Growth dynamics and structure of social networks

Kristinka Ivanova Baojun Qiu C. Lee Giles Henry C. Foley
John Yen
College of Information Sciences and Technology
The Pennsylvania State University
{kivanova, bqiu, giles, hfoley, jyen}@ist.psu.edu

Abstract

We present a study of the growth dynamics and topology of social network of scientific collaborations in the field of nanosciences. Two scientists are considered connected if they have authored a paper together and explicit networks of such connections are constructed using data from the Science Citation Index during a 26-year period (1980 – 2006). Through empirical analysis we infer the growth and topological measures that characterize the network as well as its time evolution. Recently suggested model incorporating the effects of locality, that is the shortest distance between nodes, and the attachedness, which is the degree factor, is applied and found to describe well the empirical relations.

1. Introduction

A social network is a collection of people, each of whom is acquainted with some subset of the others. Such a network can be represented as a set of nodes (or vertices) denoting people, and links (or edges) denoting the fact that they are acquainted. Social networks can be constructed for various settings and communities identified by collaboration, spread of disease, spread of ideas, or terrorist networks, to name a few. In this paper we study as an example scientific collaboration network in the field of nanosciences. The choice is justified by the fact that first, the Science Citation Index provides a reliable database that is extended over many decades and available in electronic format and second, the connectedness in such a social network is precisely defined through the papers scientists write. The methodology outlined in this paper, however, has the potential to be implemented in studies of military and terrorist relevant networks as data become available.

In accordance with the results from empirical analysis of scientific collaboration networks in nanosciences we present model simulation results from the model that incorporates both the locality and attachedness factors [1].

2. Observations

Scientific collaborations are best documented through the papers that scientists write and the Science Citation Index provides an extended electronic database of that. We focus on the nanoscience field because this is a recently emerged research field and therefore, we could have the opportunity to study the network structure and growth dynamics since the time the network occur.

We acquire the records by inquiring at the *Thomson Scientific* website [2] directly. Using keyword-based queries generated based on an iterative relevance feedback technique [3], we obtain 368511 SCI-indexed papers regarding nanosciences. The essential idea of this approach is to augment the keyword set until the returned results converge. In addition, we extract several sub-communities from the *NanoSCI* dataset using keywords such as *NanoTube*, *NanoWire*, *NanoParticle*, *Fullerene*, etc. These sub-communities vary from each other in terms of start year, the number of papers and authors. We consider that the *NanoSCI* and each of the sub-communities represent a scientific collaboration network. In each network the nodes are the scientists and two scientists are connected if they have coauthored a paper, which is represented as a link. The number of papers and the number of scientists for the *NanoSCI* and for each of the four nanosciences networks as of year 2006 are listed in Table 1.

3. Data analysis

Next we present the results from empirical analysis that provide information about the dynamics and structural mechanisms that govern the growth and topology of these collaboration networks.

3.1 Growth rate of the network

In order to assess the evolution of both the number of scientists and the number of papers that they write we plot

Table 1. Number of scientists and number of papers in different nanoscience communities as of year 2006

Dataset	$Scientists$	$Papers$
NanoSCI	292393	368511
NanoTube	31688	25285
NanoWire	86234	80645
NanoParticle	81734	69530
Fullerene	97641	96331

the number of papers as a function of the number of scientists. Results for *NanoSCI* are shown in Fig. 1. These results indicate that the growth dynamics of the collaboration networks is characterized by power law increase of the number of papers as the number of scientists in the network increases. We obtain similar power law increase for each of the four nanoscience networks as listed in Table 1. Results for the power law fit of the observational data with a line in double logarithmic scale is $E(t) = 3.05 * V(t)^{1.11}$ for *NanoSCI* network and is $E(t) = 2.5 * V(t)^{1.10}$ for *NanoTube*, where $E(t)$ denotes the number of edges (papers) and $V(t)$ denotes the number of vertices (scientists) at time t.

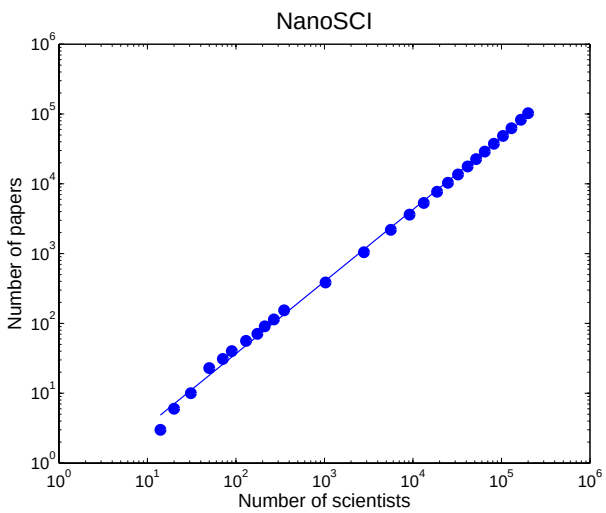

Figure 1. Growth rate of the $NanSCI$ network for the period from 1980 to 2006 expressed as the number of papers versus the number of scientists that wrote these papers.

