

# Manifestation of research teams in journal literature: a growth model of papers, authors, collaboration, coauthorship, weak ties, and Lotka's law

Steven A. Morris\* and Michel L. Goldstein†

Oklahoma State University  
Electrical and Computer Engineering  
Stillwater, OK 74078, USA

(Dated: November 12, 2005)

This paper introduces a team-based model of researchers in a specialty and investigates the manifestation of such teams in a specialty's literature. The proposed qualitative behavioral model, with its mathematical expression as a growth model, is significant because it simultaneously describes the two phenomena of collaboration and author productivity (Lotka's law) in a specialty. The model is nested: a team process models the creation of research teams and the success-breeds-success process of their production of papers, while at a lower level the productivity of authors within teams is also modeled as a success-breeds-success process. Inter-team collaboration (weak ties) is modeled as random events. This simple growth model is shown to faithfully mimic six network metrics of bipartite paper-author networks. The model is demonstrated on three example paper collections from specialties that have a wide range of degree of collaboration: 1) a distance education collection with low collaboration degree, 2) a complex networks collection with typical collaboration degree, and 3) an atrial ablation collection with heavy collaboration degree.

## 1. INTRODUCTION

### 1.1. Motivation

Collections of journal papers that cover a research specialty are collections of research reports that have been vetted by the review process. Such collections can be monitored by subject matter experts to assess the state of research progress in the specialty, and this information can be used to advise government policy makers, business leaders and leaders in research institutions (Porter *et al.*, 1991).

A research specialty is a self-organized social organization that manifests itself in its journal literature through several types of entities: 1) papers, 2) journals, 3) paper authors, 4) references, 5) terms, 6) reference authors, 7) reference journals, and more (Morris and Yen, 2004). Among these entities, paper authors, hereafter referred to simply as 'authors', represent the human creators of research. As such, their behavior and social organization drive all other aspects of the specialty. The productivity of authors, measured in the number of papers they publish, is an indicator of their standing and importance in the specialty. The social organization of researchers, and the pattern of assistance and information sharing that they provide one another, that is, their pattern of *collaboration*, form a complex network. Knowledge of the structure and dynamics of this collaboration network provides great insight into the processes driving research within a

specialty, especially the processes driving research team formation, inter-team collaboration, and growth to dominance of successful teams and successful researchers.

Coauthorship of papers by authors is an imperfect measure of the processes driving collaboration among researchers in a specialty (Melin and Persson, 1996; Subramanyam, 1983). However, there are many advantages to studying collaboration through coauthorship in collections of papers, chief advantages being that such a method is inexpensive and practical compared to conducting surveys and interviews.

A model of the growth and structure of paper-author networks allows research into methods of measuring and characterizing those networks and, by inference, the state of the specialties themselves. Furthermore, a model of paper-author network growth will help to identify indicators of trends and events in the specialty, e.g., discoveries, changes in scientific paradigms, exhaustion of research problems, flight of researchers from a specialty and migration of researchers across specialties.

### 1.2. Summary

This paper presents a team-based model of collaboration that was initially studied by Goldstein, Morris, and Yen (2005). The proposed qualitative behavioral model is significant because it simultaneously describes the two phenomena of collaboration and author productivity (Lotka's law) in a specialty, and provides great insight into the importance of teams in a research specialty.

The scope of this model is limited to collections of papers that cover a single specialty, where papers are generally written by authors all drawn from a single research team, who work on unitary research tasks.

---

\*Email:samorri@okstate.edu; Homepage:samorris.ceat.okstate.edu

† This is a working paper that has been submitted for publication. Inquiries should be directed to S. Morris.

The model is mathematically expressed using the Yule process (Simon, 1955) for the creation and productivity of teams, and the model further describes the productivity of individual members of research teams as a success-breeds-success process. Collaborations between teams (weak ties) are modeled as random invitations from teams to members of outside teams.

This simple growth model is shown to faithfully mimic six network metrics of bipartite paper-author networks. The model is demonstrated on three example paper collections from specialties with a wide range of degree of collaboration: 1) a distance education collection with low collaboration degree, 2) a complex networks collection with typical collaboration degree, and 3) an atrial ablation collection with heavy collaboration degree.

### 1.3. Organization of this paper

Section 2 establishes the foundation for the proposed model by presenting a qualitative team-based social model of research. The necessary framework for a quantitative model is given in Section 3, which describes a matrix representation of author-paper networks, and in Section 4, which discusses network metrics. Section 5 introduces a quantitative team-based growth model and discusses the derivation of model parameters from a collection of papers. As evidence to support the proposed model, Section 6 reports on modeling of three example specialties that range widely in their intensity of collaboration. Finally, Section 7 wraps up by discussing the implications of the proposed model.

## 2. RESEARCH TEAMS

### 2.1. Introduction

In this paper we will introduce a team-based model of collaboration and show that this model mimics the characteristics of authorship and coauthorship in real collections of papers. There is much evidence that team-based research drives journal literature. Price and Beaver (1966), studied an information exchange group in a biomedical specialty and found that the authors could be easily sorted into groups of various sizes, and that members of groups tended to be from the institution of the group's leading author. They also noted the existence of dominant authors within groups and the existence of dominant groups in the specialty. Mahlck and Persson (2000), mapping author networks in universities, found that departments could be broken into groups of authors dominated by lead professors. Peters and van Raan (1991) found similar structure in the coauthorship network of a chemical engineering department. In an extensive study, Seglen and Aksnes (2000) studied the coauthorship network in Norwegian microbiology and showed that the network could be divided into 180 groups with a range of sizes.

Beaver and Rosen, in a three part series (Beaver, 1978, 1979; Beaver and Rosen, 1979), describe the spread of coauthorship in journal literature through the past two centuries, and note that coauthorship increased with increasing institutionalization of research. This institutionalization accompanied the increase in funding of research through government and industrial sources. Institutionalization corresponds to the formation of research organizations. We assume that at the bottom of such hierarchical research organizations, where the actual research work is conducted, are small teams of researchers, whose members divide their labor on research tasks. As tasks are completed the contributing researchers author a report, which is submitted as a paper to a journal for publication.

Implied in this team-based model of research is the assumption that research teams are at the bottom of the organizational hierarchy, and that they cannot be broken into subgroups, that is, the next hierarchical level below a team are the individual researchers. The members of the team interact on a daily basis, with little formal organization except for the existence of a team leader. Such a team would necessarily be small, where the members could comfortably fit in a conference room for meetings and seminars. Certainly then, such teams would be less than 20 members, but might average 5 to 8 members. Steiner (1972) reports that group members seem to feel most comfortable in groups of 5 members, based on communication and conflict resolution issues. Steiner notes that optimal group size, however, is a function of the type of task for which the group exists.

Note that the majority of journal papers in many fields are written by authors from universities (Godin and Gingras, 2000). So it seems probable that many teams are comprised of graduate students and their faculty leaders. Subramanyam (1983) presents 6 types of collaboration. Three of these types can be considered consistent with a team-based model of research: 1) teacher-pupil collaboration, 2) collaboration among colleagues, and 3) supervisor-assistant collaboration. The remaining three types of collaboration can be considered as between teams: 4) researcher-consultant collaboration, 5) collaboration between organizations, and 6) international collaboration.

### 2.2. Definition of terms

As outlined by Laudel (2002), it is difficult to define many of the basic concepts of collaboration. Nevertheless, for the benefit of the reader, we will define terms as used in this paper before proceeding with any discussion. Consider the following definitions:

**project:** an organized effort, comprised of a series of tasks, conducted to solve some research problem.

**task:** a sub-problem of a research project. For our proposed model, tasks are considered unitary, that is,

tasks are small enough that they cannot be divided into sub-tasks.

