Modeling Idiosyncratic Properties of Collaboration Networks Revisited

Ergin Elmacioglu

Department of Computer Science and Engineering, The Pennsylvania State University, PA 16802 U.S.A.

Phone: (814) 865-9505 Fax: (814) 865-3176 E-mail: elmaciog@cse.psu.edu

Dongwon Lee

College of Information Sciences and Technology, The Pennsylvania State University, PA 16802 U.S.A.

Phone: (814) 865-0687 Fax: (814) 865-6426 E-mail: dongwon@psu.edu

ABSTRACT

A study on the network characteristics of two collaboration networks constructed from the ACM and DBLP digital libraries is presented. Different types of generic network models and several examples are reviewed and experimented on re-generating the collaboration networks. The results reveal that while these models can generate the power-law degree distribution sufficiently well, they are not able to capture the other two important dynamic metrics: average distance and clustering coefficient. While all current models result in small average distances, none shows the same tendency as the real networks do. Furthermore all models seem blind to generating large clustering coefficients. To remedy these shortcomings, we propose a new model with promising results. We get closer values for the dynamic measures while having the degree distribution still power-law by having link addition probabilities change over time, and link attachment happen within local neighborhood only or globally, as seen in the two collaboration networks.

INTRODUCTION

Social network analysis is an active research field in social sciences where researchers try to understand social influence and groupings among a set of people or groups. Its origin is in general believed to be due to Milgram (1967) who identified the so-called "*six degrees of separation*" phenomenon– any two people in the United States are connected through about six intermediate acquaintances, implying we live in a rather *small-world*. Since then, researchers have found evidences for a wide range of small-world phenomena arising in other social and physical networks, also known as "complex networks" (e.g., power grids, airline time tables, food chain, World-Wide Web, acquaintance network, Internet backbone).

As the data for these networks become more readily available and having more computational power, research interest has moved to the analysis of network structure. Why are people interested in the structure of such networks? Among various reasons, it is mainly because the structure of networks always affects processes occurring in the networks. For instance, the topology of the social network of interactions may affect the spread of information or disease, and the topology of the Internet may give helpful clues about the robustness, stability or cost of communication among computers world-wide. Since most real-world networks are very large in scale, analyzing their topology may reveal important properties, without looking at the actual system or dealing with complexities associated with those systems.

The analysis of the complex networks has also drawn attention to modeling of these networks. If one could come up with a good model, the entire network can be re-generated or the future status of the network can be predicted. Inspired by some of the recent attempts to apply social network analysis to scientific communities (e.g., Newman, 2001; Barabasi et al, 2002), in this article, we analyze two collaboration networks in Computing community. In particular, we are interested in finding if any existing models can re-generate two collaboration networks in precision, and if not, in improving them to do so.

Our contributions in this paper are: (1) Using two related but distinct data sets, we show that none of the currently-available generative models can correctly capture three idiosyncratic properties of collaboration network (i.e., power-law distribution, small average distance, and large clustering coefficient); and (2) By extending the preferential attachment model (Barabasi & Albert, 1999), we show how to generate synthetic networks that exhibit the three properties better than current models. Table 1 lists common terms and definitions used throughout this paper.

TABLE 1. Common terms and definitions.

Term	Definition
N	set of all nodes
N(*t*)	set of all nodes at time t
k	number of links connected to a node (i.e., degree)
k_i	degree of the node i
$k_i(t)$	degree of the node i at time t
P(i, t)	probability that node i gets a link at time t
$P(k_i, t)$	probability that a node with degree k gets a link at time t
γ(i)	clustering coefficient of node i
old node	a node that existed in a network at time t
new node	a node that joins to a network for the first time at time t
N_k	set of all nodes of degree ≤ k
geodesic	average distance between all nodes
internal link	a new link between two old nodes
external link	a new link between a new and an old nodes
m	number of links created by a new node

COLLABORATION NETWORKS

The *collaboration network* (or graph) consists of *nodes* that represent authors, and *edges* that represent co-authorship among authors. For our study, we used two well-known data sets: (1) ACM Digital Library (hereafter *ACM-DL*) has a comprehensive citation data on general computing literatures. It is automatically generated by extracting metadata from actual articles; and (2) *DBLP-DB* contains citation data on “Database, a sub-topic of the whole Computing discipline” and is a subset of the well-known DBLP data set. Its coverage is lower than ACM-DL since it covers only Database area. However, its quality is higher than ACM-DL since DBLP data set itself is manually curated by human editors. Two data sets thus represent two distinct cases – comprehensive data set with a lower quality and small but cohesive data set with a higher quality.

ACM-DL data was generated by crawling ACM Guide[1] (in January 2005) while DBLP-DB was taken from the study by Elmacioglu & Lee, 2005. For ACM-DL, we used data from 1950 to 2004, which contained 609,202 authors and 769,161 publications. Similarly, for DBLP-DB, we used data from 1968 to 2003, yielding 32,689

[1] http://portal.acm.org/guide/

authors and 38,773 papers. We built a *collaboration network* where nodes represent authors and edges represent co-authorship. Note that these data sets do not have a notion of "unique key" such as Digital Object Identifier (Atkins et al., 2000) or ISBN. Instead, they depend on the name of authors as a distinguisher. Therefore, the classical *name authority control problem* (Hong et al, 2004) may arise (i.e., same author with various spellings or different authors with the same spelling). We try to minimize this problem by conducting two experiments – one with full names (e.g., "John Doe") and the other with the first initial followed by the last name (e.g., "J. Doe"), and use these results as the upper and lower bounds of the statistics (Newman, 2001). Since the results of the both cases were similar, here, we only present the results with full names. The basic statistics related to the authors and papers of both ACM-DL and DBLP-DB are summarized in Table 2. These are cumulative statistics measured by considering the entire period. For most studies below, both data sets exhibit similar results, and therefore in the interest of space, we mainly present using the result of ACM-DL. When they show disparate results, however, we present both.