The power law dependence of the growth dynamics indicates that this is a scale-free growth and we can estimate the rate of growth of the network by taking the derivative of this power law dependence. By doing that we obtain a quantity that describes the rate of change of edges versus vertices, which is called the edge densification rate $\Delta\chi(t)$

$$\Delta\chi(t) = \frac{\Delta E}{\Delta V}. \quad (1)$$

The edge densification rates is $\Delta\chi(t) = 3.39 * V(t)^{0.11}$ for *NanoSCI* network and $\Delta\chi(t) = 2.78 * V(t)^{0.10}$ for *NanoTube* network. The edge densification rate of *NanoSCI* network is used to compare the simulation results of two network growth models, which will be discussed in section 4.

3.2. The giant component

Considering a database that spans over a long period of time (26 years) allows us to observe the occurrence and the growth of the scientific collaboration networks in nanosciences. It is known that there is a possibility of a percolation transition in social networks [4]. Usually at the beginning there are small islands of collaborations. As the total numbers of scientists and the connections between them increase the islands become more and more connected. A percolation phase transition is the sharp transition of the network from a multitude of islands of collaboration to a situation in which one large component (or cluster), comprising a substantial fraction of the scientists, emerges: the so-called gaint component. Usually in the vicinity but before the transition, the largest component has an almost linear or branched structure [5]. In their study of collaboration networks Guimera et al. [5] find that as the network grows the largest component starts to have more and more loops, and eventually, it becomes a densely connected network.

As it is done typically in the study of percolation phase transition [4], we use the fraction of scientists that belong to the largest component of the network to quantify the transition between two regimes: one is an invisible college — that is the web of connections linking the scientists across universities, and the other regime is the network of isolated schools. In Fig. 2 we present the evolution of the size of the largest component which is the number of scientists that belong to the largest component divided by the total number of scientists in the network at certain time, for three of the nanoscience networks, the *NanoSCI*, the *NanoTube*, and the *NanoWire*.

We find that the size of the largest component for these networks increases and at the end of 2006 it reaches 87% for *NanoSCI*, 81% for *NanoTube*, and 70% for *NanoWire*. It is known that a percolation phase transition occurs at around 50–60% depending on the structure of the system [4]. The results shown in Fig. 2 indicate that the scientific collaboration networks in nanosciences expand over the years and the *NanoSCI* experienced the percolation transition around 1998, the *NanoTube* and *NanoWire* went through the perco-

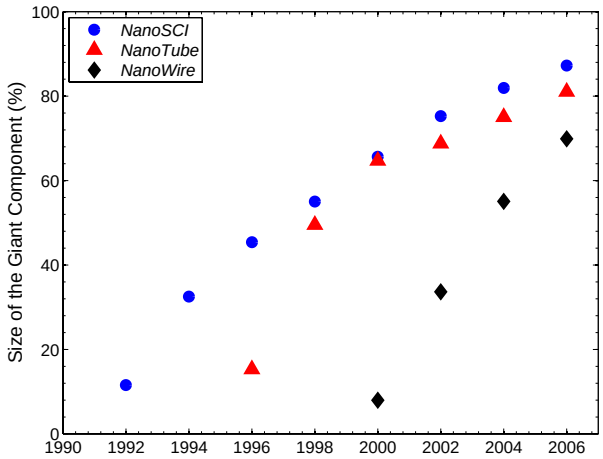

Figure 2. The evolution of the size of the giant component for the *NanoSCI* (circles), *NanoTube* (triangles) and *NanoWire* (diamonds) networks.

lation transition around 2000, and 2004, respectively. The latter two networks occur much later in the structure of the nanoscience network but expand much more rapidly.

We intend to explore further the growth dynamics of the giant component and to obtain more specific means of quantifying the percolation transition.

