Locality and Attachedness-Based Temporal Social Network Growth Dynamics Analysis: A Case Study of Evolving Nanotechnology Scientific Collaboration Networks

Haizheng Zhang
Microsoft, One Microsoft Way, Redmond, WA 98052. E-mail: haizhengzhang@gmail.com

Baojun Qiu
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802

Kristinka Ivanova, C. Lee Giles, Henry C. Foley, and John Yen
College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802

The rapid advancement of nanotechnology research and development during the past decade presents an excellent opportunity for a scientometric study because it can provide insights into the dynamic growth of the fast-evolving social networks associated with this field. In this article, we describe a case study conducted on nanotechnology to discover the dynamics that govern the growth process of rapidly advancing scientific-collaboration networks. This article starts with the definition of temporal social networks and demonstrates that the nanotechnology collaboration network, similar to other real-world social networks, exhibits a set of intriguing static and dynamic topological properties. Inspired by the observations that in collaboration networks new connections tend to be augmented between nodes in proximity, we explore the locality elements and the attachedness factor in growing networks. In particular, we develop two distance-based computational network growth schemes, namely the *distance-based growth model* (*DG*) and the *hybrid degree and distance-based growth model* (*DDG*). The DG model considers only locality element while the DDG is a hybrid model that factors into both locality and attachedness elements. The simulation results from these models indicate that both clustering coefficient rates and the average shortest distance are closely related to the edge densification rates. In addition, the hybrid DDG model exhibits higher clustering coefficient values and decreasing average shortest distance when the edge densification rate is fixed, which implies that combining locality and attachedness can better characterize the growing process of the nanotechnology community. Based on the simulation results, we conclude that social network evolution is related to both attachedness and locality factors.

Received November 13, 2008; revised August 4, 2009; accepted August 12, 2009

© 2010 ASIS&T • Published online 1 February 2010 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21225

Introduction

Nanotechnology is a highly interdisciplinary field that is generally concerned with the control of matter on the molecular level in scales smaller than 1 μm, normally 1 to 100 nm, and the fabrication of devices within that size range. This field has been growing at an astounding pace in the last decades. This is reflected in the world-wide growth of funding from both government and industry and the increasing penetration into other disciplines as well as the accelerating growth in the number of scientific publications and involved researchers. The explosive development of this field makes it ripe for in-depth scientometric analysis for this field. This article conducts a study on the evolving nanotechnology collaboration network to develop insights into its social network growth dynamics.

This study is in line with the surging interests in social network and complex network studies in the recent decades. In particular, researchers have discovered that many social networks and other real-world complex networks exhibit a set of properties that distinguish them from random networks such as the *Erdos-Renyi* model (Bollobs, 2001). These properties fall into two categories. The first category consists of a set of static topological properties that characterizes social network graphs, including (a) power-law degree distribution, (b) large clustering coefficient values, and (c) small average shortest path between two random nodes. In particular, the power-law degree distribution is a distinct feature of scale-free networks; large clustering coefficient values usually imply manifest community structures; and small average shortest path length values indicate short average separation between nodes. The latter two features constitute the so-called *small-world network properties*.

The second category is the kinetic properties exhibited in the growing process of social networks. For instance, researchers have reported the "shrinking diameter" phenomenon that the diameters of many real networks decrease over time (Barabasi et al., 2002; Kumar, Novak, & Tomkins, 2006; Leskovec, Kleinberg, & Faloutsos, 2005). These static and dynamic properties represent a significant departure from random networks. While these phenomena have been identified, there is no consensus on the cause of these features. Recently, there has been a flurry of efforts by researchers from different disciplines exploring a variety of factors, including the attachedness factor (i.e., the degree of nodes) and the locality factor, to discover the growth schemes of social networks (Barabasi & Albert, 1999; Jin, Girvan, & Newman, 2001; Kumar et al., 2000). The first section provides an overview of the existing studies.

The nanotechnology collaboration network studied in in this article, *NanoSCI*, is appealing for investigating social network growth dynamics for the following three reasons. First, collaboration networks have been widely used in scientometrics and social networks study. It has been discovered that collaboration networks possess many static and dynamic properties that are similar to other social networks. In his early work on this domain, Newman (2001, 2004) studied several large collaboration networks and concluded that these networks exhibit all the general ingredients of small-world networks, including short node-to-node distance and large clustering coefficient. Moreover, researchers have recently shown that evolving collaboration networks exhibit similar dynamic patterns as do other social networks in the growth process, such as shrinking diameters and high clustering coefficient values (Barabasi et al., 2002). Second, NanoSCI offers one of the most extensive databases to date on social networks, including 292,323 researchers and 368,511 papers that are indexed by the *Science Citation Index* (*SCI*) database (http://scientific.thomson.com/products/sci, 2006) spanning from 1980 to 2006. Finally, the history of nanotechnology research is very short, and the literature has developed so recently that the majority of it is online. Compared to other fields, even new ones such as biotechnology or super conductivity, the short history of the field combined with its fully online character facilitates this kind of metascientific study. Thus, NanoSCI provides unique research opportunities for us to investigate the characteristics of the formation stage of collaboration networks.

In particular, this article reports a set of static properties and dynamic patterns observed in the evolving nanotechnology collaboration network. Based on these observations, we explore the joint effect of attachedness (degree) factor and locality factor for network growth dynamics. This article proposes two distance-based computational growth schemes, namely *DG* (distance-based growth model) and *DDG* (hybrid degree and distance-based growth model), and compares them with other growth models. In the DG model, the probability of building a new connection between two nodes is in inverse proportion to their distance. The DDG model, similar to the *Law of Gravity*, specifies that the attractiveness between two nodes is determined by their degree and the distance. Based on the simulation results, we discover that both clustering coefficient rates and the average shortest distance are closely related to the edge densification rate, a metric that measures the relative growth speed of edges and nodes. In addition, the hybrid DDG model exhibits higher clustering coefficient values and decreasing average shortest distance when the edge densification rate is fixed, which implies that combining locality and attachedness can better characterize the growth of the nanotechnology community. To summarize, the contributions of this article include (a) exploring the locality and the attachedness-based network growth paradigm and the corresponding dynamics and patterns, and (b) discussing factors that can cause high clustering coefficient values and the "shrinking diameters" phenomenon in temporal social networks.

The rest of this article is organized as follows: First, we introduce the background of this study and give a brief review of the related works. We then define terminology and concepts used in this article The subsequent section presents several observations in social networks that motivate our work. Then, we describe two computational models that incorporate distance and attachedness factors, and present quantitative analysis on the impact of these two factors on the topological properties of graph. Finally, some potential future work is outlined.

Related Work

A proliferation of work studying the evolution dynamics of complex networks has happened in the last decade. These related studies include static analysis on social network evolution and a variety of models to reproduce the static topological properties and dynamic patterns observed in social networks. The majority of the related studies focus on either the degree factor or the locality factor. This section provides a list of representative studies along this line:

Attachedness-Based Growth Schemes

Attachedness is a concept that measures how well nodes are connected and therefore usually is reflected by the degree of nodes in complex networks. In their well-received article, Barabasi and Albert (1999) developed the notable *Preferential Attachment* theory that specifies high-degree nodes are always favored when building new connections. In their article, the authors developed a model in which new nodes are

DOI: 10.1002/asi

added to the network one by one. The authors claimed that the probability that a node v_n will be linked to a vertex v_i depends on v_i's degree, $\frac{d_i}{\sum_j d_j}$, where d_i is the degree of node v_i. Each new node attaches itself (i.e., creates a link) to one of the existing nodes with a certain probability that is proportional to the number of links that the existing nodes possess. The authors demonstrated that this simple scheme results in power-law degree distribution and "rich get richer" phenomenon. In a later article, Barabasi et al. (2006) developed a "continuum theory" based on the preferential attachment theory and used a Monte Carlo approach to simulate the network growth process. The authors showed that the clustering coefficient value can decrease or increase by adjusting the parameter that specifies the number of newly created internal links per node in unit time. However, in contrast with the decreasing average shortest distance observed in reality, this approach results in increasing average separation, which was attributed to the incomplete data by the authors. In this study, we demonstrate that the trend of average shortest distance is closely related to edge densification rate. In addition, combining both locality and attachness rate tends to result in decreasing average nodes separation.

