2.26	2.8	1.8	3.75	3.07
maximal number of authors per paper	10	7	7	5	10	7
maximal number of papers per author	14	7	14	7	6	5
single-authored papers (%)	11.0	27.5	13.5	39.2	4.9	6.9
single authors (%)	2.77	7.5				

The papers of the BCHE group constitute 73.2% of the total one; hence it is obviously clear to be expected that this group will play the most important role in the structure and characteristics of our ASCNet. In addition, this group has the higher mean authors per paper and the lower number of single-authored papers compared with the MPHCS. In the database we are considering, if an author has not any publication with co authorship, he/she won't be represented in the ASCNet. The number of these authors in percentage given in the last row of Table 1 is significantly higher for the MPHCS group.

The values in the rows "mean papers per author" and "mean authors per paper" do not differ much from each-other. To provide more information, we have plotted the distribution of the number of papers per author, and the number of co authors per paper for each group respectively (Figure 2/a), b)), using the unified data.

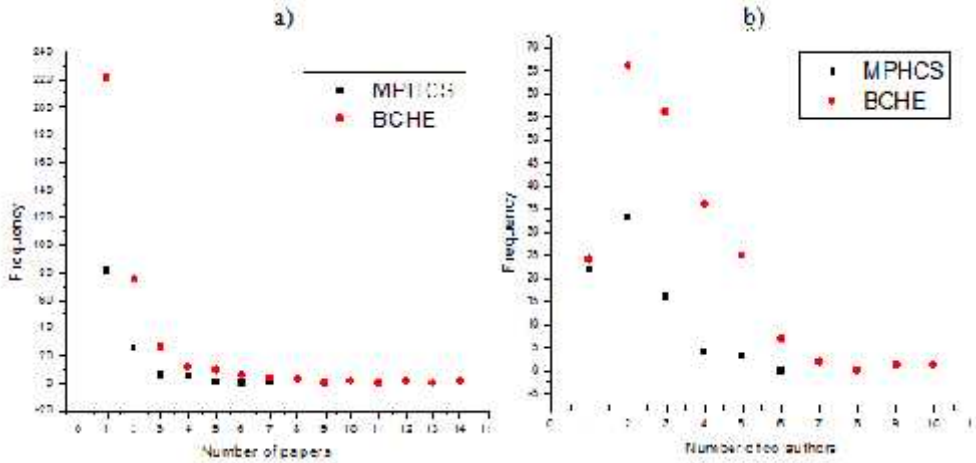


Figure 2 a) Scatter plot of the number of papers per author for each group.
b) Scatter plot of the number of co authors per paper for each group.

The figures are same for both groups, meaning that the Albanian scientists frequently work in groups with two or three members. Although, we pick out that

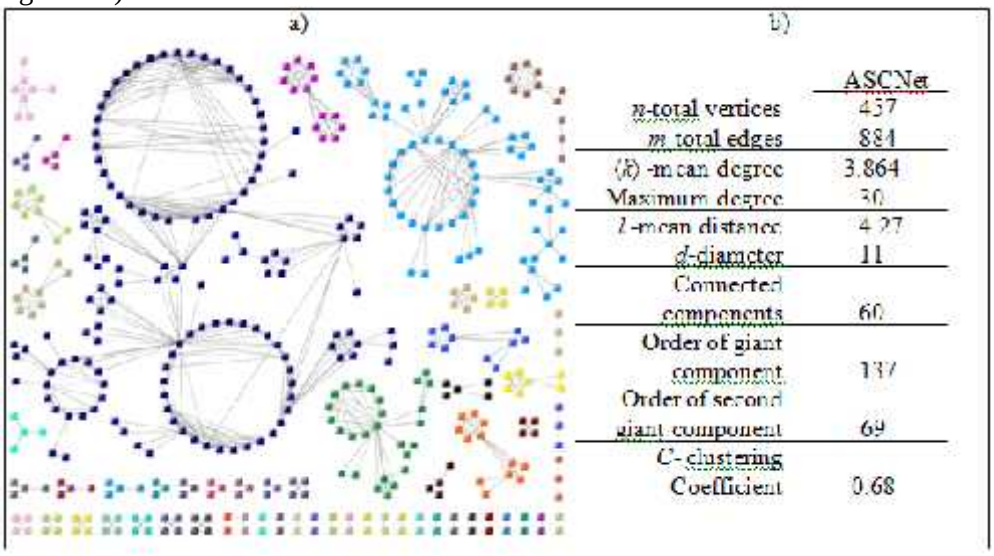
33% of the papers in BCHE group, have four or more co authors, whereas only 11.25% of the papers in MPHCS group are with more than three co authors.

We notice in the plots a high number of authors in BCHE group, which have published just one paper (61.7 %). The number of papers with three or more authors for this group, is high also (58.7 %). There is a tie between these figures. Many collaborators in a single paper, usually are founded on the papers based on experimental works, thus not only researchers, but also experimentalists, individuals from institutions with laboratories and useful data, contribute on the realization of a paper. On the other side, these individuals or most of them are not much interested to publish frequently papers. It is founded also a considerable number of foreign authors, most of them have collaborated for only one paper. All these authors do not determine the average characteristics of their group, despite the vertices, which represent them, may have high degree.

