Group-based Yule model for bipartite author-paper networks

Michel L. Goldstein,* Steven A. Morris,† and Gary G. Yen‡
Electrical and Computer Engineering, Oklahoma State University, Stillwater, Oklahoma 74078, USA
(Received 8 September 2004; published 9 February 2005)

This paper presents a model for author-paper networks, which is based on the assumption that authors are organized into groups and that, for each research topic, the number of papers published by a group is based on a success-breeds-success model. Collaboration between groups is modeled as random invitations from a group to an outside member. To analyze the model, a number of different metrics that can be obtained in author-paper networks were extracted. A simulation example shows that this model can effectively mimic the behavior of a real-world author-paper network, extracted from a collection of 900 journal papers in the field of complex networks.

DOI: 10.1103/PhysRevE.71.026108 PACS number(s): 89.75.Hc, 02.50.Ey, 87.23.Ge

I. INTRODUCTION

This paper presents a realistic bipartite model of author-paper networks, a phenomenon that has been studied since the 1920s [1]. The proposed growth model is based on modeling groups of authors using a "nested" Yule process [2], and further models "weak ties" among author groups as a Watts-Strogatz small world process [3]. The full bipartite representation of the network allows construction of many meaningful metrics to evaluate the validity of the proposed model against actual author-paper networks. Using a collection of 900 papers covering the topic of complex networks, we will show that the proposed model faithfully reproduces the characteristics of six metrics: (1) authors per paper distribution, (2) papers per author distribution (Lotka's law), (3) coauthor clustering coefficient distribution, (4) coauthorship per author pair distribution, (5) collaborator per author distribution, and (6) minimum path between author pairs distribution.

The model and the validation metrics presented in this paper are innovative when considered against previous models of Lotka's law or models of author collaboration networks. Lotka's law models deal with single authors without modeling collaboration, while collaboration models cannot describe Lotka's law and single authors. Both types of models are usually validated against simple power-law link degree distributions: papers per author for testing Lotka's law models, or collaborators per author for testing author collaboration models. Power-law link degree distributions are easy to duplicate using several types of processes [4]. Because of this, such simple models offer little insight into underlying processes that generate author-paper networks.

The proposed model, which deals with groups of authors rather than single authors, reveals the importance of research workgroups (author groups) in author-paper networks. The model indicates that publication by author groups is driven by a success-breeds-success (SBS) process, and further, that authorship by single authors within these groups is a SBS process as well. Yet, surprisingly, intergroup collaboration, i.e., weak ties, appears to be well modeled by a small world network of random interlinkages.

Bipartite author-paper networks are formed by two types of entities, the authors and papers, and the authorship links between them. There exists much analysis in the literature on the features of real-world author-paper networks. The first of these analyses was presented by Lotka [1]. His analysis, which contained a data set of journal articles compiled by hand, showed that the distribution of the number of papers per author follows a zeta distribution, a pure power law, with an exponent of approximately 2. This observation is currently referred to as Lotka's law of scientific productivity. A large number of other studies reinforced the power-law concept for the number of papers per author distribution, especially when considering only the tail of the distribution. These studies show that the observed exponent varies with the data set [5,6].

The observation of this distribution is very important, but it does not explicitly provide an insight into network dynamics. For this, a dynamic growth model is needed. Of the dynamic models in the literature, almost all are evaluated using crude comparisons to simple paper per author distributions and ignore other important metrics, such as clustering coefficient distribution or collaborator distribution. A complete and useful model must be able to mimic the real behavior of the author-paper network across many important network metrics.

This paper provides a model for the growth of author-paper networks and a step-by-step presentation of the important features of a real-world author-paper network that a model has to mimic. The proposed model, although very simple, approximates well all these features, thus building confidence in the validity of the model and the insight that the model provides into the actual dynamics of real-world author-paper networks.

II. AUTHOR-PAPER NETWORK MODELS

A number of different bipartite author-paper models exist in the literature. These models attempt to explain the process

*Electronic address: michel.goldstein@okstate.edu
†Electronic address: steven.a.morris@okstate.edu
‡Electronic address: gyen@okstate.edu

generating the power-law distribution of the papers per author distribution. They are fundamentally different from the usual preferential connection models, such as the Barabási-Albert model [4,7], because they model bipartite networks, in which one partition contains all authors and the other all papers. Although it is possible to transform a bipartite network into a simple graph by projection [8,9], this transformation removes the ability to calculate metrics to evaluate the validity of the model.