Locality-Based Growth Scheme

Degree-based models assume that the attractiveness between nodes only depends on their degrees and is independent of the distance between them. In contrast, many of the existing models explicitly or implicitly exploit the locality factor and assume that the generation of a new connection between two arbitrary nodes is related to how far apart they are in the existing topology. The following is a list of these models:

- *Copying mechanism:* This model specifies that at each time step a new node is added to the network by connecting to a constant number of existing nodes in the network (Kumar et al., 2000). The new node copies a number of links from a "prototype" node that is selected randomly from the existing nodes whereas choosing the remaining neighbors is random. The authors showed that this process can result in scale-free distribution.
- *Walking on a network:* Inspired by citation networks, Vazquez (2000) designed the *Walking on a network* scheme to simulate the graph growth process. The model specifies that a network always starts with an isolated node. At every time step, a new node v_i is added and linked to a randomly selected node v_i through a directed edge. The node v_i then mimics a "random walk" on the network by following the edges starting from node v_j and linking to their endpoints with probability p. This step is repeated for those nodes to which new connections were established, until no new target node is found.
- *Referral Model:* Davidsen, Ebel, and Bornholdt (2002) presented a simple scheme that connections are always formed between two nodes that share a common neighbor. This model emulates the real-world scenario that one person introduces two of his or her acquaintances to get to know each other. Such a simple evolution scheme is viewed as the basis of the evolution of social networks. The authors demonstrated that this simple scheme is able to reproduce major nontrivial features of social networks, including short path length, high clustering, and scale-free or exponential-degree distribution.
- *Distance Preference Model:* Jost and Joy (2002) described a purely distance-based scheme where each new node is connected to a randomly selected node and the subsequent connections are related to the distance of the destination node. This computational model resembles the DG model in our work. However, the authors focused on the discussion of degree distribution and assumed that new nodes are always connected to the rest of the networks upon joining. Thus, at any given time, there is only one Giant Connected Cluster (GCC) in the network. This assumption can lead to very different dynamics of network growth.

Similar graph growth mechanisms also include models that implicitly or explicitly rely on the locality heuristics (Guìmera, Uzzi, Spiro, & Amaral, 2005; Kossinets & Watts, 2006; Krapivsky & Redner, 2001; Leskovec et al., 2005; Liben-Nowell & Kleinberg, 2007; Watts, Dodds, & Newman, 2002) or specified feature similarity (correlation) between nodes (Xuan, Li, & Wu, 2007). In particular, Guimera et al. (2005) proposed a team assembly mechanism by investigating the interplay between "incumbents" and "newcomers" in the context of collaboration networks. This mechanism focuses on the establishment of "collaborations" rather than "links." The authors attempted to reproduce a variety of networks by adjusting the likelihood of different types of agents participating in the collaborations and evaluated the generated networks based on their degree distribution. While this novel team assembly model can be generalized to more generic networks, we study explicitly how proximity and degree play a role in network growth processes. Another more recent work on this front includes Morris and Goldstein's (2007) team-based growth model, known as the *Yule model*, for the bipartite networks that consist of both articles and authors. While both the Yule model and our article are concerned with the growth mechanism of collaboration networks, there exist three main differences between the two. First, as opposed to the Yule model that focuses on the bipartite networks, we only focus on collaboration networks; thus, our model does not model the productivity (i.e., the number of articles generated) by a team of collaborators. However, our model can be potentially applied to other complex networks, especially those in which the cost of connecting two nodes is related to their distance. Second, the team-based Yule model uses preferential attachment for within-team author selection for a new article, and random selection of new authors outside the team. Hence, it adopts a binary locality measure (i.e., whether an author is within a team or outside of a team). In contrast, the proposed hybrid model (DDG) in this article uses a continuous locality measure based on the distance between two authors in the collaboration network. The third difference between the two growth models is that the team-based Yule model does not use preferential attachment for selecting new authors outside of the team whereas our proposed hybrid model applies preferential attachments to all nodes regardless of whether they are close or far away on the network.

These articles have provided valuable insights into the dynamics of complex networks; however, there are inadequate studies in investigating how proximity and degree quantatively contribute to the network growth scheme and the implications on the static and dynamic properties of social network. In this article, we provide in-depth analysis on the growth scheme of social networks and develop network growth models that incorporate both global attachedness and locality factors. Hence, both degree and distance factors are taken into consideration in the hybrid model proposed in this article. The simulation results have indicated that the evolution process of social networks can be better characterized by combining these two factors.

Temporal Social Networks

To investigate the growth and evolution process of social networks, we define evolving social networks as *temporal social networks*. In a temporal social network $G(t) = \{V(t), E(t)\}$, the vertex set $V(t)$ and the edge set $E(t)$ evolve over time. The snapshot of a temporal social network at time t_k, $G(t_k) = \{V(t_k), E(t_k)\}$, is a static graph. The two vertices of an edge $e_i = (v_j, v_k)$ is denoted as $V(e_i) = \{v_j, v_k\}$. The set of new connections that are built at time t_k is $\Delta E(t_k) = \{e_1^{t_k}, e_2^{t_k}, \ldots, e_m^{t_k}\}$. The corresponding vertex set is $\Delta V(t_k)$. Thus, we have

$$\begin{aligned} E(t_k) &= E(t_{k-1}) \cup \Delta E(t_k) \\ V(t_k) &= V(t_{k-1}) \cup \Delta V(t_k) \end{aligned}$$

Note that in the context of collaboration networks, connections can be constructed repeatedly between the same two nodes at different times, which implies that the set $E(t_{k-1})$ and $\Delta E(t_k)$, and $V(t_{k-1})$ and $\Delta V(t_k)$ may not be disjoint. More formally, $E(t_{k-1}) \cap \Delta E(t_k) \supseteq \Phi$ and $V(t_{k-1}) \cap \Delta V(t_k) \supseteq \Phi$, where Φ represents the empty set. We also define the edge density rate of a temporal social network as $\chi(t) = \frac{|E(t)|}{|V(t)|(|V(t)|-1)}$ and the *edge densification rate* as $\Delta\chi(t) = \frac{|\Delta E(t)|}{|\Delta V(t)|}$. The edge density rate is a static concept and describes the density of edges versus nodes at a particular timestamp t. In contrast, the edge densification rate is a dynamic concept that characterizes the speed of edges' growth versus nodes' growth. As will be shown later, the edge densification rate is a crucial factor in determining topological properties of temporal social networks.

The proximity of two individual nodes in a social network is often defined in the context of the investigated application domains. The most widely used measure is the shortest distance betweeen the two nodes. In addition to shortest distance, researchers also have discovered that two additional factors can help: For instance, Koren, North, and Volinsky (2006) proposed a cycle-free effective conductance (CFEC) to measure distance between network nodes by accouting for the multiple and disparate paths between nodes. However, the pairwise CFEC computation is prohibitive for large-scale social networks; Liben-Nowell and Kleinberg (2007) showed that the number of common neighbors is a helpful proximity indicator. In the proximity-based model we develop and describe in the following sections, we adopt a shortest distance for individual proximity measure; however, the other proximity measures can alternatively fit in the model.

Accordingly, the aggregate proximity properties of a social network can be evaluated by a variety of measures, including average shortest distance, diameter, and effective diameter. The diameter dt of a social network is defined as the largest shortest path between any two nodes; that is, $dt = Max_{v_i, v_j \in V} r(v_i, v_j)$. Some researchers use *effective diameter*, a measure that is obtained by taking the 90th percentile of the largest shortest distance combined with interpolation, to reduce variance. However, the diameter, effective diameter, and average shortest distance tend to exhibit similar dynamics in our experiments. Thus, for the sake of simplicity, we use average shortest distance for measuring the aggregate proximity of social networks in the remainder of this article.