3.3. Clustering

The social networks have another important property. Real social networks are clustered, meaning they possess local communities in which a higher than average number of people know one another. One way of testing for such clustering in a network is to measure the fraction of triples in a network, also called clustering coefficient C, which for a collaboration graph is the average fraction of pairs of a person's collaborators who have also collaborated with one another. We find that the values of the clustering coefficient for *NanoSCI*, *NanoTube*, and *NanoWire* are 0.82, 0.81, and 0.87, respectively. These high values of the clustering coefficient are suggestive of a structure of the networks that contain highly dense and internally connected hubs that are sparsely connected between each other. These findings are in accordance with the evolution of the giant component that we discussed in previous subsection.

3.4. Average shortest distance

As the network evolves the distance between nodes changes. To quantify the distance $r_t(v_i, v_j)$ between any two nodes v_i and v_j at time t we consider the average shortest distance $\overline{r}(t)$ defined as

$$\overline{r}(t) = \frac{\sum_{i,j} r_t(v_i, v_j)}{V(t)(V(t)-1)}, \qquad (2)$$

where $V(t)$ denotes the set of vertices at time t.

Results for the average shortest distance inside the giant component are shown in Fig. 3 for *NanoSCI*, *NanoTube*, and *NanoWire* networks. Note that the average shortest distance reaches a maximum and then decreases with time for the entire *NanoSCI* network, as well as for each of the two sub-community networks. This is know as the "shrinking diameter" phenomenon [7].

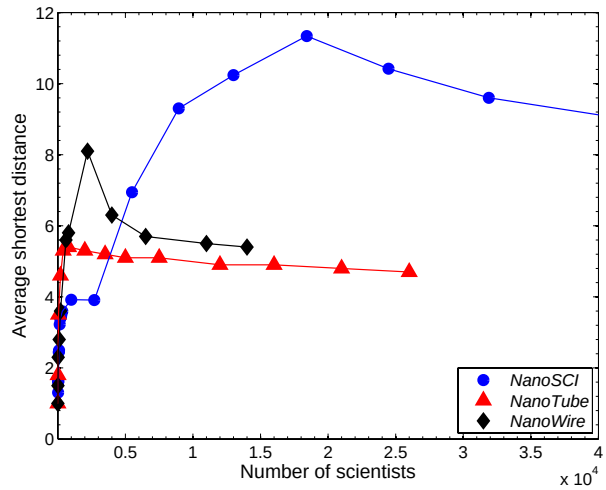

Figure 3. Average shortest distance as a function of time within the giant component for for *NanoSCI*, *NanoTube*, and *NanoWire* networks.

3.5. Degree distribution — an indicator for attachedness

In a network a randomly selected node is connected to d other nodes through d links (edges) with probability $P(d)$ which is called vertex connectivity or degree distribution. We obtain the probability $P(d)$ for each of the nanoscience networks. The results for *NanoSCI* and *NanoTube* calculated in equidistant in logarithmic scale bins are plotted in Fig. 4. Triangles mark the degree distribution of all nodes that exist in the *NanoSCI* network from its inception through the end of 2005. Crosses mark the degree distribution of all nodes that exist in the *NanoTube* network from its inception through the end of 2005. The tails of both of these distributions exhibit a behavior that is close to a power law. Networks that show such power law distribution are know as scale-free networks [6].

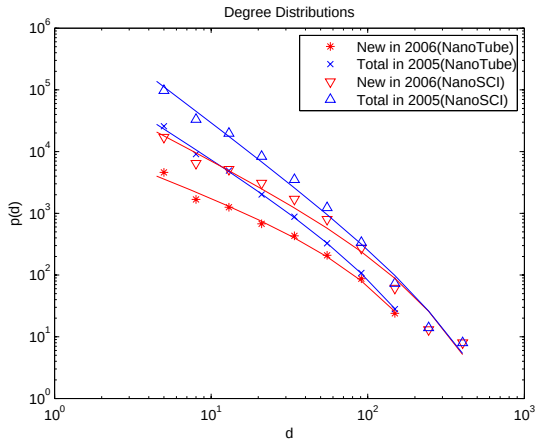

Figure 4. Degree distributions of the *NanoSCI* and *NanoTube* networks for all nodes through the end of 2005 (triangles and crosses, respectively) and for nodes that occur in 2006 only (inverted triangles and asterick, respectively). Each set of symbols is fitted with a power law with exponential cut-off. See the text for discussion and consult Table 2 for the values of the parameters of the fit.