TABLE 2. Cumulative statistics for the collaboration data as of 2004.

Entry	ACM-DL	DBLP-DB
mean papers per author	2.56	2.69
mean authors per paper	2.54	2.30
collaborators per author	3.66	3.93
total # of papers	769,161	38,773
total # of nodes (i.e., authors)	609,202	32,689
total # of edges	1,116,308	64,284
# of nodes in giant component (percentage)	346,137 (57%)	18,542 (57%)
# of nodes in 2nd largest component	66	57
clustering coefficient	0.58	0.63
geodesic	7.98	6.17
diameter	33	20

NETWORK CHARACTERISTICS

In recent years, studies of complex networks such as the Internet, WWW, or social networks have put a lot of efforts into revealing the structure and dynamics of such systems. Although there are a large number of varieties, many real-world complex networks share three important characteristics (Albert & Barabasi, 2002; Dorogovtsev & Mendes, 2002): (1) power-law degree distribution, (2) small geodesic (i.e., average distance among nodes), and (3) high clustering coefficient. In the following section, we present the results of these properties for both ACM-DL and DBLP-DB collaboration networks. Since a collaboration network may generate numerous components, we selected the largest connected component, called a giant component, to investigate these characteristics of complex system.

Degree distribution

The number of links that a node has is called "degree" of the node. Then the degree distribution $P(k)$ is the probability that a randomly selected node has k links. Until recent years, complex systems were modeled using simple random network theory in which the degree distribution is binomial or Poisson in the limit of large graph size. However, recent studies showed that many complex systems in fact follow the power-law degree distribution (Albert & Barabasi, 2002; Newman, 2003). Power law is defined as "the probability of measuring a particular value of some quantity varies inversely as a power of that value" (Newman, 2005), and the probability function is in the form of $P(k) \sim k^{-\alpha}$.

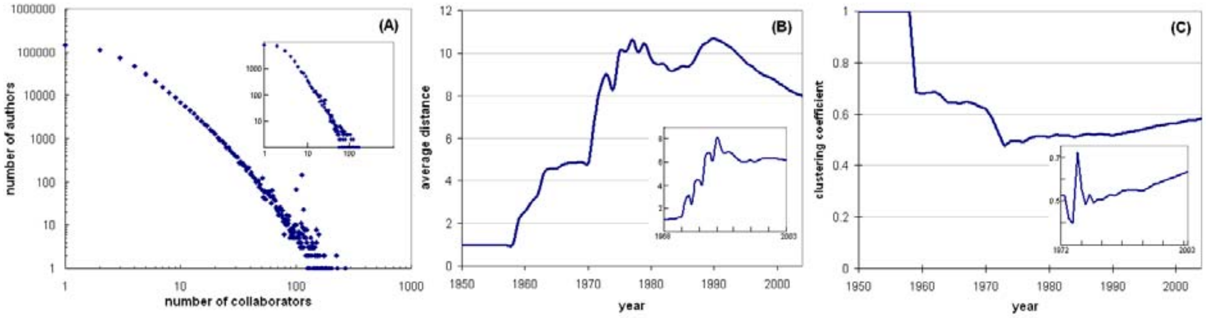

Figure 1. Static and dynamic properties of the ACM-DL Collaboration Network. Insets: the equivalent graphs for the DBLP-DB Collaboration Network.

Figure 1a shows the degree distribution of the collaboration network of ACM-DL. The horizontal axis represents degree k, while the vertical axis shows how many authors exist for each value of k. Although there is some curvative for the small values of k, the best fit for the degree distribution in the middle region and tail is power law with the exponent $\alpha = 2.84$. Consistent with the definition of power law, there are many authors in the community who has only one collaborator (link), while there are a small number of authors who have many collaborators. It was suggested that the curvative region of the degree distribution in such collaboration networks might be a signature of exponential cut off due to using a finite time frame of study for analysis (Newman, 2001).

Degree is a local quantity and characterizes individual nodes in a network. The higher value of this measure may imply higher value of importance or quality than other nodes with fewer links. Distribution of degree, in turn, quantifies diversity of nodes, spread of how many nodes are available in each importance level in a network. This may help us understand the network, i.e., we can see if nodes are mostly typical or heterogeneous; how many of them may have the key role in the network? Hence, a good model should be able to regenerate a given real-world network by creating a closer number of nodes with each degree value available in the network, in other words, the distribution generated should be a closer fit to the degree distribution curve of the actual network. Although degree is a local measure, its distribution often determines other important global characteristics in complex networks as well.

Average distance

In a collaboration network, the path with the minimum number of edges between any given pair of authors is called shortest path or geodesic of the pair. Then, the average distance in a network is the average of all pair-wise geodesics of nodes in the network. Social networks tend to have small average distances compared to the number of nodes in the networks, first described by Milgram (1967) and now referred to as "small-world effect." Figure 1b shows the evolution of the average distances in the giant component of the ACM-DL.

Each newly added link, a new collaboration in the graph, affects the average number of links between authors. There are two different kind of new links: (1) external links which are formed by an addition of a new author who collaborates first time with an existing author, or (2) internal links which are created by the collaboration of two existing authors who have not collaborated before. External links are a significant factor to increase the average distance by creating new paths to incoming authors, whereas internal links act inversely and decrease the distance among authors by creating new paths (possibly short-cuts) between existing authors.