**team:** a small group of researchers who interact on a regular basis. Teams exist to accomplish tasks, and teams may work on several tasks simultaneously.

**co-workers:** researchers who are members of the same team.

**collaborators:** researchers who work together on a task. By definition, all co-workers are collaborators, and additionally, researchers from outside a team who assist in accomplishing a task are collaborators with the members of that team.

**collaboration links:** a pair of researchers that have collaborated are said to be linked by collaboration. By definition a member of a team is so linked to all other members of the team.

**strong ties:** collaboration links between co-workers.

**weak ties:** collaboration links between researchers from different teams.

It is important to note that not all researchers that collaborate on accomplishing a task will necessarily appear as authors of the paper that is written to report on that task. In many cases, as discussed by Laudel (2002), the assistance provided by collaborators is not important enough to warrant inclusion of a researcher as an author of the paper that reports on the task. Also, not all authors of a paper are necessarily collaborators on the task that a paper is reporting about. In some cases authorship may be assigned on an "honorary" basis to reward persons who support the team's research through funding, advocacy, or lending of prestige and credibility (Laudel, 2002).

Consider the following definitions pertaining to authors:

**writer:** the writer is the researcher that performs the actual organization and writing of a paper. We assume that in most cases the writer of the paper is the lead author.

**author:** an author of a paper. Authorship bestows recognition of contribution to research and assigns responsibility for results reported in the paper

**lead author:** the author responsible for the report. This is assumed to be the researcher responsible for accomplishing the research task that a paper is reporting on. In many fields, this is traditionally the paper's first author.

**secondary authors:** authors who are not the lead author. These are assumed to have participated in accomplishing the research task that the paper reports on. The number of secondary authors on a

paper is one less than the number of authors on the paper. Single-author papers have a lead author only and no secondary authors.

**authoring team:** the team to which the lead author belongs. It is assumed that most secondary authors will be drawn from the authoring team.

**coauthor:** all the authors that appear on a paper are coauthors with each other. Coauthorship is used in the context of specific papers only, as in, "coauthors on paper X." See "cooperating author" for a more general term.

**cooperating author :** two authors that have coauthored one or more papers are cooperating authors. This term is borrowed from Glänzel (2002).

**hyperauthorship** authorship by many authors drawn from many teams simultaneously. This term is borrowed from Cronin (2001). As an operational definition, any paper with 20 or more authors exhibits hyperauthorship. Upon further research into the phenomenon of hyperauthorship, it is possible that the definition of hyperauthorship will vary from field to field.

**coauthorship links or cooperativity links:** a pair of researchers that have coauthored a paper are said to be linked by coauthorship.

Cronin (2001) provides a brief history of authorship, outlining the rise of conventions of authorship as journals started to replace correspondence, three centuries ago, as the primary means of dissemination of research results in research specialties. We assume that authorship has two functions: 1) recognition of contribution to the research effort, and 2) assignment of responsibility in case of errors or fraud. The position of the lead author in the author list may depend on the traditions of the field, but normally the first author in the author list corresponds to the lead author, and often the final author corresponds to the team leader (Subramanyam, 1983).

### 2.3. Modes of authorship

For discussion in this paper we assume three modes of authorship. These modes are assumed to characterize a specialty. In the first mode, *team authorship*, we assume that research tasks are unitary, as defined by Steiner (1972). Papers are authored by teams that perform research without equipment, e.g., mathematicians, or who perform their work in small laboratories dedicated to single teams. In the second mode, *multi-team authorship*, the tasks are divisible, as defined by Steiner, and multiple teams participate in projects and tasks and author the papers that result. The third mode of authorship, *mass recognition*, corresponds to what Price calls "big science" (Price, 1986). This mode of authorship may produce papers with hundreds of authors. As shown in Figure 1,

these modes can be considered to form a continuum ranging from 1) specialties dominated by "lone wolf" single authors in fields like mathematics, through 2) specialties, such as those from biomedical fields, dominated by "wolf pack" authored papers with 10 to 20 authors, to 3) "buffalo herd" authored papers with hundreds of authors in fields such as experimental high energy physics.

**Team authorship.** Team authorship corresponds to "little science." Papers are authored by a group of one or more authors that form a subset of a team. The authors may invite authors from other teams to collaborate in authoring a paper, but usually most of the authors of a paper are co-workers. In this mode the number of secondary authors tends to be Poisson distributed, producing a 1-shifted Poisson distribution of authors per paper. This distribution is assumed to be a result of the process that lead authors on papers use for selecting their secondary authors, as will be explained in Section 2.6.

**Multi-team authorship.** There are some fields, particularly biomedicine, where research tasks are routinely accomplished by co-operative work across two or more teams. Because of this, many papers in the field are authored by members of multiple teams, and the number of authors on individual papers can be much larger than for single team authorship. The distribution of the number of authors deviates from the 1-shifted Poisson and may approach a power-law distribution (Beaver, 1984, 2001). The tasks reported on in these papers are divisible into subtasks that can be assigned to separate teams. This multi-team mode may be the result of the need to cooperate on the use of expensive equipment, or the need for most problems to be worked on by specialists from two or more distinct disciplines (Katz and Martin, 1997).

**Mass recognition authorship.** The mass recognition authorship process corresponds to the big science model of research (Price, 1986). In this process the paper is written to report research that was performed by many teams working together. Coauthorship of the paper is used to acknowledge each team's participation in the experiment and may result in hundreds of authors on a paper (Cronin, 2001). Authorship is rewarded to each member of each team that participated in the research task reported by a paper. This differs from the multi-team mode, where authorship is assigned based on individual researchers' participation of the research task, so that many team members of the participating teams do not become authors. Mass recognition authorship in a specialty tends to produce a power-law distribution of authors per paper (Beaver, 1984, 2001).

#### 2.4. Scope of proposed model

The model proposed here is intended specifically to characterize single research specialties. As discussed by Morris (2005), a collection of papers that comprehensively samples a specialty's literature is expected to range in size from 100 papers to 5000 papers. Studies of large

collections of papers covering multiple specialties fall outside the scope of the model proposed here.

The model proposed here is intended to characterize specialties where most papers are authored by subsets of researchers from single teams. As will be discussed in Section 2.6, specialties characterized by single-team authorship are assumed to produce collections of papers with an author per paper distribution that approximates a 1-shifted Poisson distribution. As a working limit, we assume that the number of authors per paper is less than 20. Above this limit we assume that the mode of authorship is a multi-team or mass recognition mode of authorship, where numbers of authors per paper tends toward a power law. Figure 1 shows diagrammatically the scope of the proposed model compared to various modes of authorship discussed above.

In the examples included in this paper, we will model collections of papers across a range of single team authorship: 1) a "typical" single team mode example from the specialty of complex networks, 2) a marginally multi-team authorship collection from the biomedical specialty of atrial ablation, , and 3) an example with low degree of collaboration taken from the education specialty of distance education.

#### 2.5. Weak ties

Granovetter (1973) introduced the concept of *strong ties* and *weak ties*. In a social sense, an individual's strong ties denote the links to close relatives, co-workers and friends. These links are tightly knit, that is, they are to people who are quite likely to know each other, and form a tight social group. Outside of this tightly knit group, however, an individual has "weak ties" to people with whom he/she interacts only occasionally. In a social sense, these weak ties provide very useful communication links that allow individuals to reach outside their immediate group. For example, Granovetter (1973) showed that job seekers tended to seek and find jobs by using weak ties rather than strong ties.