In a temporal social network, the distance between two nodes changes over time. The shortest distance from v_i to v_j at time t_k is denoted as $r_{t_k}(v_i, v_j)$. The average shortest distance for a graph at time t_k is denoted as

$$\bar{r}(t_k) = \frac{\sum\limits_{i,j} r_{t_k}(v_i, v_j)}{|V(t_k)|(|V(t_k)| - 1)}.$$

The clustering coefficient $\aleph_{v_i}(t_k)$ for a node v_i at time t_k is defined as the proportion of links between the vertices within v_i's neighborhood divided by the number of links that could possibly exist between them. More formally,

$$\aleph_{v_i}(t_k) = \frac{|\{e_{jk}\}|}{d_{v_i}(t_k)(d_{v_i}(t_k) - 1)} \qquad (1)$$

where $v_j, v_k \in V(t), e_{jk}, e_{ji}, e_{ki} \in E(t)$ and $d_{v_i}(t_k)$ is the degree of node v_i at time t_k. The clustering coefficient of a node v_i measures how well v_i's neighbors are connected to each other. The average clustering coefficient $\overline{\aleph(t)}$ characterizes the modularity of the social network at time t (Ravasz, Somera, Mongru, Oltvai, & Barabasi, 2002).

Evolving graphs $G(t)$ usually consist of a number of isolated subgraphs. It is particularly interesting to investigate the patterns and behaviors of the largest connected cluster, the GCC, denoted as

$$GCC(t) = \{V_{GCC}(t), E_{GCC}(t)\} \text{ where } V_{GCC}(t) \subseteq V(t), \quad E_{GCC}(t) \subseteq E(t)$$

and

$$\forall v_i, v_j (v_i, v_j) \in E_{GCC}(t) \Rightarrow v_i \in V_{GCC}(t)\, and\, v_j \in V_{GCC}(t).$$

Table 1 lists a number of important notations for the concepts and terminology used in this article. Also note that we use "node" and "vertex", "edge" and "connections" synonymously.

DOI: 10.1002/asi

TABLE 1. Terminology and notations for temporal social networks.

Notation	Meaning
v_i	a vertex
e_i	an edge
$V(t)$	the set of vertices at time t
$E(t)$	the set of edges at time t
$G(t) = [V(t), E(t)]$	the graph G at time t
$\Delta V(t)$	the set of new vertices at time t
$\Delta E(t)$	the set of new edges at time t
$\chi(t)$	$\frac{\|E(t)\|}{\|V(t)\|(\|V(t)\|-1)}$ edge density ratio
$\Delta\chi(t)$	$\frac{\|\Delta E(t)\|}{\|\Delta V(t)\|}$ the densification rate of edges versus nodes at time t
$V(\mathrm{e}_i)$	the two vertices of edge e_i
$\aleph(v_i)$	the clustering coefficient of node v_i
$\overline{\aleph(G(t))}$	the average clustering coefficient of graph $G(t)$
$r_t(v_i, v_j)$	the shortest distance between nodes v_i and v_j at time t
$\bar{r}[G(t)]$	the average shortest distance for Graph $G(t)$
$d_t(v_i)$	the degree of node v_i at time t
$C_k(t)$	the expected number of vertices whose degree are k at time t

TABLE 2. Statistics for different nanotechnology communities as of 2006.

Dataset	Researchers	Articles	$\|E_{GCC}\|$	$\|V_{GCC}\|$
NanoSCI	292,393	368,511	1,836,499	268,594
NanoTube	31,688	25,285	149,138	26,849
NanoWire	86,234	80,645	435,451	77,304
NanoParticle	81,734	69,530	400,749	72,905
Fullerene	97,641	96,331	515,898	88,496

Observations and Motivations for Social Network Growth Models

The NanoSCI is a collection of nanotechnology-related articles published and indexed by the *SCI* in the 1980 to 2006 period. The records are acquired by directly inquiring at the Thomson Scientific Web site (http://scientific.thomson.com/products/sci, 2006). Using keyword-based queries generated based on an iterative relevance feedback technique (Kostoff et al., 2006), we obtained 368,511 *SCI*-indexed papers regarding nanotechnology. The essential idea of this approach is to augment the keyword set until the returned results converge. In addition, we extracted several subcommunities of nanotechnology from the NanoSCI dataset using keywords such as NanoTube, NanoWire, NanoParticle, Fullerene, and so on. These subcommunities vary with each other in terms of start year and the number of papers and authors. We consider that the NanoSCI and each of the subcommunities represent a scientific collaboration network. In each network, the nodes are the researchers, and two researchers are connected if they have coauthored an article, which is represented as a link. The number of articles and the number of researchers for the NanoSCI and for each of the four nanotechnology communities as of 2006 are listed in Table 2.

This article compares the proposed social network growth scheme with existing models using the collaboration networks constructed for these communities.

Next, we assess how the number of links (i.e., edges) between researchers and the number of researchers (i.e., nodes) increase with time. Figure 1 shows in log–log scale the edge growth versus node growth for the NanoTube and NanoSCI communities, respectively. It appears that the growth speed is almost linear in the log–log scale, which implies that the edge growth increases as power law as a function of the nodes growth. This finding justifies using the densification laws suggested previously (Leskovec et al., 2005). These regression results show that their growth rates are $|E(t)| = 2.5173 * |V(t)|^{1.1049}$ and $|E(t)| = 3.0459 * |V(t)|^{1.1141}$, respectively. Thus, the corresponding edge-densification rates for the two communities are $\Delta\chi(t) = 2.78 * |V(t)|^{0.10}$ and $3.39 * |V(t)|^{0.11}$, respectively. The edge-densification rate of NanoSCI communities is used in comparing the simulation results of different network growth models, which will be discussed later.

Figure 2 demonstrates the temporal changes of average shortest distance, $\bar{r}(GCC(t))$, of the giant connected component, $GCC(t)$ for NanoSCI, NanoTube, Fullerene, NanoParticle, and NanoWire, respectively. Note that Fullenrene and NanoWire are recently emerging communities, and the number of researchers as of 2006 is less than 20,000 according to the collected data. In this pilot study, we focus on analysis of the $GCC(t)$ and will leave exploring the entire sysytem for future research. These results clearly illustrate the "shrinking diameter" phenomenon that has been reported in Kumar et al. (2006) and Leskovec et al. (2005). This is contradictory to conventional wisdom that would predict that the diameter of growing networks shall increase over time. Leskovec et al. (2005) developed a "forest fire" network growth model, in which the diameters can either increase or decrease over time by adjusting parameters of the model. In the following sections, we show that this phenomenon can be attributed to both edge-densification rates and the way that new connections are formed (i.e., growth models) in the evolving social networks.

The average clustering coefficient, $\overline{\aleph}(G(t))$, is an indicator for the modularity of networks. Table 3 shows the average clustering coefficient for the NanoSCI, NanoTube, Fullerene, NanoParticle, and NanoWire communities. Clustering coefficient value of 0.81 for NanoTube, 0.82 for Fullerene, 0.83 for NanoParticle, and 0.87 for NanoWire communities are significantly higher than those of random networks, which are usually below 0.1 (Albert & Barabasi, 2002), and also higher than the clustering coefficient values for other collaboration networks reported (Newman, 2001). In studying the structure of scientific collaboration networks, Newman (2001) found that the clustering coefficient varies between 0.066 for *Medline* (articles in biomedical research) to 0.43 for articles published in the *Los Alamos Archive* to 0.73 for *SPIRES*