In the first 40 years of the period analyzed, the geodesic tends to increase each year with occasional fluctuations. The reason is clear that the network tries to form into a larger one with the high rate of addition of new authors and merges of small clusters to the giant component. After it finally reaches its maximum value of 10.7 in 1990, it shows continuously decreasing pattern because the increased tendency for collaboration creates new shorter paths between existing authors which prevails the effect of expansion due to the new authors. The final value of the geodesic in 2004 is 7.9 and seems to be decreasing in the following years. Compared to the size of the community, this relatively low value is probably a good sign since scientific discoveries can be disseminated rather fast (Newman, 2001). The diameter of a graph, the maximum of the pair-wise distances in the giant component, of the computing community is 33 as of 2004.

Though average distance is a simple measure, it has important implications for the dynamics of processes taking place on complex networks. In a contact network, for instance, a small geodesic implies that a rumor or a disease will spread very fast over the network, and affect a large portion of the network in a short time. Similarly, it affects how fast new ideas or inventions will disseminate and be used by others in a collaboration network, or how many hops it takes for the communication of two computers on the Internet, and so forth (Newman, 2003).

Clustering coefficient

Given a node *v*, the *neighborhood* of *v*, *Adj(v)*, is a subgraph that consists of the nodes adjacent to the node *v*. Furthermore, let us denote the edges and nodes in *Adj(v)* by *E(Adj(v))* and *V(Adj(v))* , respectively. Then, the clustering coefficient of *v*, $\gamma(v)$, is defined as $\gamma(v) = \frac{|E(Adj(v))|}{|E\max(Adj(v))|}$ where $|E\max(Adj(v))| = \frac{|V(Adj(v))|(|V(Adj(v))|-1)}{2}$ edges. Therefore, the clustering coefficient measures how many edges actually occur compared to the fully-connected case (Watts, 1999). The clustering coefficient of a graph *G*, $\gamma(G)$, is the average clustering coefficients of all nodes in *G*.

The clustering coefficient can be also viewed as "transitivity" which indicates the interactions among trios of nodes in a network (Newman, 2001) – the degree to which a scholar's collaborators have collaborated with each other. In co-authorship networks, this measure implies how much authors are willing to collaborate with each other.

The clustering coefficient of the giant component in the ACM-DL is shown in Figure 1c as a function of year. First few years result in a fully connected graph since the size of the components is small and everyone has acquaintance with each other. Over the following years, the clustering coefficient tends to decrease as the giant component expands. Starting from 1974, when the giant component becomes relatively larger, the clustering coefficient shows a steady increasing pattern until the end of the period analyzed. As of 2004, it is about 0.6. This rather high value of the clustering coefficient is expected in such a collaboration network or many other complex systems, compared to a similar random graph.

This is another simple yet important measure for characterizing a network. It defines how well a node is connected to its first degree neighbors. In a social network, higher clustering coefficient may imply individuals know each other very well and have a strong relationship with their direct contacts. Similarly, in a collaboration network, it might be an indicator of how strongly a group of researchers collaborates with each other, and shares the same interest and research subjects. It could also be regarded as a measure of network resilience in physical networks, i.e., what is the possibility that a computer can communicate with a neighbor even if the link between them fails? If the clustering around the computer is dense, there would be alternative paths to the destination.

NETWORK MODELS

A variety of network models proposed for social and other kinds of complex systems roughly follow three main types: *random networks*, *small-world networks* and finally *scale-free networks*.

Random Network Type

In a random network type, edges between nodes are distributed randomly. The first study is introduced by Erdos and Renyi (1959) - a graph with N nodes are connected by edges randomly chosen from N(N-1)/2 edges. Random networks allow us to create and study properties of real-world networks easily including small-world effect by generating small average distance. However, resulting degree distribution is Poisson for large N - majority of the nodes have the same average degree while small number of nodes have a few or many nodes. This is not sufficient to accurately model systems seen in the real world with power-law degree distribution. It is, however, possible to use general random network theory with an input parameter of desired degree distribution and define a semi-random network model to generate the target system (Albert & Barabasi, 2002).

Small-World Network Type

Another important characteristic of real world networks is their unusually large clustering co-efficient, regardless of the network size. Small-World network model, introduced by Watts and Strogatz (1998), assumes that real-world networks show some regularity along with randomness. In order to achieve this, the model starts with regular K-lattice, and rewires each link with a certain probability p in which one endpoint is kept and the other endpoint is connected to one other node with the probability p. As the rewiring probability increases, the network becomes more random. Starting with a regular network keeps clustering co-efficient high, while random rewiring process decreases average distance through shortcuts between nodes.

Scale-Free Network Type

The origin of scale-free network types was first addressed by Barabasi and Albert (1999). Their study considers dynamic growth of networks and proposes the so-called preferential attachment mechanism for link additions. According to this model in its basic form, a network starts with a small number of nodes and grows over time by addition of new nodes. A newly arriving node connects to the existing nodes but the attachment process is not entirely random. Each existing node can gather new links with a probability that increases linearly with the number of links it already has. Hence the more links a node has, the better chance it can obtain new links. This is the situation seen in most of the real-world networks. The rich or important entities (nodes with many links) usually have more power and become richer in time. Later, a large variety of improvements to this generic model have been proposed to get better fits for numerous systems by incorporating different growth constraints and identifying system-specific mechanisms for preferential attachment (Albert and Barabasi, 2002).

TABLE 3. Summary of all models experimented.