In the context of individual researchers and their social network of fellow researchers, strong ties are represented by members of his/her research team. It is true, especially today with tools like email, videoconferencing, internet collaboration tools like Netmeeting, and cheap long distance telephones, that researchers can maintain very close associations with other researchers from around the world. But there are certainly institutional barriers to strong inter-team associations. Two researchers from two different institutions are likely to be focused on two different research tasks, and each researcher is contractually obligated to spend the majority of their time focused on their own institution's project. Further, unless formal agreements are in place between institutions, there is very likely to be contractual obligations of researchers to not share intellectual property with researchers from outside institutions. It seems likely that

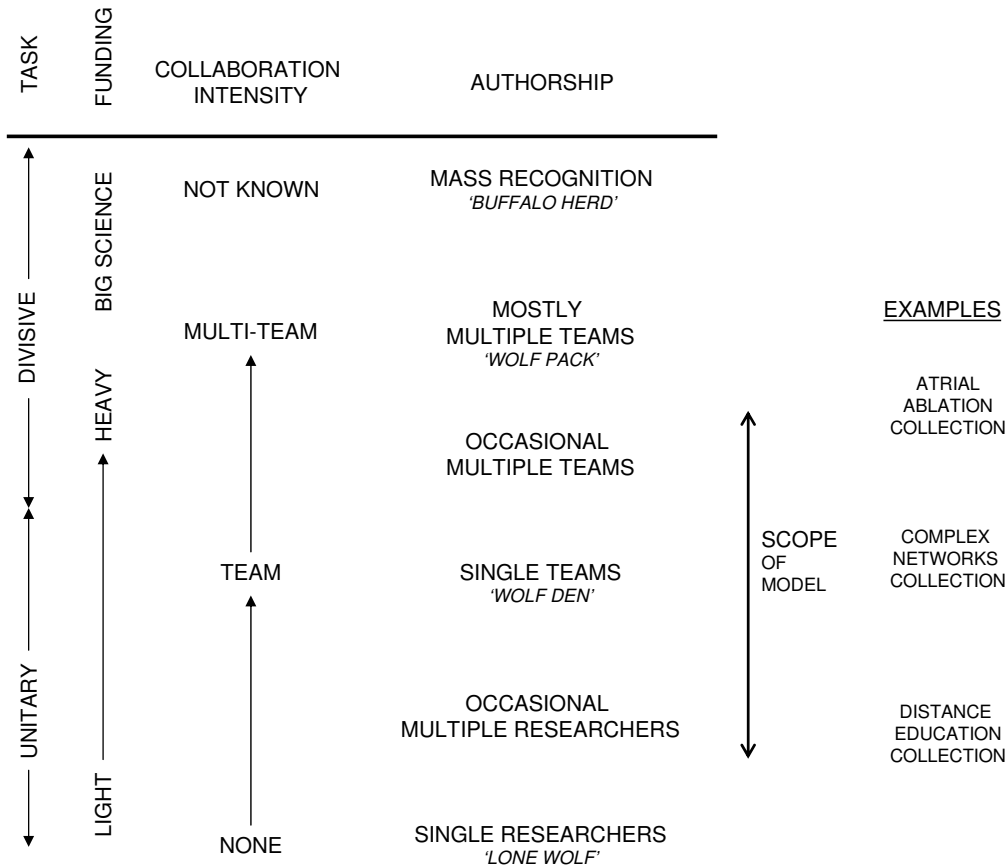


FIG. 1 Diagram of collaboration and its correlation to task type, funding, and authorship mode. The scope of the proposed model is shown on the right along with the perceived level of collaboration of the three examples included in this paper. Note that the scope of the model is for unitary tasks accomplished by single teams.

the barriers to forming strong ties outside the researcher's institution would make such associations rare.

On the other hand, weak ties between research groups, corresponding to occasional or one time collaborations on tasks, should not be particularly rare. Tasks often require expertise that is not possessed by any researchers in the team. Researchers that possess this missing expertise will be drawn from outside the team, often from other teams within the same research institution, but also from outside the institution as well.

In the model to be presented here, weak ties will be modeled as random inter-team collaborations between a team working on a task and an outside researcher. Typically, in the examples presented here, this corresponds to about 10% of authors of individual papers being drawn from outside of the authoring team's members.

## 2.6. A qualitative model of author behavior

In accordance with the definitions given above, we assume that each research team is working on some research project and that individual researchers in the team are given tasks that are part of this project. In the case of

an industry research team, the team's project will correspond to an actual project. In university research, however, it is also likely that a research project corresponds to some general research direction that aligns with the specialty of the senior faculty advisor of the team, the team being comprised of the faculty advisor, and his graduate students.

The team member who performs a research task draws upon the expertise of the other members of the team to varying degrees. The participation of other team members may range from informal conversations during coffee breaks, to full time participation in the research task.

Upon completion of the research task the researcher performing the task reports the results in a journal paper. This researcher becomes the writer of the paper and appears as the lead author. Assuming some threshold of participation in the task, those researchers whose participation in the task was above this threshold become secondary authors on the paper. Assuming a uniform distribution of participation level among the members of the team, this means that the probability that a team member is above the participation threshold is some fixed probability. This process yields a binomial distribution of the number of secondary authors, which can be approx-

imated as a Poisson distribution. From this it follows that the distribution of the number of authors approximates a 1-shifted Poisson distribution. Note that this model of paper-authoring is consistent with the concept, reviewed by Subramanyan (1983), of sub-authors. These are researchers that did not participate in the task at a level that warranted authorship, yet participated at a level that is rewarded by acknowledgement in the paper. Sub-authorship can be modeled by using a second participation threshold, below the authorship threshold, to model the distribution of the number of received acknowledgements in the paper.

Because of limitations of expertise within the team, the team will occasionally need to engage researchers from outside the team to help in accomplishing the task. These outside workers also tend to assist in the task with various levels of participation and so, by the same process for selecting team members as authors, outside researchers may become secondary authors on the paper that reports results from the task.

Teams that successfully accomplish their research goals or whose research leads to discoveries within the field will be able to acquire funds for more research projects. This process can be modeled as a success-breeds-success process, which leads to a power law distribution of team productivity of papers. The Yule process is appropriate and models both creation of new teams and the success-breeds-success production of papers within the specialty by teams.

A team member who successfully completes a task will be asked to perform additional tasks, and may even eventually assume supervisory duties of several tasks simultaneously. These successful and talented researchers will become mentors to other team members, increasing their average participation in the team's research tasks, and thus increasing the number of papers for which they become secondary authors. Thus, the productivity of individual researchers in a team can be modeled as a success-breeds-success process. Fox (1983) reviews many discussions of author productivity and notes that authors can accumulate advantage through many avenues, such as early publications, prestige of university granting their degree, and resources afforded by their funding organization. Note that the model to be proposed here is consistent with the points reviewed by Fox: teams cumulate advantage in terms of prestige and resources of their institution, while the individual authors cumulate advantage through early publication, reinforcement, and increasing access to resources as reward for success.

This qualitative model provides a basis upon which to build a mathematical model of paper-author networks. It constitutes a simple social model of team behavior, coupled with some observations about team structures, rewards, and collaboration in research. It fits the facts of research well, modeling the development of successful researchers from graduate students to junior faculty to well established faculty advisors through success-breeds-success. It further models success-breeds-success among

research teams, where success produces continued and increased funding, and acquisition of prestige.

### 3. MATHEMATICAL REPRESENTATION

The structural and dynamic description of the paper-author matrix that follows is necessary to help the reader understand the manifestation of team-based research in journal literature, and help the reader to assess the evidence presented in Section 5.1, which supports the validity of a team-based research model. Furthermore, all of the metrics of author-paper networks used in this paper, which are described in detail in Section 4, are derived from the paper-author matrix.