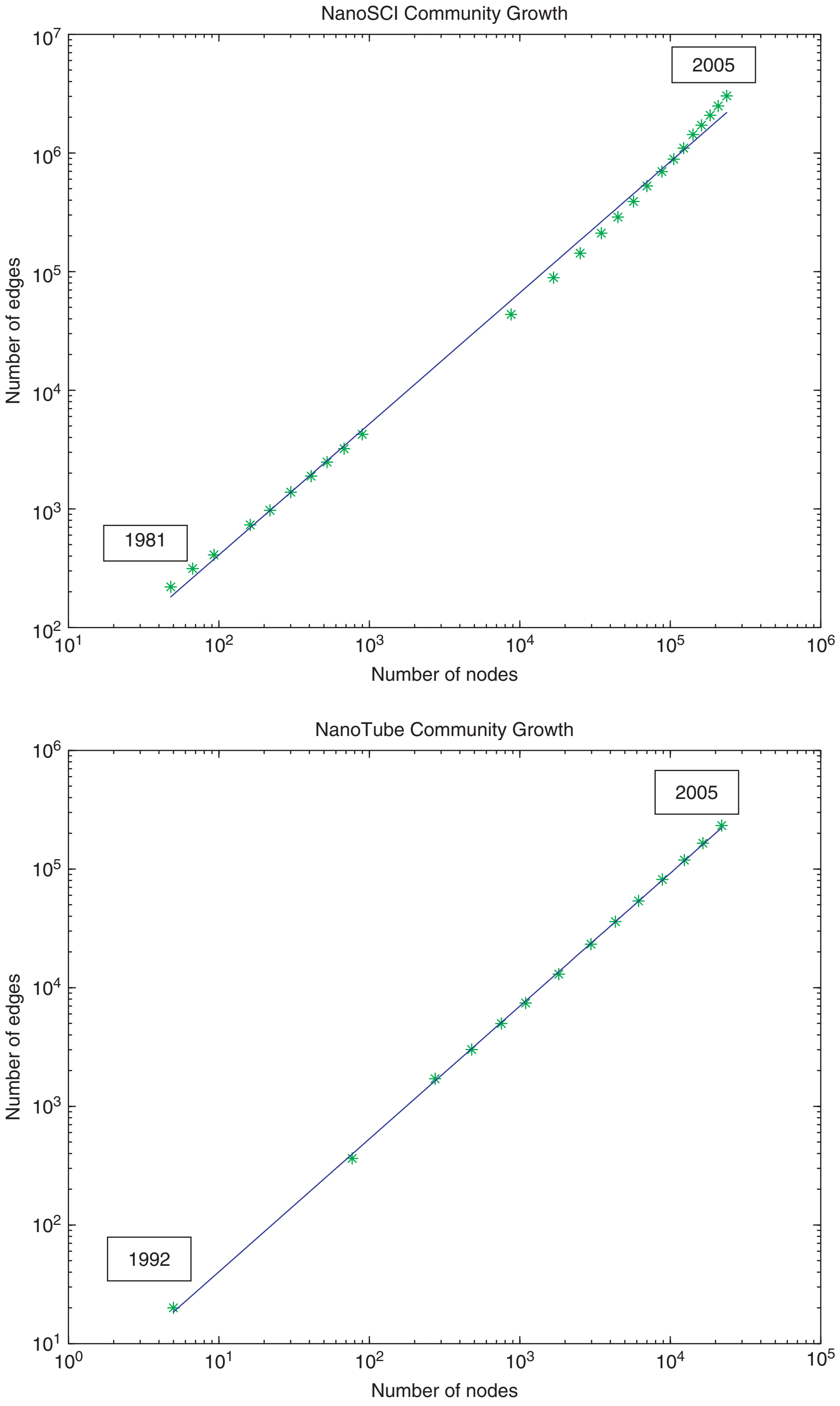

FIG. 1. The number of edges $E(t)$ versus number of nodes $V(t)$ for the NanoSCI and NanoTube communities.

(published articles and preprints in high-energy physics). Values of $\overline{\aleph}(G(t))$ obtained in this study being higher than the clustering coefficient for high-energy physics indicate even higher modularity of the nanoscience communities. Next, we compare the average clustering coefficient of the simulation results of the several network growth models with these observations from the nanotechnology communities.

To explore the causes of such intriguing phenomena, including decreasing shortest distance and high clustering coefficient values, we propose a set of computational models

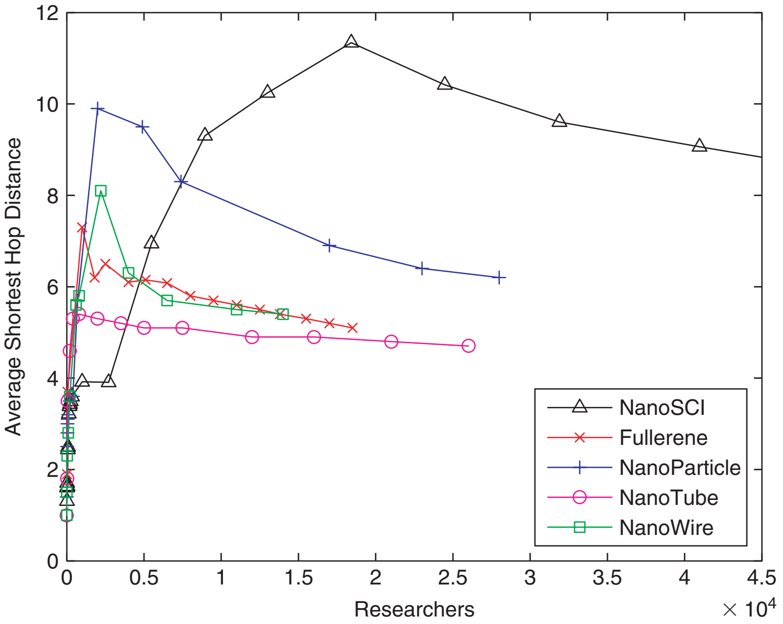

FIG. 2. The average shortest distance $\bar{r}[GCC(t)]$ of $GCC(t)$ for the NanSCI, Fullerene, NanoParticle, NanoTube, and NanoWire communities. Note that Fullenrene and NanoWire are recently emerging communities and that the number of researchers as of 2006 was less than 20,000 according to the collected data.

TABLE 3. Average clustering coefficient values for nanotecnology communities from 1980 to 2006.

Dataset	NanoSCI	NanoTube	Fullerene	NanoParticle	NanoWire
$\overline{\aleph}(G(t))$	0.82	0.81	0.82	0.83	0.87

that employ relatively simple growth schemes and explore how these growth mechanisms can affect the topological properties of the underlying temporal social networks.

Combining Locality and Global Preferential Attachment for Modeling Network Growth

The attachedness factor has been traditionally considered as a principal factor in attracting new connections. In addition to this factor, we have observed that nodes tend to connect to their peers within topologically proximity. For instance, Figure 3a shows the distribution of the shortest distance, as of time t_k, between nodes that establish the third category of connections at time $t_k + 1$ in the $GCC(t_k)$ of the pertaining social network versus the distribution of pairwise shortest distance between the nodes in $GCC(t)$. Figure 3a shows the distribution of $r_{2005}(v_i, v_j)$ where $e_l = (v_i, v_j) \in \Delta E(2006)$ and $v_i, v_j \in GCC(2005)$ for the NanoTube community. We neglect those connections which already had been in the network in previous years. This figure demonstrates that there is remarkable disparity between the two distributions. It indicates that new connections tend to be created between nodes in proximity. In particular, the vast majority of links are added between the nodes that are only two hops apart.

To demonstrate explicitly that nodes form new links inversely proportional to the topological distance, we calculate the proportion $Fr(r) = M/N$, where N denotes the number of node pairs at distance r and where M are the pairs among them that form new edges in the next time step. The results are shown in Figure 3b for the NanoSCI and NanoTube communities.

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 communities. The results for NanoSCI and NanoTube calculated equidistant in logarithmic scale bins are plotted in Figure 4a. 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 (Barabasi & Albert, 1999).