There has been a significant discussion about the degree distribution of various networks in the literature. Barabasi and Albert [6, 7] have suggested that power law distribution may apply to most of the networks of interest including social networks. They report that scientific collaboration networks in mathematics and neuroscience scale with power law exponent of 2.4 and 2.1, respectively. We find similar values of the power law exponent for nanoscience networks (see Table 2). Amaral et al., [8] present a variety of evidence for existence of different classes of networks. Newman [9] report on the structure of scientific collaboration networks and find that collaboration networks in condensed matter physics, astrophysics, high-energy physics and computer science, all can be best fit with a power law form with exponential cut-off.

Similarly to [9] we find that the degree distribution of the networks in nanosciences are best fit with a power law form with exponential cut-off

$$P(d) \sim d^{-\tau} e^{-d/d_c}, \tag{3}$$

where τ and d_c are constant whose values are listed in Table 2.

In order to quantify the growth dynamics of the network we compare the degree distribution of the nodes that occur in year 2006 only. These are scientists that join the network during 2006. Their degree distribution is marked by

Table 2. Summary of results of the analysis. The p-values of the fit for all coefficients are less than 0.0001

Dataset	τ	d_c	R^2
NanoSCI			
Total in 2005	2.21	250	0.99
New in 2006	1.77	166.7	0.98
NanoTube			
Total in 2005	1.94	108.7	1.0
New in 2006	1.41	87.7	0.99

inverted triangles for the *NanSCI* network and by asterisks for the *NanoTube* network. Again they are best fit with a power law form with exponential cut-off (3) and the values of τ and d_c are listed in Table 2. The lower values of the power law constant τ for the degree distribution of new nodes that occur in 2006 for both *NanSCI* and *NanoTube* networks indicate that these networks form new links with probability that is higher than the probability of an existing node to have certain number of degrees. This is a very specific result and is used to define one of the major characteristics of the model presented in section 4. The fact that the networks form new links with probability higher than the probability of the existing nodes is an indicator for the attachedness of the network.

3.6. Distance distribution — an indicator for locality

Topological proximity is another factor that influence the ability of the nodes to attract new links. To assess this we calculate the distribution of the distance between two nodes for each of the networks. The result for the *NanoSCI* network is shown in Fig. 5. Again we are interested in the evolution of the distribution. To estimate such an evolution we obtain the distance distribution for the network through the end of year 2005 and plot the result with circles in Fig. 5. The distance distribution for the nodes (scientists) and links (papers) that occur in year 2006 only is plotted in the same figure with asterisks. These results clearly indicate that the network forms new links with probability inversely proportional to the distance. They are as well a criterion for the locality factor which will be discussed in section 4.

4. Model

Several relationships are derived from the empirical analysis presented so far. We obtain that the scientific collaboration networks in nanosciences are characterized by scale-free growth dynamics that allow us to define the edge

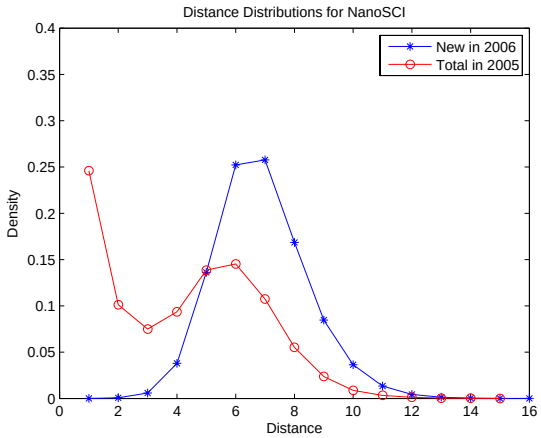

Figure 5. Distance distribution of the *NanoSCI* network. Circles denote the distance distribution of the network by the end of 2005. Asterisks denote the distribution of direct collaborations added in 2006.

densification rate which is in essence the growth rate of the network. Second, we show that the networks form new links with probability that is proportional to the degree and inversely proportional to the distance. These findings satisfy the basic rules of the recently suggested locality and attachedness based model [1].

Let us briefly describe the distance (locality) and degree (attachedness) based model, called the *DDG* model [1]. The proximity between two individual nodes can be measured by a variety of measures. We use shortest distance to measure the proximity between two nodes. In the growth process, a new vertex and l edges ($l = \Delta E(t)$) are added to the graph at each time step. The *DDG* model specifies that the two vertices of a new edge are determined in the following way: (1) one node n_s is selected uniformly from the graph as the start vertex of the new edge; (2) the probability that a node v_p is selected as the end node of the new edge is

$$p_t(v_p) = \frac{\frac{d_t^{'}(v_p)}{r_t^{'}(v_p, v_s)}}{\sum_p \frac{d_t^{'}(v_p)}{r_t^{'}(v_p, v_s)}} \tag{4}$$

where

$$r_t^{'}(v_p, v_s) = \begin{cases} R & \text{if } r_t(v_p, v_s) = \infty \\ r_t(v_p, v_s) & \text{otherwise.} \end{cases} \tag{5}$$

and $d_t(v_p)$ is defined by

$$d_t^{'}(v_i) = d_t(v_i) + 1 \tag{6}$$

where $r_t(v_p, v_s)$ is the shortest distance between nodes v_p and v_s and $d_t(v_i)$ is the degree of node v_i at time t.