#### 3.1. Paper-author matrix

Paper-author networks are most conveniently represented using a paper-author matrix,  $\Omega$ , whose rows correspond to papers in the collection and whose columns correspond to authors in the collection. The size of  $\Omega$  is  $np$  by  $nap$ , where  $np$  is the number of papers in the collection while  $nap$  is the number of authors in the collection. The matrix is binary and its elements are defined as:

$$\omega_{ij} = \begin{cases} 1 & \text{if paper } i \text{ has author } j \text{ as an author} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The structure of the paper-author matrix is identical to the structure of the paper-reference matrix presented by Morris (2005). Newly appearing papers are added to the paper-author matrix as new rows at the bottom of the matrix, and newly appearing authors appear as new matrix columns on the right. This results in a roughly lower triangular matrix with a continuous stair step of 1's down the approximate diagonal of the matrix. Figure 2 shows a diagram of a typical paper-author matrix, taken from a collection of papers on the topic of complex networks, an example that will be presented later in this paper. Section 5.1 will discuss how the specific internal structure of the paper-author matrix supports the team-based model that is proposed in this paper.

#### 3.2. Coauthorship matrix and cooperativity matrix

Two *co-occurrence matrices* are associated with the paper-author matrix: 1) the *coauthorship matrix* lists the number of papers coauthored by each pair of authors, and 2) the *author coupling matrix* lists the number of common authors for each pair of papers. These matrices are analogous, respectively, to the co-citation matrix and the bibliographic coupling matrix for paper-reference matrices (Morris, 2005).

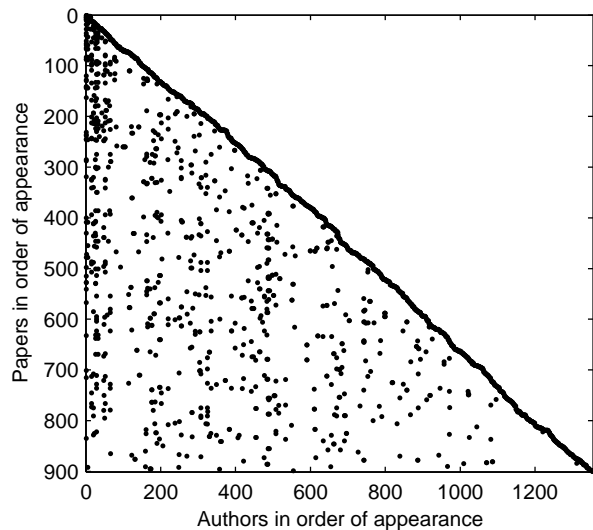


FIG. 2 Diagram of a paper-author matrix taken from a collection of papers on complex networks. Note the linear increase of authors with addition of papers, which is in accordance with the creation mechanism of the Yule process.

The author coupling matrix is useful for studying team oeuvres, that is, the body of work published by individual research teams. It is possible to include author coupling of papers in the model of paper-author networks being presented here, and it is also possible to derive network metrics based on author coupling that are similar to those that will be introduced in Section 4 that are derived from the coauthorship matrix. However, for the sake of brevity, the author coupling matrix will be excluded from the scope of this paper.

The *cooperativity matrix* is a binary matrix that lists the pairs of authors in a collection that have one or more coauthored papers. The cooperativity matrix can be obtained by setting all non-zero elements of the coauthorship matrix to unity, and setting the diagonal to zero. Coauthorship matrices tend to be very sparse, with most elements falling as blocks straddling the matrix diagonal, as shown in Figure 3. Each block on the diagonal corresponds to a group of new authors that appeared when a new paper was added to the collection.

#### 4. NETWORK METRICS

Metrics are defined here as mathematical features of a network that can be used to compare the characteristics of the network to those of other networks. Network metrics will be used in this paper to compare actual paper-author networks to networks produced by simulation of the proposed model.

For the purpose of evaluating the proposed model, we will use six metrics. These metrics serve as indicators of processes within the specialty: team size, participation threshold of authorship, researcher productivity and

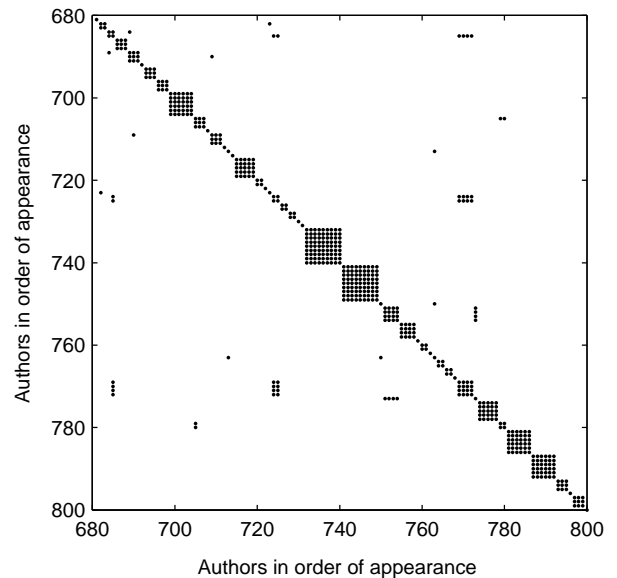


FIG. 3 Example of a coauthorship matrix. The blocks along the diagonal correspond to coauthorship links among newly appearing authors as papers are added.

longevity, level of collaboration and longevity of collaboration links, tendency to work in teams, and level of weak ties in the specialty.

#### 4.1. Authors per paper distribution

This is the distribution of the number of authors of each paper. This metric is important because it indicates the distribution of the number of researchers that work on each task in a specialty.

The description of the process generating the number of authors per paper, and the resulting authors per paper distribution, are key parts of the model being proposed in this paper. So it is important to make a detailed discussion of that distribution here. As discussed in Section 2.6, we assume a Poisson model of the number of secondary authors per paper to model the process that lead authors use to select their secondary authors when they write a paper. The Poisson distribution logically fits the proposed process, and further has the advantage of being a simple, intuitive, one-parameter distribution. Realistically, of course, the social processes being modeled are both very noisy and full of second-order effects that cannot be reliably modeled. It is understood then, that the 1-shifted Poisson distribution is proposed here as only a first order approximation of the author per paper distribution.

Despite the reservations discussed above, there is a great deal of evidence that the 1-shifted Poisson distribution approximately describes the authors per paper distribution in real collections of papers from "little science" fields. Price and Beaver (1966) note an approximate Poisson distribution of secondary authors in a biomedical

specialty. Beaver (1978) and Beaver and Rosen (1979), in their seminal papers on collaboration, assert that the distribution of the number of secondary authors tends to be a Poisson distribution for "little science". Seglen and Aksnes (2000), studying the field of microbiology, find that the distribution of the number of secondary authors in their study fits a Poisson distribution.

A large study by Glänzel (2002) of papers in the fields of mathematics, chemistry, and biomedicine, indicates, upon analysis, a 1-shifted Poisson distribution for authors per paper in these three fields, although there is a definite trend in the field of biomedicine away from that distribution. The original data from Glänzel's study, kindly supplied to the authors of this paper by Wolfgang Glänzel, is plotted in Figure 4 and compared to fitted 1-shifted Poisson distributions, showing a good match.

Several studies of the author per paper distribution have focused on trial and error fitting of various distributions to empirical data to determine the distribution that tends to fit the largest number of data sets. Ajiferuke (1991) proposed a process of multiple authoring that would produce a Waring distribution of authors per paper, but found that the inverse Gaussian-Poisson distribution tended to fit the largest number of data sets. Gupta, *et al.*, (1998) and Rousseau (1994) found that the zero truncated Poisson distribution and the geometric distribution tended to produce acceptable fits to many datasets in Library and Information Science.