Barabasi and Albert (1999) and Barabasi et al. (2002) suggested that power-law distribution may apply to most of the networks of interest, including social networks. They reported that scientific collaboration networks in mathematics and neuroscience scale with power-law exponents of 2.4 and 2.1, respectively. We found similar values of the power-law exponent for nanoscience networks (see Table 4). Newman (2001) reported on the structure of scientific collaboration networks and found that collaboration networks in condensed-matter physics, astrophysics, high-energy physics, and computer science all can be best fit

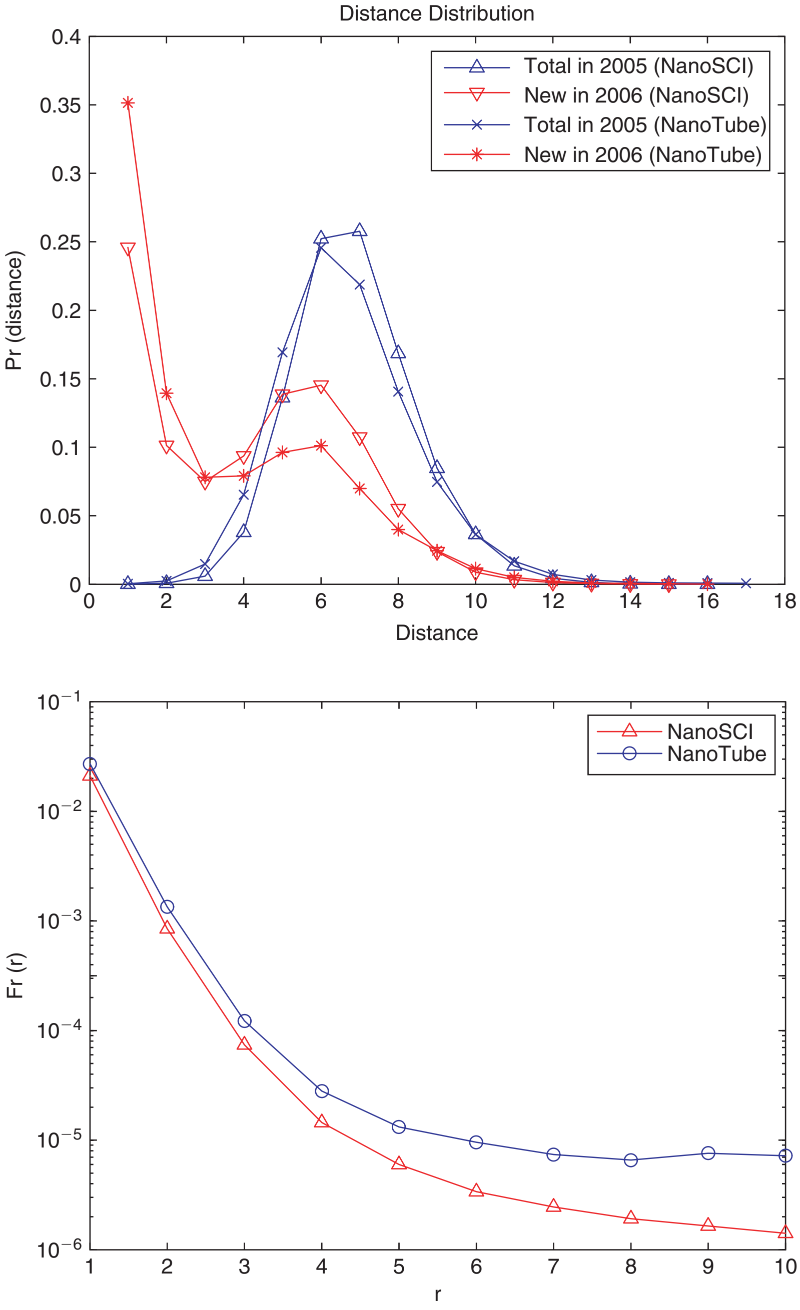

FIG. 3. The locality preference of selecting neighbors in the NanoTube community. The left curve shows the distribution of shortest distance values between those nodes that are in GCC (2005) and build direct collaboration relation between each other in 2006 [i.e., $r_{2005}(v_i, v_j)$ *where* $(v_i, v_j) \in \Delta E(2006)$ *and* $v_i, v_j \in GCC(2005)$]. The right curve shows the distribution of the pairwise shortest distribution for all nodes in $GCC(2005)$ (a). Proportion of node pairs $Fr(r)$ with certain distance forming new edges in a new time step (b).

with a power-law form with exponential cutoff. Similarly to Newman (2001), we found that the degree distribution of the networks in nanosciences are best fit with a power-law form with exponential cutoff

$$P(d) \sim d^{-\tau} e^{-d/d_c}, \qquad (2)$$

where τ and d_c are constants whose values are listed in Table 4.

To demonstrate explicitly that nodes form new links proportionally to the degree of the nodes, we calculate the

TABLE 4. 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

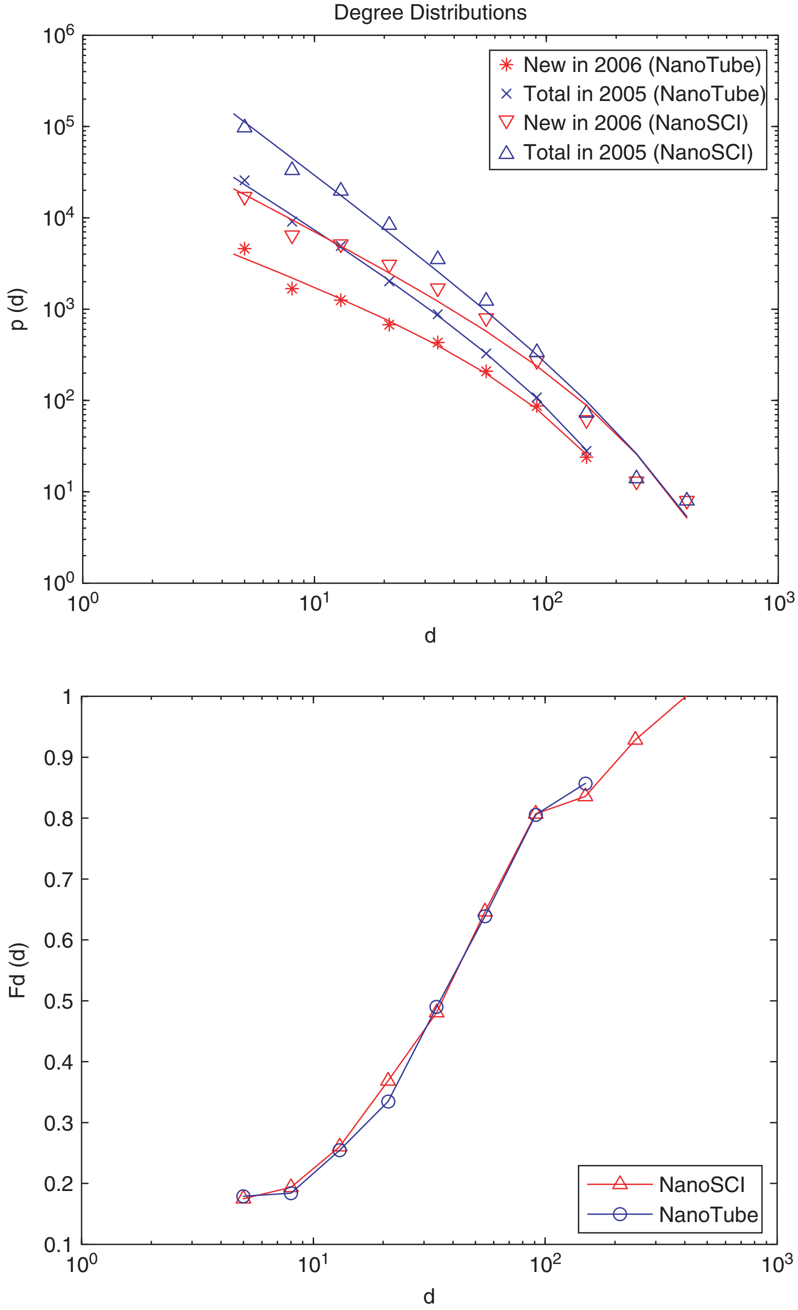

FIG. 4. Degree distributions of the NanoSCI and NanoTube networks for all nodes through the end of 2005 (triangles and crosses, respectively) and for nodes that occur in 2006 only (inverted triangles and asterick, respectively). Each set of symbols is fitted with a power law with exponential cutoff (see the text for discussion and consult Table 4 for the values of the parameters of the fit (a). Proportion of node $Fd(d)$ with certain degree attracting new edges in a new time step (b).

proportion $Fd(d)$, as a ratio between the number of nodes that form new edges at a certain step and the number of nodes with the same degree that existed at the previous time step. The results for the NanoSCI and NanoTube communities are shown in Figure 4b.