Model	Type	Network partition	Operation
ER	static	Y	all nodes are added initially, internal link addition only
WS	static	Y	all nodes are added initially, internal link addition & rewiring
Waxman	static	Y	all nodes are added initially, internal link addition only
GE	dynamic	N	node birth with m links
BA	dynamic	N	node birth with m links
GLP	dynamic	N	node birth with m links, internal link addition
PG	dynamic	Y	node birth without links, internal link addition only
AB	dynamic	Y	node birth with m links, internal link addition & rewiring
Fitness	dynamic	N	node birth with m links
Amaral	dynamic	N	node birth with m links, inactivation of old nodes
DM	dynamic	N	node birth with m links
YJB	dynamic	N	node birth with m links

Twelve Models experimented

In the analysis, we selected twelve models that belong to one of the three types – random, small-world and scale-free networks. Furthermore, each model can also be categorized as either *static* or *dynamic*. There is no network growth in static models while dynamic models start with a few number of nodes and gradually grows by addition of new nodes over time. For the static measure degree distribution, cumulative data as of 1995 is used, whereas for the dynamic measure average distance and clustering coefficient, the period from 1978 to 1990 is divided into 9 equal intervals and measured separately. At the end, the collaboration network of ACM-DL has 110,000 nodes and about 250,000 links. Table 3 summarizes all models that we experimented and their properties.

ER (Erdos – Renyi) Model

This model generates a static random graph based on Erdos-Renyi model (Erdos & Renyi, 1959). All nodes N are added initially, and then links are added between each pair based on uniform probability p, which is chosed by:

$$p = \frac{\text{expected \# of links}}{N(N-1)/2}$$

This method often partitions the graph into several disconnected components, and generates most nodes with roughly the same degree.

WS (Watts-Strogatz) Model

The basic, static small world model is used (Watts & Strogatz, 1998). All nodes are added to the system initially. N nodes are arranged on a ring d-lattice where each node is connected to its d neighbors (d/2 on either side). Each link then is randomly rewired with probability p. This process introduces pNd/2 long range edges which connect nodes that otherwise would be part of different neighbors in order to achieve small average distance. By varying p, one can observe a transition between regularity and randomness. We used d=2 and p=0.6.

Waxman Model

This is one of the first distance based models to generate Internet-like graphs (Waxman, 1988). Each node is uniformly distributed on a two dimensional plane. The probability of links between any two nodes is inversely proportional the Euclidean distance between them. We assume such a physical distance may also be regarded as social distance. The more two scientists are close to each other (being in the same institution or working on the similar subject) the more likely that both would collaborate. The edge probability between any two nodes (i, j) is given by

$$P(i, j) = \beta \cdot \exp\frac{-d(i, j)}{L\alpha}$$

where d(i,j) is the distance between nodes i and j, L is the diameter and α and β are parameters in range (0, 1] defining edge densities.

GE (Growing Exponential) Model

This model is a dynamic random network model. Network starts with a small number of initial nodes n_0. In each time step, one new node enters the network and creates m external links to old nodes which have been already in the system. An existing node *i* is selected using the uniform probability:

$$P(i, t+\Delta t) = \frac{1}{N(t)}$$

where N(t) is the number of nodes available in the system at time t. We chose m=1.35 which is the actual average value in the collaboration network. Due to the uniform attachment of links to the existing nodes, the older a node is the more chance it has to get links from incoming nodes (Park et al., 2004).

BA (Barabasi-Albert) Model

The BA model (Barabasi & Albert, 1999) is the original proposal for scale-free models and based on preferential attachment for nodes selection rather than random. It is a dynamic model, so one node joins to the network at each time step t, creating m external links to the existing nodes using the preferential attachment phenomenon (m is chosen 1.35 in the experiment). The probability that existing node i is chosen is proportional to the links it already has:

$$P(k_i, t+\Delta t) = \frac{k_i(t)}{\sum_j k_j(t)}$$

where $k_i(t)$ represents the degree of node i at time t, and the denominator is simply sum of degrees of all nodes available in the system at time t. While it keeps the simplicity, this model can capture the actual dynamics and topology of networks with power-law degree distributions. However, it is its basic form and not flexible enough to fit different power-law exponents, which has been led to many extensions and variations.

GLP (Generalized linear performance) Model

GLP is an extension of BA model. The original model is modified to fit better to some power-law exponents, i.e. Internet topology. The model is again dynamic, but has two link addition operations unlike in the generic BA model: (i) with probability p, m internal links are added between existing node pairs. For each endpoint of a links, the following attachment probability is used:

$$P(k_i, t+\Delta t) = \frac{k_i(t)-\beta}{\sum_j (k_j(t)-\beta)}$$, where $-\infty < \beta < 1$ (Bu & Towsley, 2002).

(ii) with probability 1-p, one new node and m external links are added from the new node to the existing nodes with the same attachment probability above. We used parameters as m=1.35 and β=0.712.

PG (Pretty Good) Model

PG (Pennock et al., 2002) is also an extension of the generic BA model. However, it also uses uniform attachment along with BA's strict preferential attachment. New nodes keep coming but do not generate external links. At each time step t, only internal links are created between existing nodes using the following attachment:

$$P(k_i, t+\Delta t) = \alpha \frac{k_i(t)}{\sum_j k_j(t)} + (1-\alpha)\frac{1}{|N(t)|}$$

The uniform attachment part gives additional flexibility to the model to fit different power-law exponents. Parameter m is 1.35 for the preferential attachment part and α is 0.5.

AB (Albert-Barabasi) Model

AB model is the authors' own extension to their original BA model (Albert & Barabasi, 2000). In addition to external links, internal links and rewiring of existing links are also employed. The attachment rule follows the following probability:

$$P(k_i, t+\Delta t) = \frac{k_i(t)+1}{\sum_j (k_j(t)+1)}$$

In this model three operations are available: (i) with probability p, m internal links are added. One node is selected randomly and other endpoint is selected using the attachment rule. (ii) with probability q, m links are rewired. One node, and one of the links of that node is selected randomly and using the attachment rule it is reconnected a node. (iii) with probability 1-p-q, one new node and m external links are added using the attachment rule. We did not use rewiring for generating our collaboration network since it is not an event happening in collaboration networks; and set p=0.421 and m=1.35.