For purposes of simulation here, the 1-shifted Poisson parameter is estimated from the network to be modeled and a 1-shifted Poisson random deviate generator is used to simulate the distribution of authors per paper. The authors per paper distribution is calculated by summing over the rows of the paper-author matrix to get a column vector of the number of authors for each paper, then constructing a frequency table of the elements of this vector.

#### 4.2. Papers per author distribution

This is the distribution of the number of papers authored by each author. As reported by Lotka (1926), this distribution is usually well approximated by a zeta distribution, with an exponent of approximately 2. The zeta distribution (Johnson *et al.*, 1992) is a power law which mathematically describes the "core and scatter" distribution of authors in a specialty (White and McCain, 1989). The core group is a small group of highly productive authors that are fully engaged in the specialty, while the "scatter" authors are those transient and less-productive authors that make up the rest of the specialty. This is an important metric because it describes the distribution of productivity among researchers, modeling the formation of core groups of researchers in the specialty. See Morris (2005) for a discussion of the size of core and scatter groups of authors in a specialty. The paper per author distribution is calculated by summing down the columns

of the paper-author matrix to get a row vector that lists the number of papers each author has published, then constructing a frequency table from this vector.

#### 4.3. Coauthorship per author pair distribution

For brevity, we shall refer to the coauthorship per author pair distribution as the coauthorship distribution. This is the distribution of the number of papers that each pair of authors have coauthored, taken over all pairs of authors in the collection. This is an important metric that measures the tendency of authors to work together repeatedly on consecutive tasks. The coauthorship distribution is calculated from the coauthorship matrix, discussed in Section 3.2. The distribution is found by building a frequency table of elements in either the lower or upper triangle, excluding the matrix diagonal, of this coauthorship matrix.

#### 4.4. Cooperativity distribution

The *cooperativity distribution* is the distribution of the number of cooperating authors of each author, taken over all authors in the collection. This is an important metric that measures the tendency of researchers to collaborate with others on tasks and shows the diversity of collaborators of individual authors in a specialty. Braun, Glänzel, and Schubert (2001) use this distribution to analyze collaboration of authors as a function of productivity. The distribution is calculated from the cooperativity matrix, whose construction from the coauthorship matrix is described in Section 3.2. Sum along the rows (or columns) of the cooperativity matrix to obtain a vector of the number of cooperating authors per author. The distribution is found by building a frequency table of the elements of this vector.

#### 4.5. Coauthor clustering coefficient distribution

The clustering coefficient was first introduced by Watts and Strogatz (1998) as a scalar mean clustering coefficient. Assume a network of authors linked by coauthorship, that is, a pair of authors are linked if they are cooperating authors, and are not linked otherwise. Given an author, the neighbors of that author are the set of all authors linked to him/her. The clustering coefficient for a single author is the number of links among the author's neighbors divided by the number of possible links among those neighbors. Given author  $i$ , with  $k_i$  neighbors, and  $y_i$  links among those neighbors, the clustering coefficient,  $c_i$ , of author  $i$  is:

$$c_i = \frac{y}{\frac{1}{2}k_i(k_i - 1)} \quad (2)$$

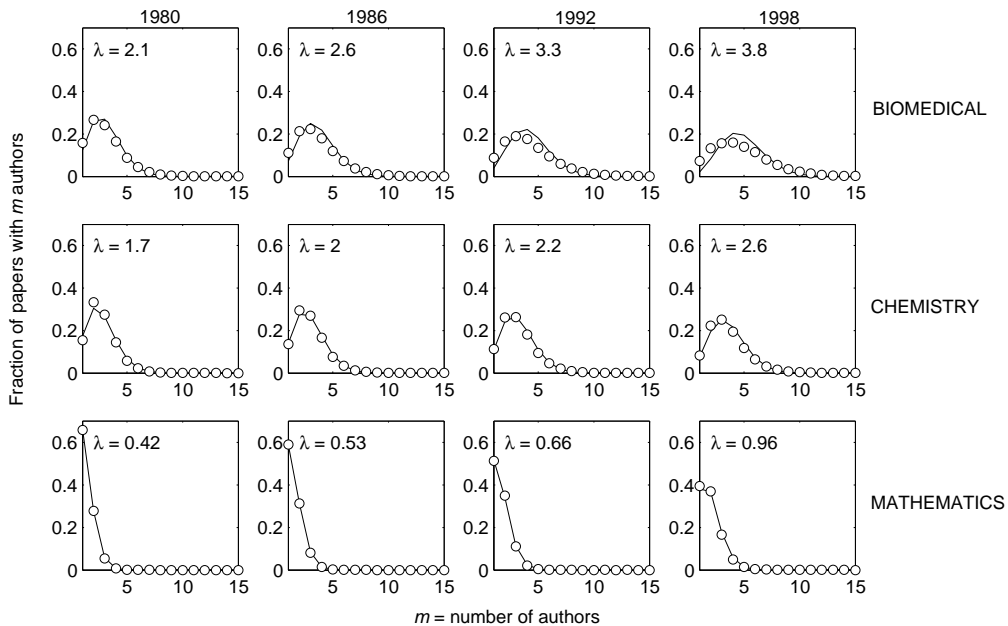


FIG. 4 Author per paper distribution of three fields of science from Glänzel (2002). The columns show data from periods ending in 1980, 1986, 1992, and 1998 respectively. The rows show data from biomedical research, chemistry, and mathematics respectively. Circles are experimental data, lines are MLE fitted Poisson distributions. The fitted distributions match the experimental data well in shape, position and scale. Note the trend in biomedical research away from the 1-shifted Poisson distribution as a function of time.

The clustering coefficient is not defined for authors with zero links or authors with only one link. Usually, in the complex networks literature (Albert and Barabasi, 2002), the clustering coefficient is taken as the mean of  $c_i$  over all nodes in the network, in this case all authors with two or more links. We define the coauthor clustering coefficient distribution as the distribution of  $c_i$  over all authors that have two or more links.

The clustering coefficient metric is important because it measures the tendency of authors to work in local groups, and is very useful for evaluating a team-based model of collaboration. The calculation method used is a simple method applied author by author. Given an author, a neighbor matrix is extracted by extracting the rows and columns of the author's neighbors from the cooperativity matrix. The fraction of off-diagonal elements of the neighbor matrix that are non-zero is the clustering coefficient for that author.

#### 4.6. Minimum path length distribution

In a network, the minimum path length between two nodes is defined as the minimum number of links that must be crossed to travel from one node to the other. This is a measure of the connectedness and ease of communication in the network. Kretschmer (2004) used the minimum path length distribution in a coauthorship network to show that highly productive authors have shorter communication paths in a specialty than low productiv-

ity authors.

For paper-author networks the minimum path length distribution is quite sensitive to the effect of weak ties in the network between author teams. As such, this metric will be used to validate the ability of the proposed model to mimic the effects of inter-group ties on communication in the collaboration network. Note that this distribution is somewhat unstable, and susceptible to artifacts caused by outlier papers that have very large numbers of authors. In the examples that will be shown in Section 6, all papers with over 20 authors were removed from the collections. Readers are advised to compare minimum path length distributions on their gross features only.