Based on these observations, we propose a novel hybrid network growing scheme that incorporates a locality element into global preferential attachment. We compare this scheme with three existing models: (a) the random growth model, (b) the preferential attachment model, and (c) the distance-based growth model. To make a fair comparison, we parameterize these models and make the edge densification rate identical to real nanotechnology-community growth rates. We next discuss the simulation results of these four network growth models and compare them with the observations from nanotechnology communities. In the following sections, we discuss the random growth model and the preferential attachment model, and then describe the distance-based model and the proposed hybrid model. Finally, we present analysis and insights based on the simulation results.

DOI: 10.1002/asi

A Random Growth Model

Callaway, Hopcroft, Kleinberg, Newman, and Strogatz (2001) proposed a simple random growth model where one new node and at most one new edge are added at a time. We slightly modify the model by parameterizing the number of the new edges as a function of the number of existing nodes. The number of existing nodes is equal to the time-stamp t; therefore, the number of new edges is denoted as $\Delta E(t)$. This model is referred as the *A-Random* model in this article.

The A-Random model involves a cycle of three steps:

1. At each time step, add one new node to the graph.
2. Randomly select two nodes and create an undirected edge between them.
3. Repeat Step 2 for $\Delta E(t)$ times.

At time t, there will be t vertices and on average $e(t) = \sum_t \Delta e(t)$ edges, where $e(t)$ is defined as $|E(t)|$. Let $c_k(t)$ be the expected number of vertices with degree k at time t. The number of isolated vertices $c_0(t)$ will increase by 1 at each time step, but decrease on average by $2\Delta e(t)\frac{c_0(t)}{t}$, the probability that a degree-zero vertex is randomly chosen as one of the ends of a new edge. Thus:

$$c_0(t+1) = c_0(t) + 1 - 2\Delta e(t+1)\frac{c_0(t)}{t} \tag{3}$$

Similarly, the expected number of degree k vertices ($k > 0$) will increase on average by an amount proportional to the probability that a degree $k-1$ vertex is chosen for attachment by a new edge, and will decrease by an amount proportional to the probability that a degree k vertex is chosen. This gives:

$$c_k(t+1) = c_k(t) + 2\Delta e(t+1)\frac{c_{k-1}(t+1)}{t} - 2\Delta e(t+1)\frac{c_k(t)}{t+1} \tag{4}$$

Note that the aforementioned equations neglect the possibility that an edge links a vertex to itself. This means that the equations are only approximate at short times, but they become exact in the limit $t- >\infty$ because the probability that any vertex is chosen twice decreases like t^{-2}.

Parameterized Preferential Attachment Model

This section describes a simple parameterized preferential attachment model (PPAM), in which a new vertex and l new edges $[l = \Delta e(t)]$ are added into the network at each time step. In building a new connection, we specify that (a) we randomly select a start node v_i, and (b) the probability that a node v_j is selected as the end node of the new edge is

$$p_t(v_j) = \frac{d'_t(v_j)}{\sum\limits_k d'_t(v_k)}$$

Where

$$d'_t(v_i) = d_t(v_i) + 1 \tag{5}$$

Thus,

$$\sum_{k=1}^{t} d'_t(v_k) = \sum_{k=1}^{t} d_t(v_k) + t = 2e(t) + t \tag{6}$$

Therefore, the likelihood of a node n_s connecting to another node n_e only depends on their degree. Note that by using $d'_t(v_i)$, the model allows zero-degree nodes to be selected as the end node.

$$c_0(t+1) = c_0(t) + 1 - \Delta e(t+1)\left(\frac{c_0(t)}{t} + \frac{c_0(t)}{2e(t)+t}\right) \tag{7}$$

The probability $SA_{k-1}(t+1)$ that a degree k $-$ 1 vertex is selected for attachment by a new edge at time $t+1$ is:

$$SA_{k-1}(t+1) = c_{k-1}(t)\left(\frac{1}{t} + \frac{k-1}{2e(t)+t}\right) \tag{8}$$

Similarly, the probability $SA_k(t+1)$ that a degree-k vertex is chosen for attachment by a new edge at time $t+1$ is:

$$SA_k(t+1) = c_k(t)\left(\frac{1}{t} + \frac{k}{2e(t)+t}\right) \tag{9}$$

Hence, the degree distribution is determined by

$$c_k(t+1) = c_k(t) + \Delta e(t+1)(SA_{k-1}(t+1) - SA_k(t+1)). \tag{10}$$

Degree-product-based PPAM (DP-PPAM). Barabasi et al. (2002) extended the preferential attachment model to take into account the degree product of both nodes in the network. Thus, we also compare our distance-based models to the following updated degree-product-based PPAM model. In particular, in building a new connection, we specify that (a) the start node v_i is selcted based on its degree; the probability is defined as $\frac{d'_i}{\sum_k d'_k}$; and (b) the probability that a node v_j is selected as the end node of the new edge is

$$p_t(v_j) = \frac{d'_t(v_j)}{\sum\limits_k d'_t(v_k)}$$

where

$$d'_t(v_i) = d_t(v_i) + 1 \tag{11}$$

DG model

This section describes a simple proximity-based growth model in which the likelihood of building a connection between two nodes depends only on their proximity. Note that the proximity between two individual nodes can be evaluated by a variety of measures, including shortest distance, CFEC (Koren et al., 2006). In this article, 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 two vertices of a new edge are determined in the following way: (a) one node n_s is selected uniformly from the graph as the start vertex of the new edge; and (b) the probability that a node v_p is selected as the end vertex of the new edge is:

$$p_t(v_p) = \frac{\frac{1}{r'_t(v_p, v_s)}}{\sum\limits_p \frac{1}{r'_t(v_p, v_s)}} \tag{12}$$

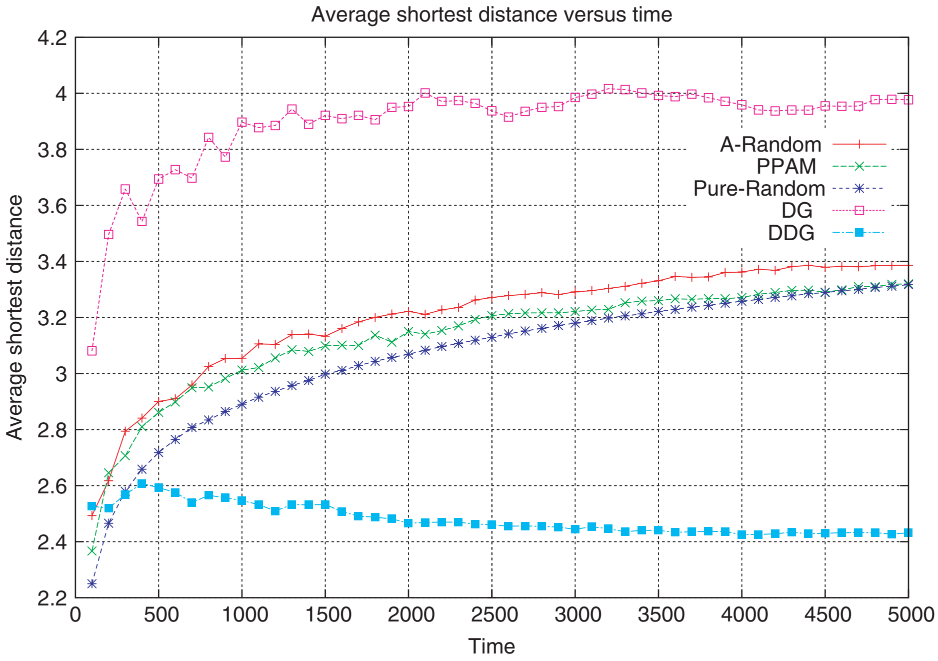

FIG. 5. The average shortest distance versus time. The *edge densification rate* is fixed at $\Delta\chi(t) = 3.39 * |V(t)|^{0.12}$

where

$$r'_t(v_p, v_s) = \begin{cases} R & \text{if } r_t(v_p, v_s) = \infty \\ r_t(v_p, v_s) & \text{otherwise.} \end{cases} \tag{13}$$

Equation 13 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 nonzero.