In the model presented by Newman *et al.* [10], the goal was to enable the generation of any degree distributions, such as Poisson, exponential, and power-law, for simple, directed, and bipartite graphs. The proposed method is very general but it is mainly focused on predicting three features: the average degree, the clustering coefficient [3], and the degree distribution of the projected graph. The model is able to effectively predict the features for a network of company directors; however, it fails to approximate the features of authorship networks.

Huber [11] presents a model of authors to predict five different features: the rate of production, career duration, randomness, Poisson-ness distribution (related to the variance of the author's productivity through time), and the distribution of papers per author. Huber's model is complex and involves distributions of career durations (assumed exponential) and Poisson distributed counts of papers, based on the author's productivity. Although this model predicts very well the features of interest, its major drawback is that it does not model the existence of coauthors. In the model, each author is "evolved" individually. A useful model must have the ability to predict collaboration patterns.

Recently, Börner *et al.* [12] presented a model in which the author network and the reference network evolve simultaneously. This study is an important acknowledgment that multiple interconnected networks exist in collections of journal papers, and that the challenge of modeling such paper collections is to find the basic rules of author behavior that produce the growth characteristics of the multiple interconnected networks contained in them. Börner *et al.*'s main goal was to predict the evolution of the number of papers, authors, and citations in a large and heterogeneous collection of journal articles, such as all of the papers published in the *Proceedings of the National Academy of Science* from 1981 to 2001. The paper includes a detailed set of proposed author behavioral rules and predicts gross measures of author, paper, and reference growth well, but the study does not discuss detailed metrics of network characterization.

One major disadvantage of all models found in the literature is the inability to predict most of the features of real-world networks. The prediction of only one or two features greatly weakens the usefulness of such models as models of real-world behavior.

III. PROPOSED GROUP-BASED YULE MODEL

A Yule model is a preferential connection process first proposed as a model of biological evolution by Yule in 1924 [2]. Our model uses a Yule process to model the growth of author groups in the author-paper network. The proposed model is based on the observation that usually authors are part of a research group. Most of the papers they write are coauthored with other members of their group. Collaboration between research groups happens, but multigroup papers are far less common than in-group papers.

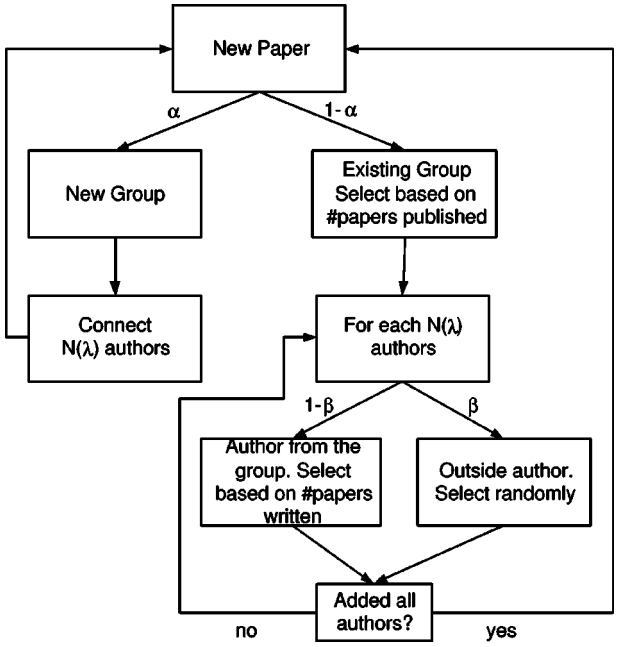

FIG. 1. Diagram of the proposed group-based Yule model for author-paper networks.

A diagram of the model can be seen in Fig. 1. When a paper is created there is a probability α that a new author group is created with N_g all new members, where N_g is a constant. The number of authors of the paper, $N(\lambda)$, is the first author plus a Poisson-distributed number of additional authors. This 1-shifted-Poisson distribution has parameter λ. The probability distribution of the 1-shifted-Poisson, $p_{sp}(k)$, is given in Eq. (1):

$$p_{sp}(k) = \frac{\lambda^{(k-1)}e^{-\lambda}}{(k-1)!}, \quad k = \{1,2,\ldots\}, \tag{1}$$

where k is the number of authors and $p_{sp}(k)$ is the probability of a paper having k authors.