## 5. GROWTH MODEL OF TEAMS

### 5.1. Introduction

The proposed growth model is based on several elements: 1) linear growth, 2) teams, 3) success-breeds-success of teams, 4) success-breeds-success of authors, and 5) random weak ties among teams. There is qualitative evidence for each of these elements in the structure of the paper-author matrix. Figure 5 shows a detailed view of a section of a paper-author matrix from a collection of papers on the topic of angiogenesis, a biomedical research topic. Rows in this matrix correspond to papers, arranged in order of publication, while columns correspond to authors, in order of appearance. Dots on

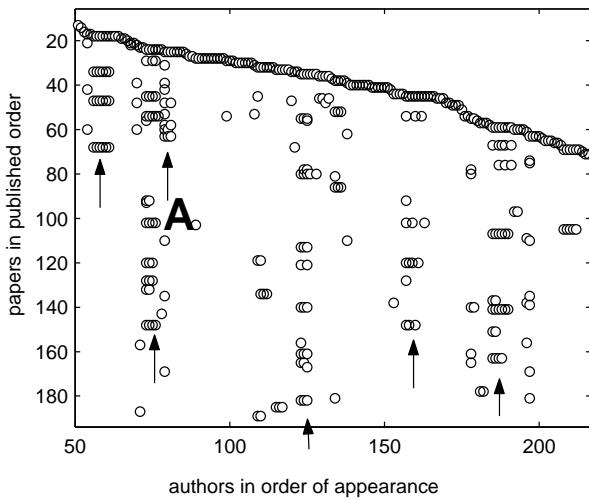


FIG. 5 Diagram of part of a paper-author matrix. Arrows point to matrix columns that show groups of authors that repeatedly coauthor papers. Note that authors from these groups usually appear as subsets of larger groups, and that most groups have dominant authors that publish many papers, for example, the first author of group "A."

the diagram correspond to ones in the matrix, a dot exists at position  $(i, j)$  of the diagram if author  $j$  was an author of paper  $i$ . The stair-step pattern of ones along the top of the diagram corresponds to the appearance of new authors in the collection, above this stair step there are no ones in the matrix. Notice that almost all papers produce some new authors, and that, while the number of new authors varies from paper to paper, the number of authors appears to be a roughly linear function of the number of papers. This is especially evident in Figure 2, which clearly shows the linear growth of authors in a collection with the addition of new papers. *This clearly supports the idea of a linear growth-based model of the paper-author network.*

Notice in Figure 5 that there are groups of adjacent ones in the matrix, identified by arrows in the diagram, that repeat vertically as more papers are added to the collection. These repeat groups correspond to groups of authors that repeatedly coauthor papers. In most of the repeating groups the specific authors from the group tend to vary from paper to paper. Also, there appear to be dominant authors in each of these groups that consistently appear in that group's papers. Note, for example, the group identified as "A" in the diagram. The first paper from this group has 7 authors, but various subsets of the first three authors of this group repeat four times. The first author repeats 11 times, often appearing with groups of new authors. *These observations of repetition of subsets of authors, and the existence of dominant authors, support the presumption that researchers work on tasks as teams and publish as teams, that papers are published by subsets of researchers from teams, and that productivity of authors inside of teams is driven by a success-breeds-success process.*

## 5.2. Proposed team-based Yule model

Yule originally proposed a model of biological evolution based on the principle of success-breeds-success (Yule, 1924). The Yule model was analyzed by Simon and applied to word frequencies (Simon, 1955). The model was also used by Price to model the growth of reference networks and other bibliographic networks (Price, 1976). The Yule model has also been repeatedly applied to modeling of author productivity (Chen, 1989, 1994). Several refinements to the Yule model have been suggested over the years to provide a better match of experimental data to the theoretical curve (Vukovic, 1998). For a good review of the Yule model, see Newman's review of complex network theory (Newman, 2003).

In the model proposed here the Yule model is applied to teams of authors of fixed size. Figure 6 shows a diagram of the proposed team-based Yule process for modeling paper-author networks. The growth of the network proceeds one paper at a time. The number of authors of each paper is a random number,  $m(\lambda)$ , drawn from a 1-shifted Poisson distribution with parameter  $\lambda$  :

$$p_m(k) = \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}, \quad k = \{1, 2, \dots\} \quad (3)$$

Where  $p_m(k)$  is the probability that a paper will have  $k$  authors. When a new paper appears there is a probability,  $\alpha$ , that the authoring team is a newly appearing team. In this case a new team is created with a fixed number of researchers,  $N_G$ . If the authoring team is an existing team, then the existing team is selected from the set of existing teams with probability:

$$p_t(q) = \frac{q}{np} \quad (4)$$

where:

- $q$  is the number of papers that the team has published
- $np$  is the number of papers in the collection
- $p_t(q)$  is the probability that the team will author the newly appearing paper.

As in the case of a new team,  $m(\lambda)$  authors, drawn from the 1-shifted Poisson distribution, are selected to author the paper. For each author there is a probability,  $\beta$ , that the author is drawn from outside the team. In such a case the author is drawn randomly from the set of all outside team members in the collection, whether they have authored a paper or not. If a researcher outside the team is not chosen as an author, the selection from within the team is done by another preferential attachment process, modified to allow selection of authors that have never published a paper. The probability of selecting an author  $i$  in the team is:

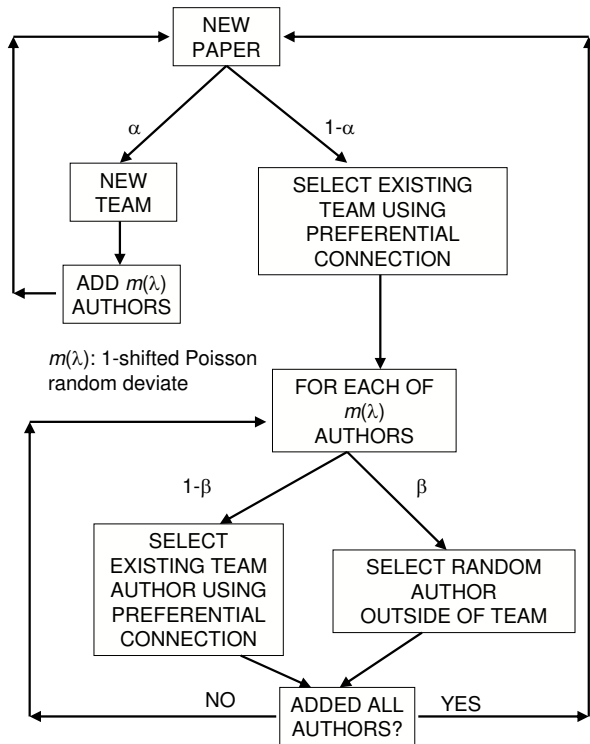


FIG. 6 Diagram of team-based nested Yule process of growth of paper-author networks. The branch on the left under "new paper" corresponds to creation of teams. The branch on the right under "new paper" corresponds to selection of an existing team using success-breeds-success. In this branch, success-breeds-success is used to select authors in the team, with a branch for random selection of outside team members as authors.

$$p_a(i) = \frac{k_i + 1}{\sum k_j + N_G} \quad (5)$$

where  $k_i$  is the number of papers written by author  $i$ ,  $\sum k_i$  is the sum of the number of authorships among the team, and  $N_G$  is the number of authors in the group. This is a preferential attachment process that favors authors by the number of papers they have published. The paper creation cycle described above, and shown in Figure 6, proceeds until the desired number of papers has been added to the collection.

In summary, the model has four parameters:

- $N_G$  - team size
- $\lambda$  the 1-shifted Poisson parameter for authors per paper
- $\alpha$  the Yule parameter for the team selection process
- $\beta$  the probability that an author on an existing team's paper will be from outside the team

### 5.3. Discussion of the applicability of the proposed model

The growth model that was introduced in Section 5.2 is a fairly simple model that is easy to understand. It incorporates three processes that are often discussed in the field of complex networks: 1) linear growth, 2) preferential attachment, and 3) random connections as weak ties.