Hybrid DDG Model

This section describes a hybrid DD*G* model in which the distance and attachedness (i.e., degree) factors are combined to determine the likelihood that a node is selected to form a new edge. Similar to the aforementioned models, one new vertex and $l = \Delta E(t)$ edges are added at each time step in the graph growth process. The DDG model specifies that the start node v_s is selected randomly from the graph, but the probability that a node v_p is selected as the end node of the new edge is

$$p_t(v_p) = \frac{\frac{d'_t(v_p)}{r'_t(v_p, v_s)}}{\sum_p \frac{d'_t(v_p)}{r'_t(v_p, v_s)}} \tag{14}$$

where $r_t[p(v_p, v_s)]$ is defined by Equation 13 and $d_t(v_p)$ is defined by Equation 11.

Analysis and Simulations

To evaluate the proposed network growth model, we compare the simulation results of the hybrid model together with those of the other three models with observations regarding the NanoSCI, NanoParticle, and NanoTube communities. As a comparison, we also calculate the topological properties of pure random graphs with the same number of nodes and edges as a baseline to compare with other approaches. This baseline approach is denoted as *PureRandom* in this section. As a reminder, *A-Random* refers to the quasirandom approach described earlier; *PPAM* refers to the previously described preferential attachment model. In the rest of this section, we analyze and compare these network growth methods from two important perspectives: (a) the average shortest distance of the networks generated by these models over time and (b) the average clustering coefficient of these networks.

Temporal distance analysis. We now compare the average shortest distance between the different models described previously in this section. Figures 5, 6, 7, and 8 illustrate the results of a set of experiments obtained by varying the densification rates and growth dynamics. In particular, Figure 5 shows the average shortest distance versus time using the same node growth rates $(\Delta|V(t)|)$ and the same edge densification rate $(\Delta\chi(t) = 3.39 * |V(t)|^{0.12})$. Figure 7 shows the temporal patterns of average shortest distance for *A-Random* approaches at different edge densification rates $\Delta\chi(t) = \{0.02{*}|V(t)|^{0.69};\ 0.02{*}|V(t)|^{0.6};\ 3.39{*}|V(t)|^{0.12};\ 0.02{*}|V(t)|^{0.5}\}$, respectively. Figure 7 indicates that the kinetic properties of the average shortest distance metric are closely related to the ratio of edge growth versus node growth. When the edge densification rate is $\Delta\chi(t) = 3.39{*}|V(t)|^{0.12}$, the average shortest distances of *A-Random* increase slowly over time. However, when the edge densification rate is $\Delta\chi(t) = 0.02{*}|V(t)|^{0.69}$, the average distance for *A-Random* decreases over time after a sharp increase in the early stage of network growth. Similar results also are observed for other growth models. In general, when the edge

DOI: 10.1002/asi

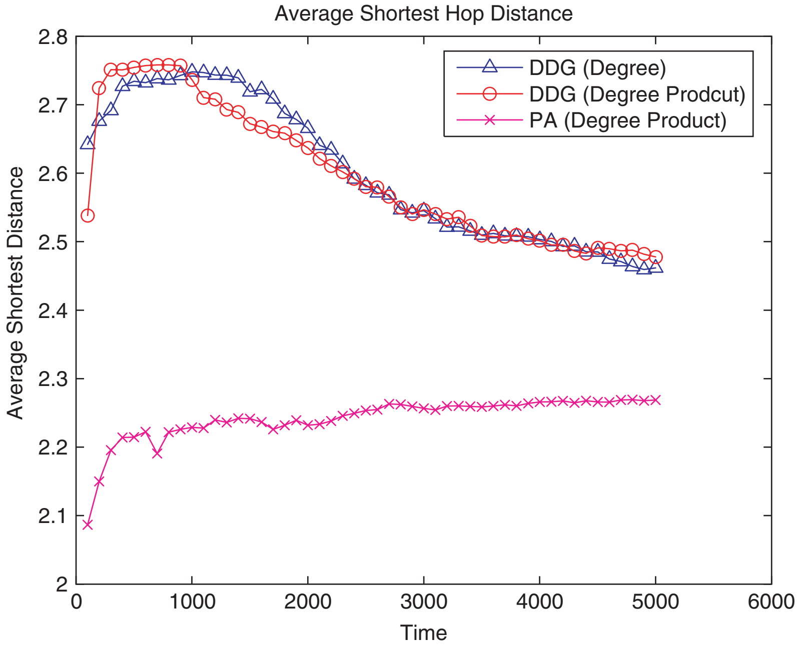

FIG. 6. The average shortest distance versus time simulation results for three models; triangles mark the result for the DDG model, the circles denote the result for DDG model with degree-product, and the asterisks mark the degree-based preferential attachment model.

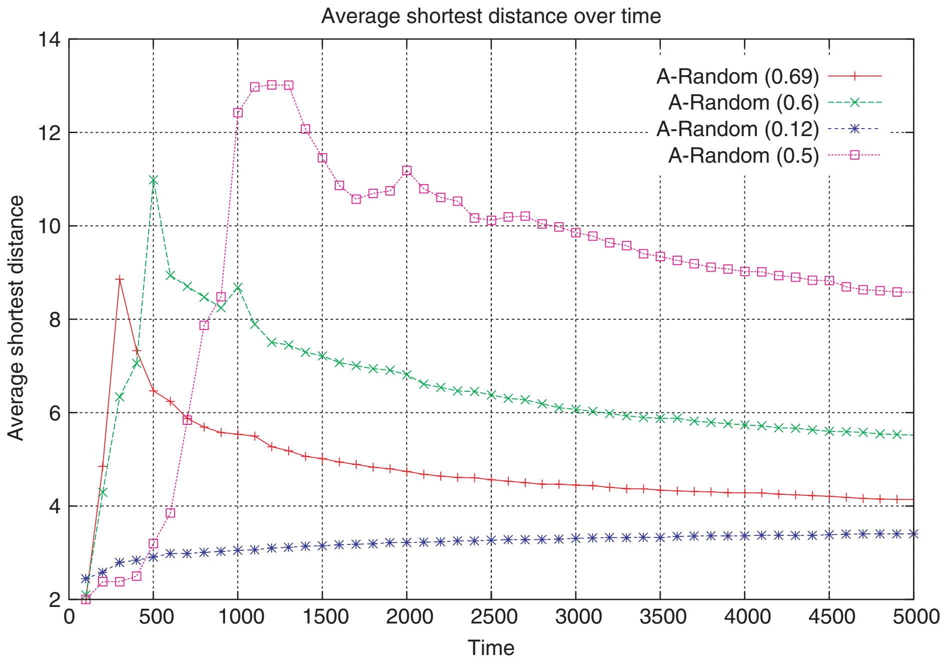

FIG. 7. The average shortest distance versus time $\{\Delta\chi(t) = [0.02 * |V(t)|^{0.69}; 0.02 * |V(t)|^{0.6}; 3.39 * |V(t)|^{0.12}; 0.02 * |V(t)|^{0.5}]$ respectively$\}$.

density is higher, the average shortest distance for these models at a particular time point is smaller. Figure 5, however, shows an interesting result from the perspective of modeling the growth dynamics of the nanotechnology community. By adopting the edge densification rates of the NanoSCI community (i.e, $\Delta\chi(t) = 3.39^{*}|V(t)|^{0.12}$), the simulation results of these models resulted in rather different growth behavior.