Fitness Model

This model is another extension to the BA model assuming there is also a competition among nodes (Bianconi & Barabasi, 2001). In the original idea of BA, the older nodes tend to gain more links compared to new ones since they have more chance to get links. However, certain nodes even if they are new, may be very proliferate and get many links in a short time. Fitness model also considers this aspect of real world networks and combines nodes' individual ability to get links with preferential attachment. In this model each node is assigned a fitness parameter n_i which does not change in time. This parameter is chosen from a uniform distribution between 0 and 1. Then the attachment probability for incoming nodes changes to:

$$P(k_i, t+\Delta t) = \frac{n_i k_i(t)}{\sum_j (n_j k_j(t))}$$

Amaral Model

Amaral et al. (2000) explicitly introduces aging and cost (or capacity) into the original BA model. They argue that as nodes in many real-world networks get older, they are unable to get new links after a point. In other sense, as they get some number of links it may be hard for them to get more new links since they aged, or their capacity has been filled. In their proposal, they use the idea of active and inactive nodes. At any given time, there are both active and inactive nodes in the system, but only active nodes can get new links from incoming new nodes. Moreover at each time step, with a constant probability p, each active node may become inactive and can no longer get more links. This constraint can also be applied as a cost issue. As nodes can reach a certain number of edges, they can not make new connections. In the experiment, 10% of the existing nodes become inactive in each time step.

DM (Dorogovtsev-Mendes) Model

This is another model that considers aging, gradually. The probability that a new node connects to existing node k depends on the degree of k as well as its age, decaying exponentially with v:

$$P(k_i, t+\Delta t) = \frac{k_i(t)\cdot(t-t_i)^{-v}}{\sum_j (k_j(t)\cdot(t-t_j)^{-v})}$$

where v is a tunable parameter, and chosen as 0.5 in the experiment. They show that power-law scaling in degree distribution is available only for when v<1 (Dorogovtsev & Mendes, 2000).

YJB (Yook-Jeong-Barabasi) Model

This model is another distance based model particularly for Internet-like graphs in which the placement of links is driven by a competition between preferential attachment and linear distance dependence between nodes. In this model, the probability that a new node links to a node with k_i links at distance d_i from the new node is

$$P(k_i, d_i, t+\Delta t) = \frac{k_i(t)^{\alpha} / d_i^{\sigma}}{\sum_j (k_j(t)^{\alpha} / d_j^{\sigma})}$$

where α and σ are pre-assigned exponents, governing preferential attachment and the cost of the node-node distance. Increasing α will favor linking to nodes with higher degree, while a higher σ will discourage long links (Yook, Jeong and Barabasi, 2002). As in Waxman model, we regard distance as the social distance between two individuals.

ANALYSIS

In this section, we compare all the models experimented for generating the collaboration network of ACM-DL according to the three basic characteristics of the two collaboration networks. We consider degree distribution as a static metric since it does not change in time; hence the final snapshot of the collaboration network is used to assess this metric. On the other hand, average distance and clustering coefficient metrics tend to evolve over time so periodical snapshots are required to measure these metrics.

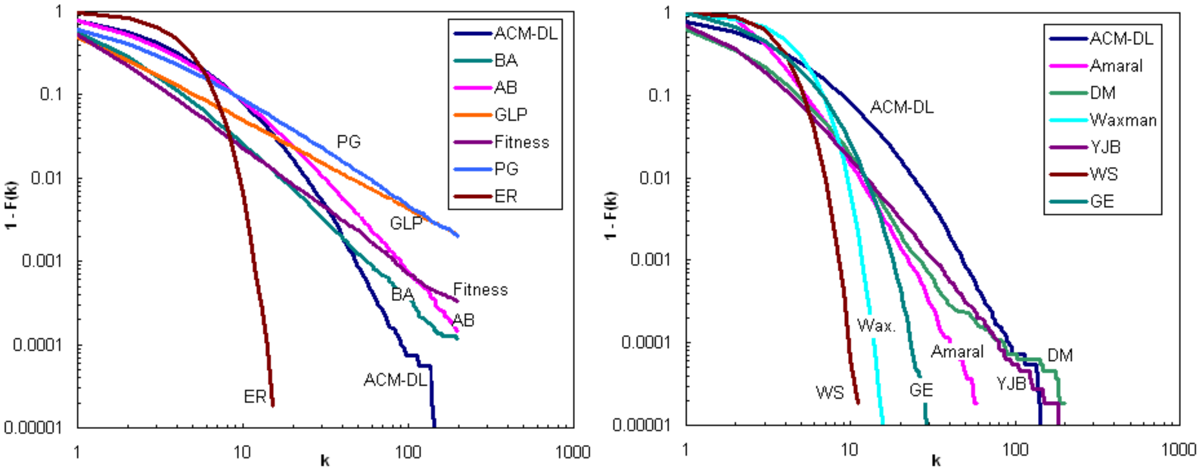

Figure 2. Degree distribution of all generative models and the collaboration network of ACM-DL

Figure 2 illustrates degree distribution performance. Here in order to eliminate the noise, cumulative degree distribution is used, which is defined as follows: Let N be the set of all nodes in the graph, and N_k be the set of nodes of degree equal or less than k. Then, F(k) = $|N_k|$ / |N| is the cumulative degree distribution. On the figures, the horizontal axis is the degree of nodes and the vertical axis represents 1-F(k).