*Preferential attachment* is simply a model of the success-breeds-success process, modeling the probability of selection of an entity as proportional to the number of times the entity has been selected before, mathematically expressed using Equation 4. The Yule model is probably the original mathematical expression of preferential attachment. For unipartite networks the Barabasi-Albert model (A. Barabasi and Albert, 1999) is probably the most well-known preferential attachment model, and was applied to collaboration networks by Barabasi, *et al* (2002). *Unipartite networks* are networks whose nodes are exclusively of one entity type. Examples are networks of authors connected by coauthorship, or networks of references connected by co-citation links. Paper-author networks are *bipartite networks*, consisting of two partitions of nodes, one partition contains papers, the other partition contains authors. All links in a bipartite network are across the network partition, in this case, links are always between authors and papers.

*Small-world networks* were first proposed by Watts and Strogatz as a method of modeling weak ties (Watts and Strogatz, 1998). A small world network is a regular lattice structure, where the mean minimum path length between nodes is large. Random links are added to the lattice, which does little to change the structure of the network, but which drastically shortens the mean minimum path length between nodes. In the context of our paper-author network, the lattice is comprised of a regular structure of teams that are isolated from each other. Considering links as coauthorships, then the authors are isolated from all other authors except their co-workers. The addition of random inter-team collaboration links significantly shortens path length between authors not of the same team.

In the context of complex network theory, the proposed model can be considered a composite model. The model is nested in the sense that preferential attachment of authors is modeled within teams which are preferentially attached at a higher level. The model is additionally a Watts-Strogatz small world model in the sense that random connections are used to model weak ties between teams of researchers.

### 5.4. Estimation of model parameters

The first parameter,  $\alpha$ , is obtained by determining the probability of new team creation. This probability is estimated by making a paper-by-paper pass through the network to determine the fraction of papers that appear

with a completely new set of authors.

The parameter  $\lambda$  is calculated by dividing the total number of authorships by the total number of papers and subtracting 1 (1-shifted Poisson estimate.) The number of authors per team,  $N_G$ , is chosen heuristically as 20, which is assumed to be the upper limit on the number of researchers that can effectively interact as a team. One of the examples presented here, however, gave slightly better results by using  $N_G = 10$ . This particular example, distance education, is a specialty with very little collaboration and no core authors. The "weak tie" parameter,  $\beta$ , is estimated by matching the coauthorship distribution to the actual data using trial and error. A value of 0.1 for  $\beta$  gives consistently good results.

To summarize, the parameters  $N_G$  and  $\beta$  are usually set to 20 and 0.1 respectively, and the parameters  $\alpha$  and  $\lambda$  are estimated from the actual data using the methods outlined above. From our limited experience it appears that the parameters  $\alpha$  and  $\lambda$  are negatively correlated, that is, as  $\lambda$  increases,  $\alpha$  decreases. This implies that as team size increases ( $\lambda$  increases), probably as a result of increasing institutionalization in the specialty, then author teams tend to be less transient, and stay in the specialty longer, because decreased  $\alpha$  means the percentage of papers authored by existing teams increases. This appears logical, because increased institutionalization reflects the availability of more research funds in a specialty, enabling teams of researchers to remain in the specialty longer.

## 6. EXAMPLES

The examples test the proposed model across a range of collaboration intensities. The first example, complex networks, is from the field of physics, and is presented as "typical" team-based research, with a Collaboration Coefficient ( $CC$ ) (Ajiferuke *et al.*, 1988) of 48%. It is not a specialty that has a significant number of multi-team projects. The second example, distance education, is from the field of education, and is extreme in its lack of collaboration. It has  $CC$  of 30% and has no significant core of highly productive authors. The third example, atrial ablation, is from the biomedical field. It has a great deal of collaboration, with  $CC$  of 72%, and has a well defined set of core authors that publish heavily in the specialty.

We will simulate each of the collections using the proposed model and compare metrics from those simulations to the metrics calculated from the actual collections of papers. It is quite difficult to make quantitative comparison of these metrics, nor is it very useful. The model is chiefly proposed as an instrument of insight into the manifestation of team-based research in paper-author networks, and is not expected to match actual data closely. It is only desired that the comparisons demonstrate that the model can produce simulations that reasonably mimic the characteristics of the paper-author

networks they model simultaneously across several non-trivial metrics. Because of this, comparisons of metrics will be limited to visual comparisons of metrics from the simulated data and with those of actual collections.

### 6.1. Complex networks literature

The complex networks collection was gathered on September 8th, 2003, from ISI's Web of Science product. This collection was used by Morris (2005) in a study of growth modeling of paper-reference networks. It was also used in a preliminary study of the paper-author growth model being proposed in this paper (Goldstein *et al.*, 2005). The collection was gathered by finding all papers that satisfy the following queries:

- cites references with (author=BARABASI-AL AND year=1999 AND journal=SCIENCE)
- cites references with (author=WATTS-DJ AND year=1998 AND journal=NATURE)
- cites references with (paper=ALBERT-R AND year=2000 AND journal=NATURE)
- cites references with (paper=ALBERT-R AND year=2002 AND journal=REV-MOD-PHYS)
- cites references with (paper=DOROGOVTSSEV AND year=2002)

The queries above yielded 832 papers. An additional 386 papers were added from a previously studied collection (Morris and Yen, 2004) that was gathered on February 2nd, 2003, using the above queries and additional queries of all papers citing authors NEWMAN-ME and PASTORSATORRAS-R. After elimination of duplicates, 932 papers remained. These were clustered using bibliographic coupling and 30 papers in clusters that were obviously off-topic were discarded. Two papers having more than twenty authors were discarded. In total there are 900 papers in the collection, linked by 2274 authorships to 1354 authors.

**Model parameters.** The model parameters for the complex networks collections were estimated using the techniques outlined in Section 5.4 with  $\hat{\alpha} = 0.33$ ,  $\hat{\beta} = 0.1$ ,  $\hat{\lambda} = 1.53$ , and  $\hat{N}_G = 20$ . These parameters indicate that about a third of new papers correspond to the appearance of a new team in the specialty, and that about a tenth of the authorships are inter-team collaborations. The mean number of authors per paper is about 2.5.

**Comparison of model simulation to actual data.** Figure 7a shows the comparison of simulated authors per paper frequencies against the actual frequencies. The overall match is good, with minor discrepancies between actual and simulated frequencies, probably from deviations in the actual data from the 1-shifted Poisson distribution. The comparison of papers per author distribution is shown in Figure 7b. Note there is a fairly close

