Network Structure and Emergent Collaboration in a Research Network

Molka-Danielsen, Judith; and Søvik, Bernt Louis Berge

Abstract— ***Clusters of associations and even friendships can foster a sharing of ideas that lead to the co-production of scholarly works. This paper studies the emergence of a network of research collaborations within one small academic institution. We study which of these associations are beneficial to research production and what is the relationship to the structural connectivity of the network. The nodes of this social network are 92 researchers at Molde University College in Norway, and the links are their co-author associations that produced 1480 unique scholarly works. The source of data is the ForskDok database of documented research in Norway that contains over 160,000 publications. We examine global network characteristics: the average degree distribution, and the diameter of network collaborations. We study network neighborhood characteristics: a comparison of percentage of collaboration takes place within and between faculties and that which is external to the school. From these findings we discuss the emerging collaboration and conclude whether faculty members that collaborate have higher quantities of publications. Finally, based on our study of this network data, we discuss the meaning of collaborative efforts for researchers and their institutions.***

Index Terms—*Collaborative networks, degree distribution, network diameter, and productivity.*

1. INTRODUCTION

SOCIAL networks are information networks of humans interacting and creating relationships. A link between two persons in the network can be a close friendship, partnership, or working relationship. There are many studies that examine the structural and dynamic network characteristics of such social networks, and in particular of emerging researcher networks.[4][5][12][14] This study also examines a network of researchers, in particular, the collaborative co-authorship network at Molde University College in Norway. We explore the associations between network characteristics, the benefits of collaboration and the quantities of publications. While "the network of researchers" is becoming an important resource also to small educational institutions, a problem for these is they usually to not have the established researcher networks of larger institutions. Other studies support the need for this research. In particular, a study on betweeness centrality correlation in social networks has shown that "each person in surrounded by almost the same influential environments of people no matter how influential the person may be."[10] This indicates that improving the collaborative environment can assist all member of that environment. We are therefore motivated to discover how the connectivity of the network emerges; because it can assist in the network's further evolution and indicate how to improve the potential productivity of the institution.

2. BACKGROUND

2.2 Prior Research of Information Networks

Recent studies have identified a "small world" phenomenon in social networks.[17][18] This phenomenon states that average distances across networks can be very small, that one must traverse only a few links on average to go between any two nodes of the network.[1] That is all of the members of the network are on average close to all other nodes of the network. Why do some networks exhibit this phenomenon? Small world networks have been characterized to have a few members (or nodes) with many connections to other members. The popular members are called "hubs". We see hubs in technology networks also, such as routers on the Internet. The routers (nodes) are connected by transmission lines (links). But, hubs exist also in certain social networks, such as networks of researchers connected by collaborative works. Is the researcher network at Molde University College a "small world" network? The data here will support that MCs network shares some of these characteristics. The characteristics that we study are: the average degree distribution of the nodes, the "betweenness" positioning of nodes, the clustering coefficient of node neighborhoods, and the direction of collaboration. Finally, we will look at what these characteristics mean to the researchers.

3. RESEARCH METHOD

3.1 Data

The Data source for the experiment was the ForskDok database publicly available from the library pages at the Molde University College: (http://www.himolde.no). ForskDok defines itself: "ForskDok consists of two (2) databases, FORSKPRO and FORSKPUB. In addition is a

register of Norwegian research institutions, a register for topic disciplines and a person register. FORSKPRO consist information about more than 5000 ongoing or finished FoU-projects. FORSKPUB consists of information about more than 160,000 publications and other results from FoU-works."

From the ForskDok database we obtained a listing of research works given constraints from input specifications. The listing for this experiment was requested on the 19th of January, 2005. The query response consisted of 1781 publications for college faculty researcher consisting of 1480 unique publications.[4]

Open source Java Universal Network/Graph (JUNG) framework for analyzing networks and graphs was implemented to produce the visualization of the network data results. In particular, JUNG was used in creating the network map and for collecting data about; the diameter of the network, degree distribution and betweeness values.

3.2 Network Structural Characteristics

In the network studied authors (researchers) are defined as nodes and links are created between authors in the network when they collaborate in a publication. Links are not redundant in that there can only exist one link between two (2) unique authors. A "unique co-author" link is defined as the first occurrence of two authors working together on a publication. For example, if author A works with author B on publication-2001 and publication-2002, there is only one unique link established. If new author C also works on publication-2002, then two new unique links are established A-C and B-C.