The ER, GE, WS and Waxman models seem not to be able to generate a close degree distribution as the actual network's. The reason is clear, due to the mechanism they employ they can generate poison-like graphs which is insufficient for representing the most real-world networks including social networks. Whereas, the others all of which involves a variation of preferential attachment were able to generate good fits for actual power-law degree distribution. Yet, many of them show deviation from the actual network. The best fit is achieved by AB Model; especially for the lower values of k it shows a perfect match. One reason for AB Model to emerge among others is that it employs both internal and external link additions according to the preferential-attachment mechanism which is the real world situation comparing to only external link additions employed by the most of the remaining models.

Figure 3 presents the performance for the both dynamic measures. For the average distance measure, all the methods show a small average distance implying that all can generate small-world property well. However, if we look at the tendency the results seem not that promising. Particularly for the second half of the period analyzed in which the network has been more mature and seems to continue stabilized, the average distance of the network decreases. Nevertheless such a tendency can be imitated by none of the models experimented. They all show a regular increasing pattern due to the regularity in the attachment rule. Similarly, the clustering coefficient is high for the collaboration network, and yet it increases as the network grows. This measure is imitated even worse than average distance by the models. The pattern is decreasing in all models as well as the absolute value is dramatically small than the actual network's clustering co-efficient, meaning that those methods are all blind in re-generating clustering coefficient.

To summarize the results, among the models experimented, the ones that employ preferential attachment mechanism can be said to closely generate power-law degree distribution seen in the collaboration network of ACM-DL. However, the other two important properties are not captured satisfactorily. The results of all methods showed a small average distance, but the decrease in tendency is not seen as the network grows. Moreover, they all show too poor performance in terms of clustering coefficient, and its tendency. The collaboration network shows very high clustering (around 0.6) and this seems to continue to increase. Yet, all methods generated very small clustering coefficient with a decreasing trend.

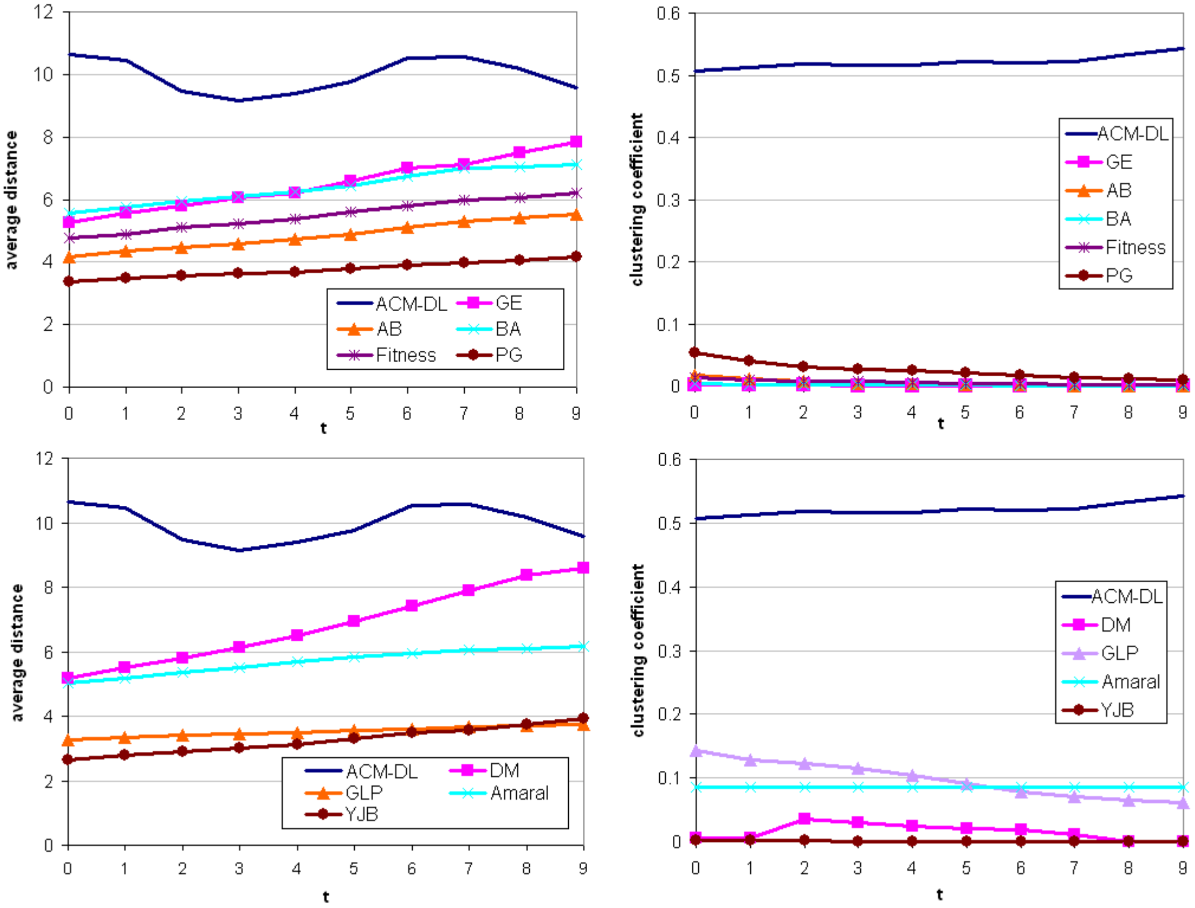

Figure 3. Results of the dynamic measures of all generative models and the collaboration network of ACM-DL

NEW MODEL

This section will describe a new model in order to capture all three characteristics of the collaboration networks by considering the problems in the existing models. Since the degree distribution is correctly generated, our goal is to try to improve both average distance and clustering coefficient results. Even though absolute values may differ, it is important for a good model to regenerate those two properties at least in parallel with the actual network's results while keeping the power-law degree distribution. We try to modify AB model which shows the best performance in terms of degree distribution to achieve this goal.