If a new group is not created, an existing author group is chosen using the following probability distribution:

$$p_g(q) = \frac{q}{N_p}, \tag{2}$$

where q is the number of papers that this group has published, N_p is the total number of papers in the network, and $p_g(q)$ is the probability of an existing group authoring a paper. This is the Yule process, which favors groups in proportion to the number of papers they have published.

When an existing group is selected, it is necessary to select the authors within the group that author the paper. The number of authors of the paper is modeled as a 1-shifted Poisson distribution. In order to model interconnection between groups ("weak ties"), for each author, there is a prob-

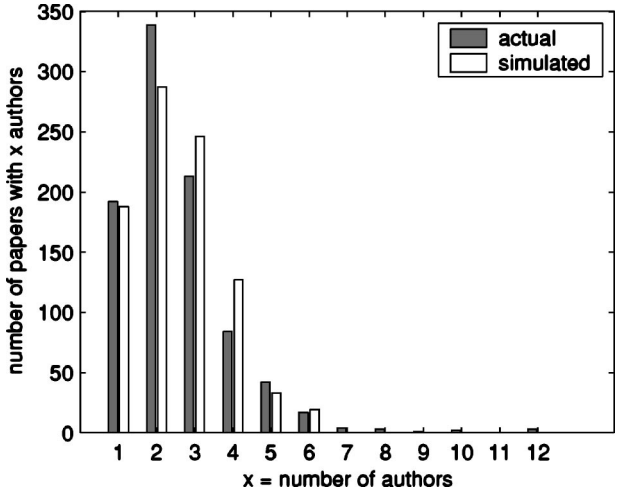

FIG. 2. Frequency distribution comparison for the number of authors per paper between the actual distribution and the simulated distribution. $\lambda_{actual}=1.527$, $\lambda_{sim}=1.651$.

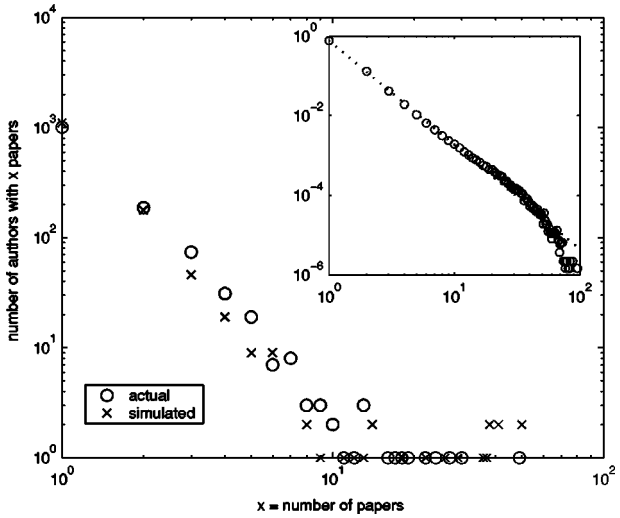

FIG. 3. Frequency distribution comparison for the number of papers per author, Lotka's law, between the actual distribution and the simulated distribution. The actual distribution has a power-law exponent of $\gamma=2.544$ and the simulated distribution has $\gamma=2.770$. The inset shows the model-predicted paper per author distribution, which fits a zeta distribution.

ability β that this author is from another group. If so, the author is selected randomly from among all authors in the network, whether they have authored a paper or not. If an outside author is not chosen, an author from the selected group is chosen. This selection is done by another preferential connection process, modified to allow selection of authors that have never published a paper. The probability of selecting an author i in the group is

$$p_a(i) = \frac{k_i + 1}{\sum k_j + N_g}, \tag{3}$$

where k_i is the number of papers written by author i, Σk_j is the sum of the number of authorships among the authors in the group, and N_g is the number of authors in the group. This is a preferential attachment process which favors authors by the number of papers they have previously published.

The paper creation cycle of Fig. 1 repeats until the desired number of papers is added to the network.

In summary, this model has four parameters: the group size N_g, assumed always constant for this simple model; the probability of creating a new group, α; the probability of choosing an author from another group, β; and the Poisson parameter that defines the distribution of number of authors per paper, λ. Given a data set to be modeled, it is easy to analytically determine α and λ.

The following section presents methods for obtaining these parameters to model a real-world network. Methods for correctly validating the model are also presented, by analyzing network metrics. It is important to note that the proposed model assumes that the groups are working on a single research specialty. For modeling multiple research specialties at once it would be necessary to restrict intergroup publishing between "related research," but this is beyond the scope of this study.