Since some of the analysis was to be based on authors from the staff at Molde College (MC) it was necessary to be able to identify the nodes as either internal or external. Internal Links (IL) are co-authorships between two persons that are both faculty within one department at MC. Between Links (BL) are between 2 MC faculty from different departments. External Links (EL) are between persons where one person is not on faculty of the college. There were 92 internal researchers. Links are discussed later and listed in Table 5. It was also an objective to be able to study the evolution over different time periods, so identification of the age/date of the publication/collaboration link was collected.

[4] Table 1 has the number of publications or scholarly works that the researchers of the various departments have been a part of. This total of 1781 is different from 1480 because authors A and B can be from two faculties and have worked on the same publication P. The count of publications per researcher is therefore different than the total number of unique publications (1480). The number of co-authorships is also greater than 301 (=1781-1480) because some coauthors are not on faculty within the college.

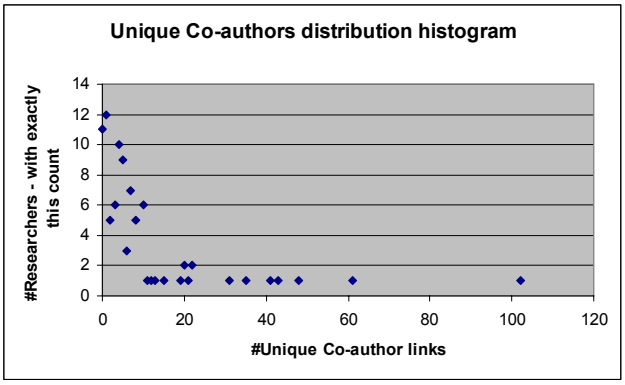

Fig.1. Unique Coauthor distribution histogram

TABLE 1
PUBLICATIONS AND CO-AUTHORSHIP

Faculty	#Pubs for researchers	#Unique Co-Authors	#Total Co-Authors
Health Sc.	194	80	142
Informatics	324	165	420
Social Sc.	271	152	326
Logistics	517	223	660
Economics	475	217	618
Sum	1781	837	2166
Average	19.36	9.10	23.54
Std. Dev.	35.8	14.71	49.33

4. FINDINGS

Table 1 summarizes the data by faculty of the total number of publications per researcher, the number of unique coauthors and the total number of coauthor collaborations. In the dataset the average number of publications per researchers is 19.36. The average number of unique associations per researcher is 9.1. The average total sum of unique coauthors that a researcher would have ever worked with (on all accumulated works) is 23.54. The average number of co-authors on one given scholarly work is 1.21.

One objective of this project was to find out whether the data summarized in Table 1 would in more detail characterize the researcher network at MC as a "small world" network. In summary we find it can be called a "scale free" network with some characteristics of a "small world" network. We demonstrate in more detail in Figure 1 and explain herein.

4.1 Degree Distribution

In Figure 1 the number of unique person-to-person links follows an exponential distribution or power law. In other words, there are a few authors with many different collaborators. This degree distribution follows the shape of a "scale

free network" rather than a "random network." The degree distribution of random networks follows a bell shaped curve. So, if one plotted the number of nodes with links versus the number of unique links there would be a bell shaped distribution. That would mean that most nodes have the same number of links and that nodes with very large number of links do not exist. Random network theory was introduced by Paul Erdős and Alfréd Rényi in 1959. [7]