One problem with the existing methods is that they use fixed parameters for the probability of both internal and external link additions, no matter at what stage the network is. However, the collaboration network of ACM-DL does not show such regularity in terms of both link types. In the initial stage where the network is relatively small and tries to enlarge, the external links dominates the network. Many new nodes compared to the initial population enter the system, and connect to the existing nodes by creating external links. In the mean time, collaboration between old nodes is uncommon and decreasing the probability for internal link additions. This situation results in less number of short cuts between existing nodes and increases the average distance over the initial years. Nevertheless, this rate does not keep up at the same level. As the network becomes more mature, the probability of existing nodes to collaborate with each other increases significantly and later catches the same rate or sometimes even larger in the recent years of the period analyzed. This is the most important factor why we see a stable decrease in the later stage of the collaboration network. Hence we can also capture this situation by not changing the model but adjusting the parameters dynamically for the probabilities of internal and external link additions. We use AB model and compute the probability for internal link additions p(t) with a linear increase as follows:

$$p(t + \Delta t) = p(t) + \alpha \cdot p(t)$$

where initial internal link probability p(0) = 0.143 and the increase rate α = 0.15, chosen by considering the actual collaboration network's behavior.

Another problem with the existing methods is that they employ preferential attachment mechanism globally. In other words, if a node creates a link with a preferential attachment, it can connect to any of the remaining nodes in the entire network, with some probability less or more. This is not the case seen in a collaboration network. One person usually gets new acquaintances within a limited community, i.e. a friend of a friend, another person in the same company etc. So the attachment rule works within a limited range rather than for the entire network. This creates a hierarchical structure in the collaboration network, which is responsible for high value of clustering coefficient in such networks. A collaboration, or in general, a social network is in fact fundamentally modular; one can easily identify groups or communities that are highly interconnected within the group but have only a few links to nodes outside of the group to which they belong to. Hence, there are usually many nodes that have dense connections only within their community, where as relatively small number of other nodes creates the hierarchical topology by having connections between communities. As a result, high clustering coefficients of those densely connected small groups cause the entire network to have a high average clustering coefficient (Ravasz & Barabasi, 2003). Researchers have also tried to cope with the dramatic difference between the clustering coefficient values of the theoretical models and real-world networks. Among the ones that directly aim clustering coefficient issue, Holme and Kim (2002) introduce a mixture of global preferential attachment and local uniform attachment to increase the number of connected trios of nodes in the original BA model. Ravasz & Barabasi (2003) proposes a new hierarchical network model that combines scale-free property with high degree of clustering. In their model, a network starts with a small cluster of five densely linked nodes, and then four replicas of this module are created and connected to the central node of the old cluster. The same operation recursively continues to enlarge the system to the desired size. Frobczak et al. (2003) define a higher order clustering coefficient metric for complex networks claiming that it would be more appropriate metric to give insight into the modeling of clustering mechanism rather than using standard clustering coefficient. Higher order clustering coefficient measures connectivity between a node's immediate and mode distant neighbors to a specific distance.

To get better results for clustering coefficient, we use a similar scheme as in (Holme & Kim, 2002). We use the preferential attachment mechanism in all cases and limit the range for a new link addition for both internal and external linking in the AB model as follows:

(i) A new node creates its first external link according to the preferential attachment. However if it creates more external links, each consecutive one must be within the 2-node neighborhood of that node with probability p, and within the entire network with probability 1-p.

(ii) An existing node creates its internal links to connect to the other old nodes according to the preferential attachment rule. However, each internal links must be within the 2-node neighborhood of that node with probability p, and within the entire network with probability 1-p.

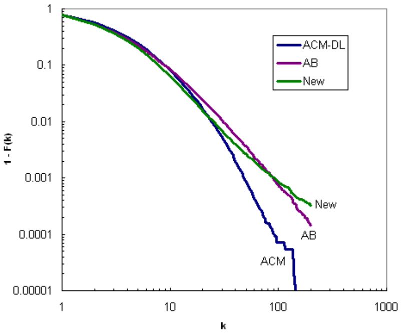

Figure 4. Results of the new model for ACM-DL

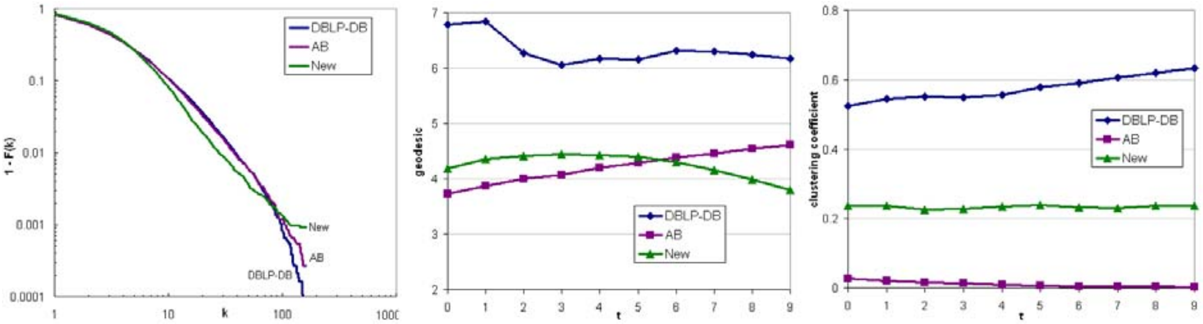

Figure 5. Results of the new model for DBLP-DB

The result of this model is shown in Figure 4. We experimentally select p = 0.5. It creates the power-law degree distribution since the mechanism is still preferential attachment with a slight deviation. Moreover it achieves the similar pattern with the average distance and clustering coefficient, and generates closer absolute values to the actual collaboration network in terms of clustering coefficient performance.