Equation 5 specifies that if two nodes are disconnected, the distance between them is a large number R. This way, the probability of building a new connection between any two nodes is non-zero.

Simulation results for the average shortest distance produced with the *DDG* model are shown in Fig. 6. Note the maximum in the *DDG* curve. We compare this result with simulation output from another model. This is the parameterized preferential attachment model *PPAM* [7] in which the probability that a node is selected as the end node v_j of the new edge is

$$p_t(v_j) = \frac{d_t^{'}(v_j)}{\sum_k d_t^{'}(v_k)} \tag{7}$$

where $d_t^{'}(v_i)$ is defined as in (6).

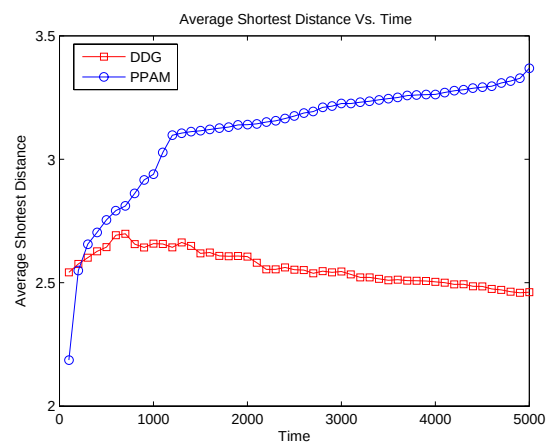

Figure 6. Model results for average shortest distance as a function of time for edge densification rate $\chi(t) = 3.39 * V(t)^{0.11}$ using DDG (squares) and PPAM (circles) models.

Simulation results from the preferential attachment model calculated with the same diversification rate are plotted with circles in Fig. 6 and show an increase and a tendency to saturate, but do not demonstrate a maximum in the evolution of the average shortest distance as do the *DDG* model results.

5. Conclusions and future work

We present a study of the growth dynamics and structure of scientific collaboration network which can serve as a prototype of a complex evolving network with wide applications, including military ad hoc communication network and terrorist networks. The empirical analysis of real-world data provided the main characteristics and relation-

ships that describe a specific network. We find that collaboration networks in nanosciences are characterized by scale-free growth dynamics that allow us to define edge densification rate. Typical for these networks are the quickly increasing giant component and high clustering, which are indicative of a structure of dense hubs sparsely connected between each other known as assortive networks [10]. Next we test the vertex connectivity by assessing the degree distribution and find a power law dependence (with exponential cut-off) which is a characteristic of scale free networks. The exponential cut-off is more pronounced for the distribution of links that occur in 2006, and therefore might be caused by the finite time window. Furthermore, we present model simulation results that show the "shrinking diameter" feature of the real data.

We intend to focus on a detailed study of the evolution of the giant component and especially to quantify the percolation phase transition as this is of paramount importance for investigating the growth dynamics of social networks.

References

[1] H. Zhang, B. Qiu, K. Ivanova, C. L. Giles, H. C. Foley, and J. Yen. Locality and Attachedness-based Temporal Social Network Growth Dynamics Analysis. *J. Am. Soc. Inform. Sci. Tech. submitted*

[2] http://scientific.thomson.com/products/sci. 2006.

[3] R. N. Kostoff, J. A. Stump, D. Johnson, J. S. Murday, C. G. Lau, and W. M. Tolles. The structure and infrastructure of the global nanotechnology literature, 2006.

[4] D. Stauffer, A. Aharony. *Introduction to Percolation Theory* (Taylor and Francis, London), 2nd Ed.

[5] R. Guimera, B. Uzzi, J. Spiro, and L. A. N. Amaral. Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, 308(5722):697–702, 2005.

[6] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[7] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:3, 2002.

[8] L. A. N. Amaral, A. Scala, M. Barthelemey, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA*, 97:11149, 2000.

[9] M. E. J. Newman. The structure of scientific collaboration network. *Proc. Natl. Acad. Sci. USA*, 98:404, 2001.

[10] S. Redner. Testing out the missing links. *Nature*, 453:47, 2008.