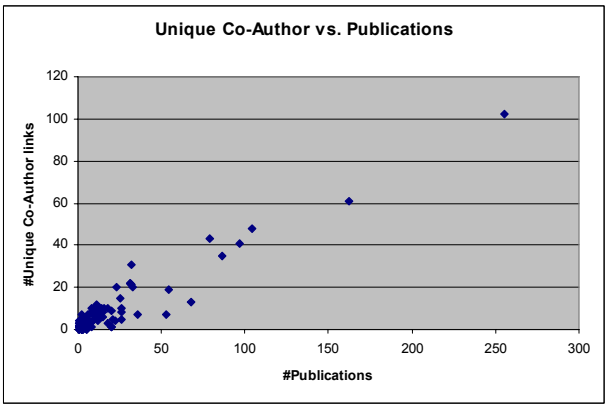

Fig.2. Unique Coauthor vs. Publications Count

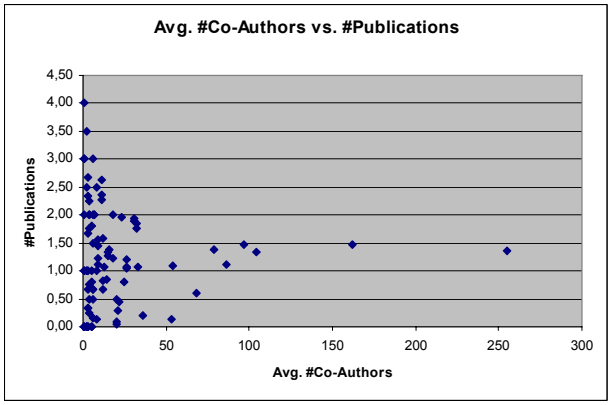

Fig.3. Average #Coauthors vs. Publications Count

Another kind of network, described by Barabási and colleagues, is called a scale free network. [3] This type of network follows a power law distribution such that most nodes have only a few links, while a few nodes have many links. The nodes with many links are called hubs. An example of this is an air traffic system where many small airports are connected by a few hubs. Social networks have also been found to follow a scale free distribution. For example, a study was made by Malcolm Gladwell to measure how social a person is. [9] His conclusion was there are a few people that have a knack for making friends and acquaintances and these are connectors. Similarly, our network of researchers appears to contain a few connectors, as is seen

TABLE 2
BETWEENESS OF NODES

Faculty	Betweeness	
	Rank	Score
Health Sc.	124.88	1643
Informatics	90.41	4032
Social Sc.	71.07	4270
Logistics	36.67	16102
Economics	87.76	3865

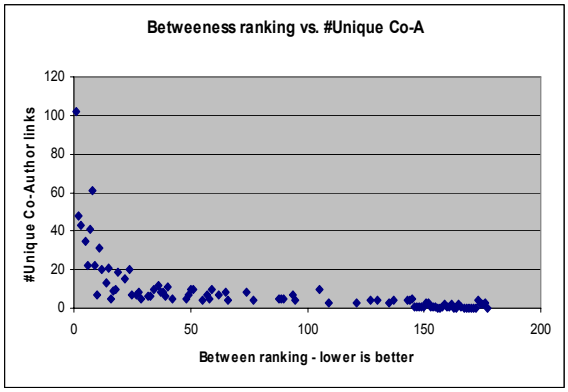

Fig. 4 Betweeness ranking vs. #Unique Co-Authors

in Figure 1 and summarized in Table 1, the standard deviation from the average number of unique coauthors is broad.

Another characterization of networks is that they change or evolve over time. [4] Presently, the Molde College researcher network has some demonstrable "hubs" of researchers with many connections. But, the number and size of hubs can continue to increase over time. Similarly, Albert and Barabási study of the Internet and the www network of web pages showed that though one might expect that there would be only a few hub web pages with very many links in and out, that was not the case. Rather, they found the hubs on the web are not so rare, and that new hubs can come into existence even while other established hubs already exist. Also although the number of hubs increased, that the network still followed a power law and that growing networks often follow a power law distribution.[2]

In our network, we may also have researchers that will change into connectors or hubs over time. That is, they may produce more publications and create more connections (new associations) over time. The creation of hubs usually arises out a differentiating factor. For example, a researcher with a research grant might make that researcher a popular coauthor. This phenomenon has been called "preferential attachment."[3] Preferential attachment explains that any researcher would prefer to work with a more famous, established, or experienced researcher than with an unknown. The tendency, also known as "the rich get richer", explains the growth of hubs in some networks. Others from

1955 until the present confirm that "growing networks are self-organized into scale-free structures because popularity is attractive."[6][16] Causality can be bidirectional, as will be discussed further later.

Is there an association between the amount of collaboration and the number of works produced by researchers of the studied network at Molde College? Figure 2 shows that working with a high number of unique coauthors are associated with a high number of publications. Figure 3 shows that the average number of coauthors per article is not necessarily associated with a high total number of works produced. We may conclude from these data that it helps to have many associations, but it does not help to have many names on one article. The data in these figures are based on the data collected from ForskDok and summarized in Table 1.