To justify the results of the new model, we repeat the model on generating the collaboration network of DBLP-DB. Data from 1985 to 2003 is divided into nine equal intervals similar to the one of ACM-DL, in which there are a total of around 18,500 nodes and 51,000 edges available at the end of the period analyzed. AB model is also experimented for comparison, where the parameters p=0.455, m=1.5, without rewiring. We set p(0) = 0.2 and α=0.11 for the dynamic parameters and use p=0.5 for the local preferential attachment probability. The results of the real data set, AB model and the new model are shown in Figure 5 on the three characteristics. AB model almost perfectly matches on the degree distribution while giving disappointing results on the remaining two metrics, like in the ACM-DL case. Despite a small deviation on the degree distribution, our new model is again able to regenerate the desired situation for the average distance and the clustering coefficient for this collaboration network as well.

Using dynamic network parameters and a mixture of local and global preferential attachment are surely two important factors in the modeling of collaboration network evolution but might not be enough as seen in the results. Many other such phenomena hidden in the actual collaboration patterns of scholars should be discovered and used along with these two factors for more accurate models. A possible future study may enrich the proposed model by considering the network to be composed of different sub networks each of which might refer a different field, study, location, etc. Each sub network may have its own events and attachment rules within the sub network and with the other sub networks, as seen in the actual collaboration networks. Incorporating more actual events happening in actual networks into a model will result in more accurate generation of such networks.

CONCLUSION

Most complex networks share similar properties that characterize the topology of a network: (i) having a power-law degree distribution, (ii) small average distance between all nodes, and (iii) having highly clustered. In this paper, we analyze a type of complex network: the collaboration network of scientists publishing the computing literature. We find that the collaboration network of ACM-DL has all these typical structural properties following a power law distribution with exponent -2.84, capturing “small-world” property with a small average distance of 7.9 and being highly-clustered with clustering coefficient 0.6 for the largest connected component of the network which has 57% of all authors. We also review different types of network models to generate complex network topology, and present the results of several generic network models for our collaboration network. The results show that although they are sufficiently able to regenerate power-law degree distribution of the network, not enough for capturing the other two dynamic properties. Finally, we present our model that modifies the original AB-model for generating a closer match and tendency for the average distance and clustering coefficient dynamism seen in our collaboration network. As a future direction, the model can be experimented on other types of complex networks and seen if it provides a good generative model for any network.

REFERENCES

Albert, R., & Barabasi, A.-L. (2000). Topology of evolving network: local events and universality. Physical review letters, 85(24):5234-5237.

Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. Rev. of Mod. Phys. 74, 47.

Amaral, L., Scala, A., Barthelemy, M., & Stanley, H. (2000). Classes of small-world networks. Proc. Natl. Acad. Sci. USA, 97:11149.

Atkins, H., Lyons, C., Ratner, H., Risher, C., Shillum, C., Sidman, D., Stevens, A., & Arms W. (2000). Reference Linking with DOIs: A Case Study. D-Lib Magazine.

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, 590-614.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286:509-512.

Bianconi, G., & Barabasi, A.-L. (2001). Competition and multiscaling in evolving networks. Europhys. Lett. 54 436.

Bu, T., & Towsley, D. (2002). On distinguishing between Internet power law topology generators. INFOCOM.

DBLP, Computer Science Bibliography. http://www.informatik.uni-trier.de/~ley/.

Dorogovtsev, S. N., & Mendes, J. F. F. (2000). Evolution of networks with aging of sites, Phys. Rev. E. 62 (2) 1842-1845.

Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks. Advances in Physics 1079-1187.

Elmacioglu, E., & Lee, D. (2005). On six degrees of separation on DBLP-DB and more. ACM SIGMOD Record, Vol. 34, No. 2, page 33-40.

Erdos, P., & Renyi, A. (1959). On random graphs. Publ. Math. Debrecen, 6:290-297.

Fronczak, A., Holyst J. A., Jedynak, M, & Sienkiewicz J. (2002). Higher order clustering coefficients in Barabasi-Albert networks. PHYSICA A 316 (1-4): 688-69.

Holme, P., & Kim, B. J. (2002) Growing scale-free networks with tunable clustering. Phys. Rev. E Vol 65, 026107.

Hong, Y., On, B.-W., & Lee, D. (2004). System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach. ECDL, Bath, UK.

Milgram, S. (1967) The small world problem. Psychology Today 2, 60-70.

Newman, M. E. J. (2001). Who is the best connected scientist? A study of scientific coauthorship networks. Phys. Rev. E64 016131; Phys. Rev.E64 016132.

Newman, M. E. J. (2003). The structure and function of complex networks. SIAM Review 45, 167-256.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323-351.

Park, S.-T., Pennock, D. M., & Giles C. L. (2004). Comparing static and dynamic measurements and models of the Internet's AS topology. INFOCOM.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: characterizing the competition for links on the web. PNAS vol. 99, no. 8 p. 5207-5211.

Ravasz, E., & Barabasi, A.-L. (2003). Hierarchical organization in complex networks. Phys. Rev. E 67 026112.

Watts, D. J. (1999). Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton University Press, Princeton.

Watts, D.J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature (London) 393, 440.

Waxman, B. M. (1988). Routing of Multipoint Connections. IEEE Journal on Selected Areas in Communications, 6(9).

Yook, S. H., Jeong, H., & Barabasi, A.-L. (2002). Modeling the Internet's large-scale topology. Proc. of the National Academy of Sciences (PNAS) 99 13382-13386.