IV. EXAMPLE

The example is a collection of papers covering the specialty of complex networks. This data set, collected from the Science Citation Index, contains 900 papers, 1354 authors, and 2274 authorships linking authors to papers. Despite not being a large data set, its size is compatible to the assumption of the single specialty in which the proposed model operates. Moreover, being a rather new specialty, there are very few inner specialties that would need to be manually removed to fit with these assumptions.

The data set was obtained by obtaining all papers from ISI's Web of Science that satisfied the following queries: (a) cites references with (author=BARABASI-AL AND year =1999 AND journal=SCIENCE); (b) cites references with (author=WATTS-DJ AND year=1998 AND journal =NATURE); (c) cites references with (author=ALBERT-R AND year=2000 AND journal=NATURE); (d) cites references with (author=ALBERT-R AND year=2002 AND journal=REV-MOD-PHYS); (e) cites references with (author=DOROGOVTSEV AND year=2002).

The queries above yielded 832 papers. Additional papers were added manually from a list of papers citing additional authors NEWMAN-ME and PASTORSATORRAS-R, collected previously [13].

The first parameter α is obtained by determining the probability of new group creation. This probability is estimated using a paper-by-paper pass through the network to determine the fraction of papers that appeared with a completely new set of authors.

The parameter λ is calculated by dividing the total number of authorships by the number of papers and subtracting 1 (1-shifted Poisson estimate). The number of authors per group, N_g, was chosen heuristically as 20, which is assumed as the upper limit of the number of researchers that can efficiently interact as a group.

The "weak tie" parameter β is estimated by matching the coauthorship distribution (the distribution of the number of times pairs of authors have coauthored) by trial and error.

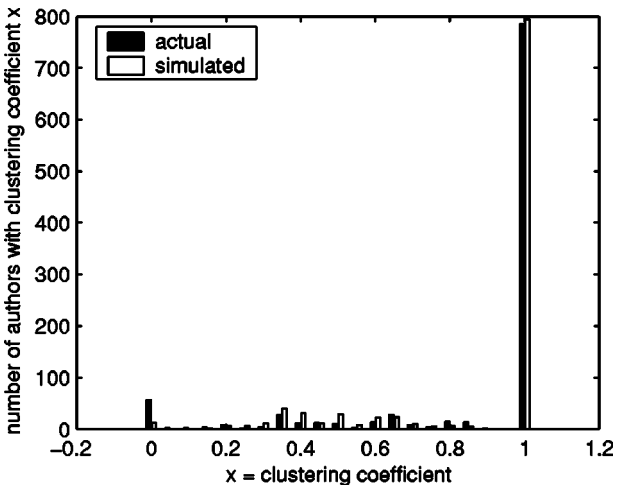

FIG. 4. Frequency distribution comparison of the clustering coefficient. $C_{actual}=0.867$, $C_{sim}=0.881$.

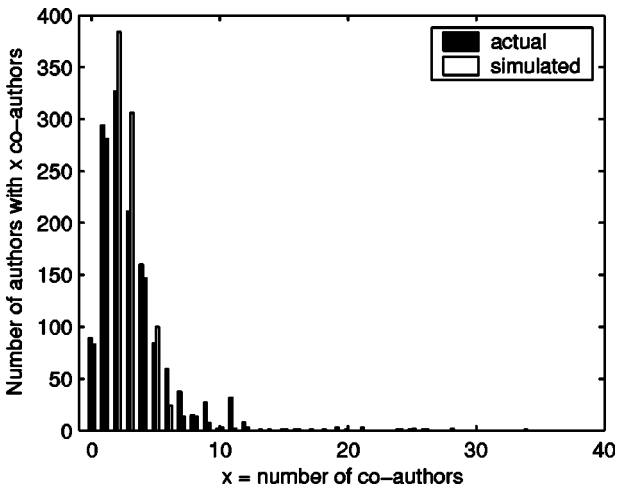

FIG. 5. Frequency distribution comparison of the number of collaborators per author. $\mu_{actual}=3.15$, $\mu_{sim}=2.82$.

The matching of the coauthorship distribution will be explained below.

The parameters estimated for the example network are $\hat{\alpha}=0.33$, $\hat{\beta}=0.1$, $\hat{\lambda}=1.527$, and $N_g=20$.

To validate the model, several metrics are used to compare model simulations to the actual network. The following metrics are used for comparison.