TABLE 3
PUBLICATIONS GROWTH

Period	#Publications added	#Publications accumulated
1991-93	35	35
1993-95	80	115
1995-97	164	279
1997-99	374	653
1999-2001	250	903
2001-03	378	1281
2003-01/ 2005	199	1480

TABLE 4
DIAMETER AND CONNECTIVITY

Year From 1991 - to 12/31/yy	Diameter of network	# Connected authors	#Links in network	# Clusters
1997	8	148	333	16
1999	10	286	674	12
2000	10	339	817	14
2001	10	375	889	14
2002	10	416	981	15
2003	10	444	1059	17
2004	12	462	1119	16
2005	12	494	1214	15
2006	12	494	1214	15

TABLE 5
EXTERNAL LINKS TO INTERNAL LINKS RATIO

Faculty (FAC)	#IL within FAC	#BL between FAC	#EL out of MC	B-to-I ratio	E-to-I ratio
Infor.	19	14	113	.74	5.95
Econ.	20	35	139	1.75	6.95
Health	12	17	39	1.42	3.25
Social Sc	19	18	94	.95	4.95
Logistics	6	28	182	4.7	30.33
MC total	132	←	567		4.30

4.2 Betweeness

Table 2 and Figure 4 tell us something about how the collaboration network is connected. It follows the shape of exponential distribution but important to note is the nodes not in the direct path of the shape. These are nodes that most likely connect larger clusters together making them a preferable location to traverse from one cluster to another. [3] This could be perceived as the strength in weak links which relates to the early studies of Mark Granovetter who as a graduate student at Harvard University tried to find out how people "network" to get new jobs. [11] The authors/nodes in the network with the "best" ranking can be called connectors or hubs since they often are connecting clusters together. These connectors or hubs can be observed in the visualization of the complete network in Figure 6.

4.3 Degree Exponent

As stated previously, the link distribution in a scale free network follows a power law.[15] That is the probability of a node being connected to k other nodes can be expressed as $P(k) = C * k^{-Y}$ where Y is the degree exponent. The degree exponent for our network is .8 and is shown in Figure 5. We found this to be lower than expected for social networks in general. Previous research by Barabási found higher degree exponents in social networks of scientific collaboration. They found it was Y=2.1 in a Natural Science collection, and Y=2.4 in a Mathematics collection. Their collections spanned 8 years. Although these exponents are considerably higher than ours, it should be noted that the number of researchers in the prior studies are also considerably larger. The mathematics database contained 70,975 different authors and 70,901 papers. The natural sciences database contained 209,293 different authors and 210,750 published papers. [4]

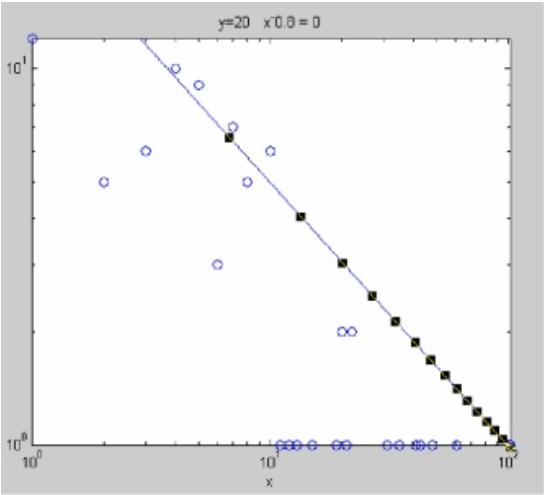

Fig.5. Degree Exponent

In summary, our network had too few participants to represent the end of the scale that is supposed to be rare occurrences. Of the 92 researchers, only 66 researchers were part of the connected network (had coauthors). A network of more researchers is needed to map the degree exponent. Also, and importantly, the former studies concentrated on specific fields of science while our data was cross-disciplinary and there would seemingly be less collaboration across fields. Last, the degree exponent can be different from that of a larger institution because the quantity of research activity or unique collaborative ties established by member (nodes) of one small college may be far less than that of a larger.

4.4 Analysis of network evolution

The growth of the network is depicted in Table 3 and Table 4. The registration of publications in ForskDok started in 1991. As links are added to our network the number of separate clusters reduces and the size (number of nodes) of the largest cluster increases. The diameter of a network is the "longest shortest path" or the largest number of links that must be traversed to get from any node to any other node in the network using the shortest path.[5] Our program computes the diameter and also the average of "longest shortest paths" that is the average distance. Peterman distinguishes between the definition of "small world" and "scale free" networks. He says "small world" implies that the average distance between nodes of the network increases at most logarithmically with the number of nodes.[6]

[5] One estimate for the diameter of the web is D=2log(N) where N is the number of web pages and 2 is an estimate of the inward degree exponent.