If we would neglect the authors, which have published just one paper, the others, in the BCHE group, have written averagely 3 papers.

III. The construction and analysis of the ASCNet

A Collaboration Network is an undirected graph, where each vertex represents a paper's author and there exists an edge between two given vertices if and only if the corresponding two persons have coauthored at least a paper. The ASCNet constructed, based on this definition and the available data, is shown in Figure 3/a).



	ASCNet
n-total vertices	437
m total edges	884
⟨*k*⟩ -mean degree	3.864
Maximum degree	30
l -mean distance	4.27
d-diameter	11
Connected components	60
Order of giant component	137
Order of second giant component	69
C-clustering Coefficient	0.68

Figure 3 a) The visualization of the ASCNet.
b) The quantitative measures of the ASCNet.

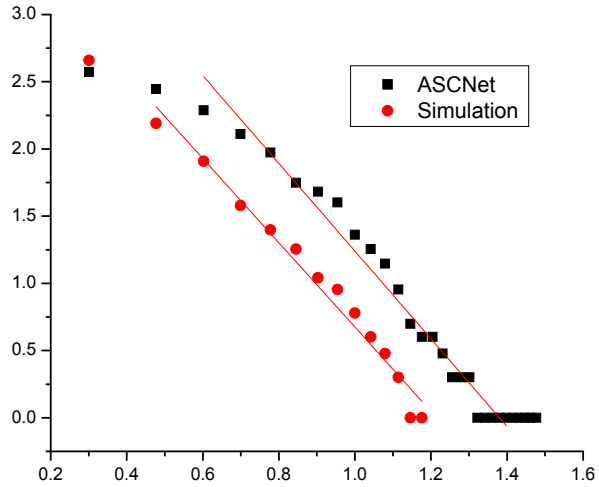


Figure 4 Cumulative degree distribution in logarithmic scale of both networks;
ASCNet (black dots), Simulated Network (red dots), and their linear fittings.

Making the cumulative degree distribution plot of the ASCNet, shown in Figure 4, we find that the ASCNet is scale-free.

The slope γ is founded using the technique described in Introduction, and the result is γ=3.7. To be convinced for the correctness of our result, we simulate also a scale free network with the same γ, its cumulative degree distribution is plotted in the same graph with the ASCNet.

To gain further insight on the structure properties of the ASCNet, let us analyze the quantitative measures presented in the table of the Figure 3/b).

The mean degree in our terms is the average total number of individuals with whom a scientist has collaborated. As is seen, each author has collaborated averagely with 3-4 others, during the period of study. An author has maximally collaborated with 30 individuals.

The Collaboration Networks may reasonably consider as acquaintance networks. If two persons have coauthored a paper, then they know each other. A feature that characterized the acquaintance networks is the low mean distance between their vertices. This is called the "small world effect" or as is known with the phrase "Six degree of separation" [6].

The distance or "geodesic" between two vertices in a network is the shortest path length between them. The small world effect describes the fact that despite the large size of the network, there is relatively a short path between any pair of vertices. An estimation of the presence of the small world effect in a network is the mean distance l for all vertices' pairs. [7] In our ASCNet the l = 4.27, which means that the "small world effect" is evident. Notice that, our network is not connected,

thus the quantity l is calculating only for the pairs of vertices, which have a connecting path. The vertex-vertex distance in the context of Collaboration Networks gives the rate of the information transition, from one to another scientist, such as new theories, experimental results and possibilities of establishing new contacts for future collaborations. In our ASCNet, the information started from an individual has to pass through averagely three persons to arrive to everyone else in the community. This estimation is more valuable for the Giant Component, because of the fragmentation of our network in many small components, which is treated in the following.

The network transitivity or also called clustering is a property that differs in various networks. If a given vertex A is adjacent to vertices B and C, then the clustering coefficient C gives the probability that B and C to be adjacent [8]. In our ASCNet, the clustering coefficient C is very high (68%), that is the average fraction of pairs of authors, which have collaborated with a common individual and with each other also. This coefficient measured only for the vertices, which belong to the Giant Component, is higher (76%). This may be the consequence of the fact that a considerable number of authors come from the Faculty of Natural Science, so they work and experimentalize together, and they know well each-other, enabling them to collaborate in their researches and to publish their common works.

As we mentioned above, it is easily distinct in Figure 3/a) the presence of small groups of vertices that have a high density of edges within them, and a lower density of edges between groups. This means that our network show a "community structure". It is a matter of common experience that researchers do divide into groups along lines of interest, job's environments, and friendships too. But on the other side this occurrence indicates a poor intense of collaborations even also between researchers within the same discipline. If we look at the data that belongs to mathematics area, we find only 39 papers, 20 of them with single author. The Collaboration network of mathematicians, it would be composed of 12 connected components with at most 4 vertices each of them. Furthermore the mathematicians are not connected at all with others. Let us evolve some reasons of this situation:

Firstly, we also find a similar occurrence in the networks studied from Newman in [4] from databases that cover theoretic fields, like Physics and Computer Science. To the contrary, the experimental databases (in Biomedicine, condensed matter physics, and astrophysics) show a large number (hundreds or thousands) of collaborators in a single paper. This was the reason we divided our data in two groups. The basic statistics given in Table 1, would be approximately equal for each discipline within its group.