Authors per paper. The distribution of the number of authors per paper. As discussed above, this is simulated as 1-shifted Poisson distributed. Note in Fig. 2 the close match of actual to simulated frequencies, further confirming the 1-shifted Poisson assumption presented. This metric is important because it relates directly to the number of participants on projects within the group, an important measure of interaction within workgroups.

Papers per author distribution (Lotka's law). This is the distribution of the number of papers that each author published. Note in Fig. 3 the close match of simulated frequencies to actual frequencies for this metric. This metric is important because it measures the distribution of productivity among authors in a specialty, modeling the formation of core groups of researchers in a specialty. The inset in Fig. 3 shows the model-predicted paper per author distribution, generated by gathering statistics from 1000 simulations. The predicted distribution certainly models Lotka's law, producing an excellent fit to a zeta distribution with an exponent of 2.77. Fitting was done using maximum likelihood expectation and the fit passed a Kolmogorov-Smirnov (KS) test [14] at an observed significance level (OSL) of $10\% < \mathrm{OSL} < 1\%$, $T=0.0031$, $N=1.3\times10^6$. The KS is a commonly used goodness of fit test whose test statistic is based upon the maximum deviation of the cumulated experimental distribution from the proposed distribution. For details, please see [15].

Coauthor clustering coefficient distribution. The clustering coefficient was first introduced by Watts and Strogatz [3] as a scalar mean clustering coefficient. However, when observing the distribution of the clustering coefficients, a very interesting characteristic is found in coauthor networks: a large spike at unity. Therefore, it is imperative to use the distribution as the metric rather than the mean, which effectively hides unity spike behavior. For example, although author networks usually have a mean clustering coefficient of 0.8, comparable to that of citation networks [4], the distribution of the coauthor network clustering coefficient is fundamentally different from the distribution of clustering coefficient in citation networks [16]. Newman *et al.* discuss this distribution in [10] and model it, with limited success. Note in Fig. 4 that simulation using the proposed model fully mimics the distribution of the clustering coefficient. This metric is important because it measures the tendency of authors to work in local groups.

Collaborator distribution. The distribution of the number of unique coauthors to each author in the network. Newman *et al.* attempted to model this distribution with only partial success [10]. Note the close match of the simulated to actual coauthorship frequencies in Fig. 5. This metric is important because it measures the tendency of authors to work with other authors.

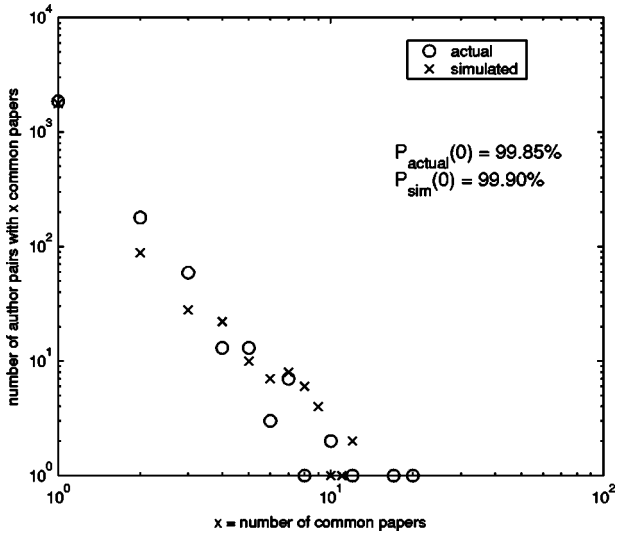

FIG. 6. Frequency distribution comparison of the coauthorship distribution, showing the number of papers coauthored by each pair of authors.

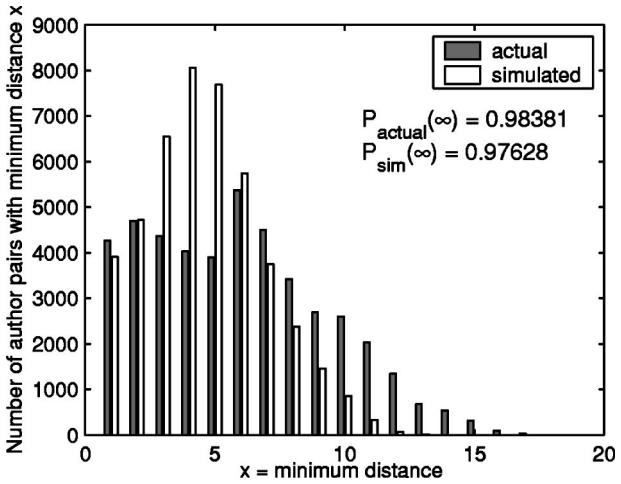

FIG. 7. Frequency distribution comparison of the minimum distance between authors, i.e., the minimum number of links between each pair of authors.