[6] Peterman defines: Scale-free refers to the lack of an intrinsic scale in some of the properties of the network. In particular, degree (or connectivity) distribution: the degree k of a node is the number of other nodes it has links to and the degree distribution P(k) is simply the histogram of the number of nodes with a given degree k. Scale-free networks exhibit a power-law behavior of the distribution $P(k) \sim k^{-Y}$, with Y values often between 2 and 3." [15]

TABLE 6
CLUSTER COEFFICIENT OF THE CONNECTED NETWORK

Faculty	# IL	# Nodes	# Possible Links	CC	CC Random
Infor.	19	20	190	.1000	.0475
Econ.	20	14	91	.2198	.1020
Health	12	10	45	.2667	.1200
Social Sc	19	10	45	.4222	.1900
Logistics	6	6	15	.4000	.1667
MC total	76	66	2145	.0615	.0303

In our network the average distance between any two nodes is 4.4 links in 2005 and it was the same in 1999. The growth is too small to draw conclusions. Similarly, if the average distance increases with a logarithmic law, we would also expect the diameter of the network will increase slowly and it does as shown in Table 4.

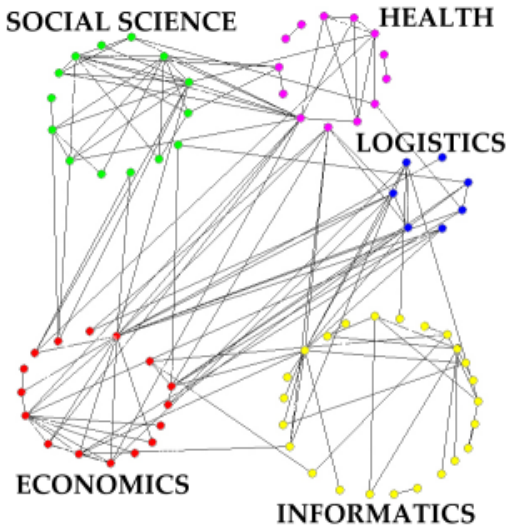

Fig. 6. Molde College Researchers Network Map

4.5 Where the links are established

Some communities are sometimes identified by the number of links within a group. For example, this method was used to identify communities on the web. The study by Flake, Lawrence and Giles, identified "documents belong to the same community if they have more links to each other than to documents outside the community." [8]

But, researchers it is hoped do not work dominantly within their own department. We examined collaboration by department and confirmed a greater establishment of links outside MC. In Table 5 is the ratio of external links to internal links for all faculty departments is greater than 1. To compare the extent of collaboration within departments, we can look at

the clustering coefficient (CC) of the connected network of researchers. The connected network is the group researchers that have collaborated with someone. The "clustering coefficient" is the number of existing links divided by the number of possible links within a grouping. [13][17] In Table 6 we see that the Social Sciences and Logistics departments are the most collaborative among researchers within their department and Informatics is least collaborative. Figure 6 is a network map of the links within and between departments. The node set in the figure includes the group of researchers with at least one coauthor link.

However, even though the Logistics faculty collaborates within, they also have a high ratio for establishing external contacts. We can also ask if a greater ratio of external associations contributes to a higher number of publications. A study by Newman found the probability of acquiring new collaborations increases with the number of past collaborations. [14] We can look at the association of the publication count to past collaborations. Table 5 shows that while all departments are collaborating more often externally, with coauthors in an outward direction rather than within the school, the Logistics department has the greatest ratio and the greatest number of publications in Table 1. This would indicate that external collaboration helps in the production count.

5. CONCLUSIONS

In the first part of this study we analyze the properties of the social network of researchers at Molde University College and recognize that the degree distribution of collaborations is similar to other scale free networks. This is significant because understanding how the members of the network connect to each other and the importance of "connectors" or hubs within this network may encourage the members of this network or other small institutions to come in contact with other researchers and to form new associations or links. The hubs in Figure 6 are apparent and traceable.

Second we analyzed the connection between researcher productivity as measured in the total number of publications and the existence of coauthor collaborations. It was asked if one can boost the total number of publications by increasing the number of coauthors on each paper. The answer seems to be no. The data in Figure 3 supports this conclusion. In short the average number of coauthors per publication across all researchers was 1.21 coauthors per work with a .9 standard deviation. That is, the author with the most publications (data point with count of 261) had an average of 1.36 coauthors per publication. Another researcher with 26 publications had an average of 1.19 coauthors per publication. Although the publication count is an order of magnitude different, the average numbers of coauthor links are almost alike. We conclude that listing many coauthors on research cannot be the only factor in contributing to a high number of publications. Figure 2 however, indicates another factor contributes to a greater number of publications that is the number of unique associations.