The average shortest distance of the network generated by the hybrid DDG model decreases over time after an early stage increase, which is similar to what we observed in the actual nanotechnology community (see Figure 2). In contrast, the average shortest distance of the networks generated by the random growth models and the PPAM increases over time. Results for a modification of the PPAM, which assumes that two nodes connect with each other proportionally to the product of their degrees, are plotted in Figure 6. The average shortest distance first increases and then keeps almost constant value similar to the the average shortest distance of the local distance-based model (DG), which increases first then oscillates around a convergence point. The average shortest

DOI: 10.1002/asi

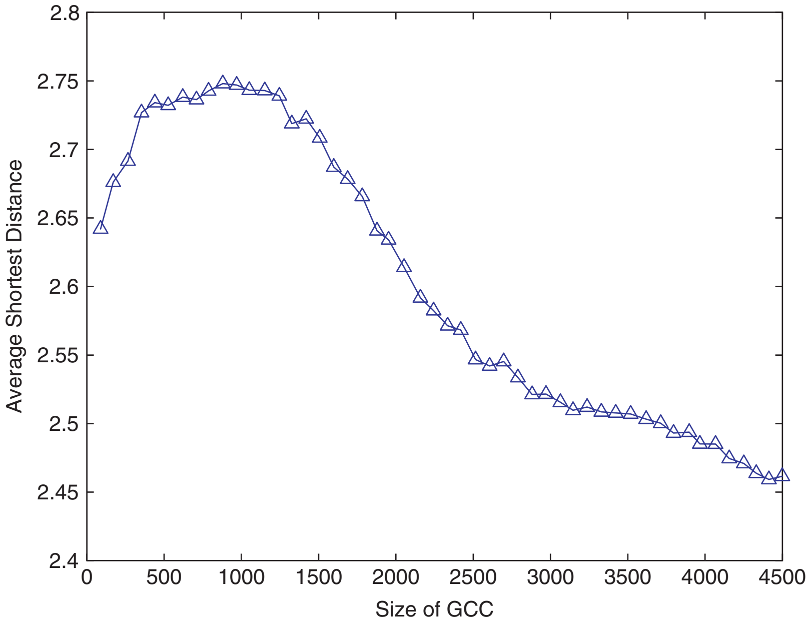

FIG. 8. The average shortest distance versus GCC size for DDG-degree algorithm.

TABLE 5. Average clustering coefficient values for proposed computational models. The clustering coefficient values are the average of clustering coefficient values for snapshots obtained over 5,000 time cycles.

Dataset	A-Random	DG	PPAM	DDG
$\overline{\aleph}(G(t))$	0.02	0.05	0.03	0.45

distance produced by the DDG model better describes the observations than does the preferential attachment degree product model. Figure 8 demonstrates that the average shortest distance first increases and then decreases with increasing the size of the GCC. We also noted that the effect of adding a local factor into the global attachment-based model reduces the average shortest distance while the purely locality-based model results in a larger average shortest distance than does the global attachment-based model. This growth behavior suggests that there is a synergistic effect in proximity and attachedness factors.

Clustering coefficient analysis. A node's clustering coefficient measures the connectivity among this node's neighbors. Large cluster coefficient values indicate that the neighbors of the node in question are well connected to each other. Table 5 shows the clustering coefficient values for networks generated by different models using the edge densification rate of the NanoSCI community (i.e., $\Delta\chi(t) = 3.39 * |V(t)|^{0.12}$). This table indicates that the hybrid DDG model clearly has significantly higher clustering coefficient values than the preferential attachment, the *A-Random* growth, and the local distance-based models. More important, the simulation results suggest that the hybrid DDG model generates networks with a clustering coefficient that is much closer to those of the collaborative network of the nanotechnology communities, shown in Table 3.

In conclusion, based on the simulation results, we observe that the hybrid DDG model is able to produce networks with clustering coefficient values closer to what was observed in the nanotechnology community. Furthermore, this model generates networks whose average shortest distance decreases over time when the edge densification rate of the NanoSCI community is used. Hence, it is more suitable as a model for the collaborative network of the nanotechnology community than are either the global preferential attachment model or the local distance-based model.

Conclusions and Future Work

The explosive development of nanotechnology research calls for in-depth scientometric study of this field. In this article, we conducted a case study on the evolving nanotechnology collaboration networks and concluded that both locality and attachedness play significant roles in the social network growth process. In particular, this article expands the definition of temporal social networks and demonstrates that a science-based collaboration network is similar to other real-world social networks. Furthermore, the nanotechnology collaboration networks studied exhibit an intriguing set of static and dynamic properties. Inspired by the observations that in collaboration networks new connections tend to be augmented between nodes in proximity, we explored both locality and attachedness factors in growing networks and proposed two distance-based computational growth schemes, namely DG and DDG. The DG model considers only the locality element while the DDG is a hybrid model that factors into both locality and attachedness elements. We discovered that the dynamic patterns of average shortest distance and clustering coefficient are closely connected to the edge densification rates as well as specific growth mechanisms. In addition, we discovered that when we use when the edge densification rate of the NanoSCI community, (a) clustering coefficient rates of the DDG model were closer to those of the nanotechnology community, and (b) the DDG model exhibited a decreasing average shortest distance phenomenon, which also was observed in the collaborative network of the nanotechnology community. These simulation results suggest that the hybrid approach that combines locality and attachedness can better characterized the growth of the nanotechnology community.

The results of this study also inspire us to investigate related questions in our future work. For instance, what are the general characteristics of social networks that are best characterized by a hybrid network growth model? How can variations in hybrid growth models be compared to each other? What insights can be obtained from different hybrid approaches to model network growth? Future research that addresses these and other related questions may not only improve our understanding about the dynamic behavior of network growth but also lay the foundation for providing deeper insights on social network analysis.

References

Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74, 47–97.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509.

Barabasi, A.-L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. PHYSICA A, 311, 3.

Bollobs, B. (2001). Random graphs (2nd ed.). Cambridge University Press.

Callaway, D.S., Hopcroft, J.E., Kleinberg, J.M., Newman, M.E.J., & Strogatz, S.H. (2001). Are randomly grown graphs really random? Physical Review E, 64, 041902. Available at: http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0104546

Davidsen, J., Ebel, H., & Bornholdt, S. (2002). Emergence of a small world from local interactions: Modeling acquaintance networks. Physical Review Letters, 88, 128701.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. (2005). Team assembly mechanisms determine collaboration network structure and team performance. Science, 297 (5586), 1551–1555.

Jin, E.M., Girvan, M., & Newman, M.E.J. (2001). The structure of growing social networks. Physical Review E, 64, 046132.

Jost, J., & Joy, M.P. (2002). Evolving networks with distance preferences. Physical Review E, 66, 036126.1–036126.7.

Koren, Y., North, S.C., & Volinsky, C. (2006). Measuring and extracting proximity in networks. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 245–255). New York: ACM Press.

Kossinets, G., & Watts, D.J. (2006). Empirical analysis of an evolving social network. Science, 311, 88–90.

Kostoff, R.N., Stump, J.A., Johnson, D., Murday, J.S., Lau, C.G., & Tolles, W.M. (2006). The structure and infrastructure of the global nanotechnology literature. Journal of Nanoparticle Research, 8(3/4), 301–321.

Krapivsky, L., & Redner, S. (2001). Organization of growing random networks. Physical Review E, 63, 066123.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 611–617). New York: ACM Press.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. (2000). The web as a graph. In Symposium on Principles of Database Systems (PODS)(pp. 1–10). New York: ACM Press.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (pp. 177–187). New York: ACM Press.

Liben-Nowell, D., & Kleinberg, J. (2007). The link prediction problem for social networks. Journal of the American Society for Information Science and Technology, 58(7), 1019–1031.

Morris, S.A., & Goldstein, M.L. (2007). Manifestation of research teams in journal literature: A growth model of papers, authors, collaboration, coauthorship, weak ties and lotka's law. Journal of the American Society for Information Science and Technology, 58(12), 1764–1782.

Newman, M.E. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, USA, 98, 404–409.

Newman, M.E. (2004, April). Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences, USA, 101(Suppl. 1), 5200–5205.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., & Barabasi, A.-L. (2002, August). Hierarchical organization of modularity in metabolic networks. Science, 297(5586), 1551–1555.

Vazquez, A. (2000). Disordered Networks generated by recursive searchers. Europhysics Letters, 54(4), 430–435.

Watts, D.J., Dodds, P.S., & Newman, M.E.J. (2002). Identity and search in social networks. Science, 296, 1302–1305.

Xuan, Q., Li, Y., & Wu, T.-J. (2007, May). A local-world network model based on inter-node correlation degree. Physica A Statistical Mechanics and Its Applications, 378, 561–572.

DOI: 10.1002/asi