Secondly, we must emphasize that the two studied journals are not the only ones where Albanian researchers publish their works. This may be another reason that the mean number of papers per author is low.

Usually, most of the publication in Biology, Environment and Chemistry are based on experimental researches; therefore this group in ASCNet reflects a more collaborative community, more members, and much productivity estimated by the number of publications. The giant component (vertices with dark blue color in Figure 3/a)) consists of collaborations only in chemistry area, 30% of the total

vertices. This low percentage of the number of vertices in the giant component is because of the multi-disciplinary nature of our network.

The second giant component (vertices with light blue color in Figure 3/b) presents a special interest. It covers a variety of disciplines, Biology, Environment, Physics and Computer Science and a numerous collaborators from different areas which we just mentioned. 90% of researchers in this component come from the Faculty of Natural Science. Certainly, this is a good thing, although the spread of such collaborations is located on a low percentage of the total number of authors.

Recently, the scientific research fields have become so fluid. Individual researches are substituted from the collaboration teams which are becoming a prominent means on high quality production. Multi-authored papers are more frequently cited than single-authored publications. For a better productivity performance, it is necessary that also our researches to enable themselves the assembly in research teams. Using the network analysis, Guimerà at [9] founded out that large teams with teammates from various disciplines, the combination of the significant presence of long experienced researchers with newcomers with fresh ideas, and lower tendency to "over-repeat" collaborations, are the required factors to build successful teams which impact directly on quality production. See also [10].

IV. Conclusions

We have studied the Albanian Scientific Collaboration Network, using the database, which integrates the publications of the Bulletins of Faculty of Natural Sciences and Aktet of AlbShkenca Institute, in a span 2004-2010 and 2008-2010, respectively.

We have analyzed the basic statistics of the data, divided in two groups, BCHE, and MPHCS, comparing the respective measurements. The authors of the BCHE group publish more papers, often with larger groups of collaborators per single paper; they are enable to create a larger community.

We have constructed the ASCNet, and by the cumulative degree distribution, we argue that ASCNet is a Scale Free network with a slope $\gamma = 3.7$. Simulating a scale free network with the same γ, we have proved that the slope founded is correct.

In the ASCNet is present the "small world effect". The information started from an individual has to pass through averagely three persons to arrive to everyone else in the community.

ASCNet has the property of clustering, meaning that two scientists with a common collaborator, have much higher probability to have collaborated, than two others chosen randomly from the Scientific Community.

We have founded the presence of various interdisciplinary Collaborations, in Biology, Physics, and Computer Science. These collaborations don't have a wide spread to all the Natural Sciences' disciplines; their frequency also is not in the accordance to the demands of the evolution and development dynamics that the sciences are asking for.

The recipe is: Self-assembly in large research groups, involving individuals from various disciplines, experienced scientists and newcomers with fresh ideas.

References

[1] Dorogovstev, S. N. Mendes, J. F. F. (2003) Evolution of Networks. From Biological Nets to the Internet and WWW. OXFORD UNIVERSITY PRESS.

[2] Hoxha, F. Xhaja, E. (2009) Computer Simulations of dynamic processes in Complex systems. BSHN (UT), 7.

[3] Li, L., Alderson, D. Doyle, J. C. Willinger, W (2005) Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications, Internet Mathematics Vol.2, No. 4: 431-523.

[4] Newman, M. E. J. (2001) The Structure of Scientific Collaboration Networks, Proc. Natl. Acad. Sci. USA 98, 404-409.

[5] Newman, M. E. J. (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Physical Review E, Volume 64 (016132).

[6] Albert, R. Barabàsi, A. (2002) Statistical mechanics of complex networks. Department of Physics. University of Notre Dame. Notre Dame. Indiana 46556.

[7] Costa, L. da F. Rodrigues, F. A. Travieso, G. Villas Boas, P. R. (2006) Characterization of Complex Networks: A Survey of measurements, arxiv:cond-mat/0505185v5 [cond-mat.dis-nn],.

[8] Newman, M. E. J. (2003) The Structure Function of Complex Networks. Society for Industrial and Applied Mathematics, SIAM REVIEW, Vol.45, No. 2, pp. 167-256.

[9] Guimerà, R. Uzzi, B. Spiro, J. Nunes Amaral, L. A. (2005) Team Assembly Mechanism Determine Collaboration Network Structure and Team Performance. SCIENCE vol. 308, 697-702.

[10] Stvilia, B. Worrall, A. Kazmer, M. M. Hinnant, C. C. Burnett, G. Schindler K. Burnett, K. Marty, P. F. (2010) Composition of Scientific Teams and Publication Productivity, Proceedings of the American Society for Information Science and Technology, Volume 47, Issue 1, pages 1-2.