The relationship between the unique associations and the number of published works cannot be concluded to be a one way causal relationship. As others found that two factors are present in networks that exhibit the scale free structure that is growth and preferential attachment. [3] This was also observed in the study of a citation database. [14] Our data also supports this point: while more associations can lead to more production, it is also true that more production can lead to more associations. For a young researcher, it is only important how they can join this virtuous cycle. Therefore, we finally looked at the role of the "connectors" or hubs to see the type of connections made. In addition to acting as connector to between research faculties, the connectors appear to have many external contacts. As indicated in the betweeness data points in Figure 4. Also Table 5, establishes that external contacts seems to contribute to the total number of publications for the faculties and the school.

Last the cluster coefficients in Table 6 for the entire MC and for the faculties demonstrate that there is collaboration within faculties, but it may not be as commonplace for faculty between faculties to work together. This is in agreement with other studies that working partners are not chosen randomly, and is also supported by the network's degree distribution and confirms this network as a scale free network.

In summary, the most important factor in network productivity is that faculty should establish "new" contacts. The addition of new unique coauthors seems to contribute most to the growth of publications.

REFERENCES

[1] Albert, R., Jeong, H. and Barabási, A.L., "Diameter of the World Wide Web,"NPG, Nature, 401, 1999, 130-131.

[2] Albert, R., and Barabási, A.L., "Statistical Mechanics of Complex Networks," Amer Physical Society, Reviews of Modern Physics, Vol 74, 2002, 1-51.

[3] Barabási, A. L., "Linked", Perseus Pub. Cambridge, MA, USA, 2002.

[4] Barabási, A. L, H.Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vissek, "Evolution of the Social Network of Scientific Collaborations," Physica Scripta, Physica A 311, 2002, 590-614.

[5] De Castro, R. and Grossman J.W., "Famous trails to Paul Erdos", Springer New York, Mathematical Intelligencer, 21(3), 1999, 51--63.

[6] Dorogovtsev, S.N., Mendes, J.F.F., and A.N. Samukhin, "WWW and Internet models from 1955 till our days and the popularity is attractive principle," Cornell University Library, Physics E-print Archive: arXiv:cond-

mat/0009090 v1, 2000.
[7] Erdős, P. and Rényi, A., “On Random Graphs I”, Math. Debrecen, vol. 6, 1959, 290-297, In Karonski, M. and Rucinski, A., “The Origins of Theory of Random Graphs, “ in The Mathematics of Paul Erdős, eds. Graham, R. L. and Nesetril, J., pub. Berlin_ Springer, 1997.
[8] Flake, G. W., Lawrence, S. and Lee Giles, C. “Efficient Identification of Web Communities”, ACM, Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining, Boston, Mass, USA, August, 2000, 156-160.
[9] Gladwell, M., “The Tipping Point,” Little Brown, New York, NY, USA, 2000.
[10] Goh, K.-I., Oh, E., Kahng, B., and D.Kim, “Betweeness centrality correlation in social networks,” Amer Physical Society, Physical Review E 67 017101, 2003, 1-4.
[11] Granovetter, M. “The Strength of Weak Ties,” University of Chicago Press, American Journal of Sociology 78, 1973, 1360-1380.
[12] Grossman, J. W., website that explains the Erdős number. (Access: http://www.oakland.edu/enp/), 2004.
[13] Hannemann, R.A. “Introduction to Social Network Methods”, University of California, http://faculty.ucr.edu/~hanneman/SOC157/TEXT/C6Centrality.html, 2004.
[14] Newman, M., “Clustering and Preferential Attachment in Networks”, Amer. Physical Society, Physical Review, E64, 2001, 25102.
[15] Petermann, T. and De Los Rios, P., “Exploration of Scale-Free Networks”, Ecole Polytechnique F´ed´erale de Lausanne, 1015 Lausanne, Switzerland, Manuscript Number: arXiv:cond-mat/0401065 v1, to appear Eur. Phys. J. B, 2004.
[16] Simon, H.A., “On a class of skew distribution functions, “ Cambridge U. Press, Biometrica, 42, 1955, 425-440.
[17] Watts, D. J., and Strogatz, S. H., “Collective Dynamics of ‘Small-World’ Networks,” NPG, Nature 393, 1998, 440-442.
[18] Watts, D. J., “Small-Worlds”, Princeton University Press, Princeton, NJ, USA, 1999.