Coauthorship distribution. This is the distribution of the number of common papers between pairs of authors, across all author pairs in the network. Figure 6 shows that the proposed model matches the actual distribution well. This is an important metric because it measures the tendency of pairs of authors to repeatedly work together on individual projects.

Minimum distance distribution. Figure 7 shows the distribution of the minimum distance between pairs of authors in the network, i.e., the minimum length of the path of coauthorships between them. This metric is important because it measures the tendency of groups to invite outside workers onto projects.

For additional discussions of network metrics applicable to author-paper networks, see Newman [6], who discusses several of the metrics used here.

All metrics shown above present a close match between the real-world network and the model simulation. As an exception, the minimum path distribution shows a fair amount of deviation, but this distribution appears to be unstable and tends to change greatly from simulation to simulation. The actual minimum path length distribution is probably unstable as well, but investigation of that hypothesis is outside the scope of this paper.

V. CONCLUSIONS

This paper proposes a very simple model for author-paper networks by introducing the concept of preferential attachment of group authoring of papers. Adding this simple concept to a Yule-type process it was possible to obtain very similar behavior using multiple metrics, when comparing to a real-world network. This suggests that, in the real world, the modeling of research groups is essential to understanding the dynamics of paper authoring. Analysis of single authors or random connections between authors, as proposed by previous researchers, does not provide a reasonable model of reality.

Another important conclusion drawn from this model is that "weak ties" between groups are well modeled by simple random intergroup coauthorships. This implies that group collaboration does not actually work by establishing formal long-term commitments, but by single collaborations, possibly from informal meetings at conferences, or e-mail discussion lists. Multiple collaboration with outside groups may happen in real life, but such collaborations are uncommon and do not affect the gross characteristics of the network. This model further implies that outside collaboration is not dependent on the amount of work that the outside person has done in the field.

Note that while there is local preferential connection of authors within groups, and global preferential connection of the groups themselves, the intergroup linkage approximates a Watts-Strogatz small world process. The model here is really a hybrid, being a "nested preferential connection, global small world" model.

We also showed that using only a single metric, such as the distribution of papers per author, or a single mean value for the clustering coefficient, incompletely validates a model. Analyzing multiple metrics allows validation against specific behaviors that fully characterize the network.

It is important to note that this model only accounts for the behavior of authorships in a collection of papers. To actually understand the nature of collections of journal papers it would be necessary to implement and discuss the interaction of this author-paper bipartite network with the other bipartite networks in the paper collection, such as the paper-reference network [16–18], paper-journal network (Bradford's law) [19], and paper-term network (Zipf's law) [19]. The analysis of their complex interaction will certainly shed light on a large number of open questions regarding the growth and mapping of information structures.

[1] A. J. Lotka, J. Wash. Acad. Sci. **16**, 317 (1926).
[2] G. U. Yule, Philos. Trans. R. Soc. London, Ser. B **213**, 21 (1924).
[3] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[4] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[5] M. L. Pao, Inf. Proc. Mngt. **21**, 305 (1985).
[6] M. E. J. Newman, in *Complex Networks*, edited by E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai (Springer, Berlin, 2004).
[7] A.-L. Barabási, H. Jeong, R. Ravasz, Z. Nida, T. Vicsek, and A. Schubert, Physica A **311**, 590 (2002).
[8] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[9] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **101**, 5200 (2004).
[10] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[11] J. C. Huber, J. Am. Soc. Inf. Sci. **53**, 209 (2002).
[12] K. Börner, J. T. Maru, and R. L. Goldstone, Proc. Natl. Acad. Sci. U.S.A. **101**, 5266 (2004).
[13] S. A. Morris and G. G. Yen, Proc. Natl. Acad. Sci. U.S.A. **101**, 5291 (2004).
[14] M. L. Goldstein, S. A. Morris, and G. G. Yen, Eur. Phys. J. B (to be published).
[15] W. J. Conover, *Practical Nonparametric Statistics* (Wiley, New York, 1999).
[16] S. A. Morris, J. Am. Soc. Inf. Sci. Technol. (to be published).
[17] S. Naranan, J. Doc. 27, 83 (1971).
[18] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[19] H. D. White and K. W. Mccain, Annu. Rev. Info. Sci. Technol. **24**, 119 (1989).