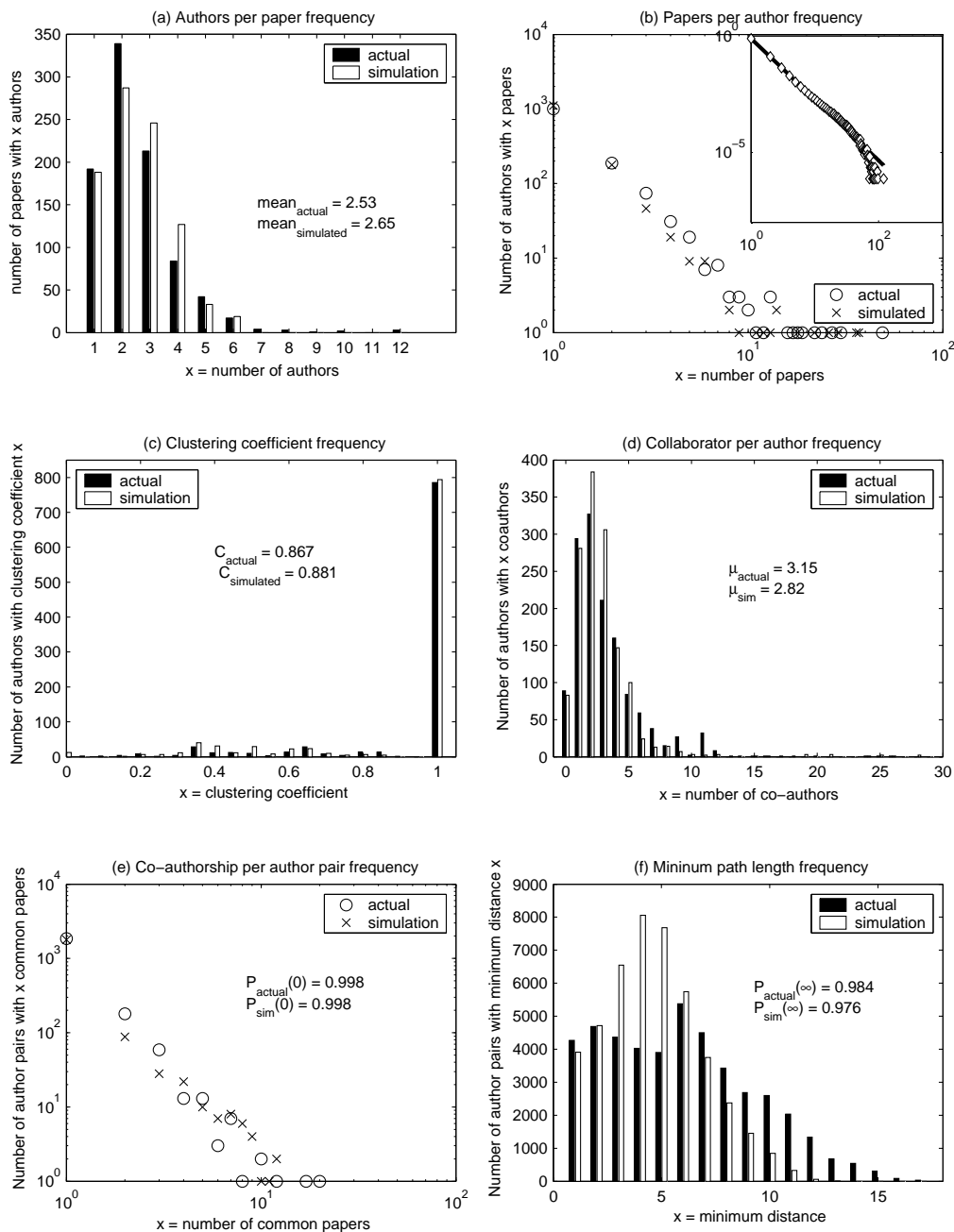


FIG. 7 Comparison of model simulation metrics against actual metrics from a collection of papers covering complex networks theory.

match of the model simulated frequencies to the actual frequencies. The estimated power-law exponent for actual data is 2.54, while the estimated power-law exponent for the simulation is 2.77. The inset of Figure 7b shows the model-predicted paper per author distribution, generated by gathering statistics from 1000 simulations. This predicted distribution models Lotka's law, producing an excellent fit to a zeta distribution with an exponent of 2.77, which is plotted in the inset. The effect of truncating the simulations at 900 papers is evident in the divergence of the simulation distribution from the zeta

distribution in the extreme tail of the distributions. Figure 7c shows a comparison of the coauthor clustering coefficient distribution for the complex networks collection. The simulation mimics the actual data well, including the spike at unity. Figure 7d shows the comparison of the cooperativity frequencies. There is good general agreement between actual and simulated frequencies in this plot. Figure 7e shows a comparison of the coauthorship distribution from simulated and actual data, presented on a log-log plot. Visually, this distribution appears to be a power-law. The plot shows the percentage of un-

connected author pairs as  $P(0)$ , with both simulation and actual data showing this fraction at about 99.8%. The frequencies of coauthorships also agree well between actual and simulated data. Figure 7f shows a comparison plot of minimum path length distribution. The plot shows the percentage of author pairs,  $P(\infty)$ , that have no path of coauthorship links between them. Both the simulation and the actual data show that this percentage is about 98%. The general shape and scale of the distributions agree, though there are distinct differences in frequencies at path lengths from 3 to 5, and the actual distribution appears to have a longer tail than the simulated distribution.

## 6.2. Distance education

The distance education collection was gathered from ISI's Web of Science product on May 16th, 2002. The papers were gathered using the query: (topic = "distance education" OR topic = "distance learning" OR topic = "e-learning"). The collection is 1391 papers linked by 2838 authorships to 2472 authors. It should be noted that the top journal for this specialty, defined as the journal receiving the most citations, is not indexed by the Science Citation Index. Because of this, no papers from this journal, *The American Journal of Distance Education*, are included in this collection. The EBSCO database service lists 170 papers in this journal from its founding in 1987 until May, 2002. These missing 170 papers probably lowered the published paper counts of core authors in the specialty in this collection.

**Model parameters.** The model parameters for the distance education collection were estimated using the techniques outlined in Section 5.4, with  $\hat{\alpha} = 0.65$ ,  $\hat{\beta} = 0.1$ ,  $\hat{\lambda} = 1.04$ , and  $\hat{N}_G = 10$ . These parameters indicate that about two thirds of new papers correspond to the appearance of a new team in the specialty, and that about a tenth of authorships are inter-team collaborations. The mean number of authors per paper is about 2.

**Comparison of model simulation to actual data.** Figure 8a shows the comparison of simulated authors per paper frequencies against the actual frequencies. While the frequencies generally agree in shape and scale, there are differences between actual and simulated frequencies, probably from deviations in the actual data from the 1-shifted Poisson distribution. Note that the actual data has significantly more single authors than the simulation. It is possible that the selection of secondary authors is a mixture of a Poisson process and a second process that doesn't produce secondary authors. The comparison of papers per author distribution is shown in Figure 8b. There is a fairly close match of the model simulated frequencies to the actual frequencies. The estimated power-law exponent for actual data is 3.7, while the estimated power-law exponent for the simulation is 3.55. Such a high exponent indicates a specialty with no core researchers. The inset of Figure 8b shows the

model-predicted distribution, generated by gathering statistics from 1000 simulations, showing that the model's predicts Lotka's Law with some divergence in the tail due to truncation at 1391 papers. Figure 8c shows a comparison of the coauthor clustering coefficient distribution. The simulation mimics the actual data fairly well. Figure 8d shows the comparison of the cooperativity frequencies. While there is crude agreement in shape and scale of simulated frequencies against actual frequencies, there are some differences. Figure 8e shows a comparison of the coauthorship distribution from simulated and actual data, presented on a log-log plot. There is some difference between simulation frequencies and actual frequencies, but not in the slope of the curves, although it is difficult to estimate slope of curves from only the four points on the plot. Both simulation and actual data show that 99.9% of the author pairs have zero coauthorships. Figure 8f shows a comparison plot of minimum path length distribution for the distance education collection. The plot shows the percentage of author pairs,  $P(\infty)$ , that have no path of coauthorship links between them. Both the simulation and the actual data show that this percentage is about 99.9%. The general shape and scale of the distributions agree.

## 6.3. Atrial ablation

The atrial ablation collection was gathered from ISI's Web of Science product on July 14th, 2003. The papers were gathered using the query: (topic = atrial AND topic = ablation), which gathered 3095 papers. Among these, 4 papers were discarded as having over 20 authors. The collection is 3091 papers linked by 16,785 authorships to 6409 authors. This is an example of a specialty where tasks often require more than one team to accomplish, and much collaboration is present. It is perceived as being on the high end of the scope of the proposed model as shown in Figure 1.

**Model parameters.** The model parameters for the atrial ablation collection were estimated using the techniques outlined in Section 5.4:  $\hat{\alpha} = 0.22$ ,  $\hat{\beta} = 0.1$ ,  $\hat{\lambda} = 4.43$ , and  $\hat{N}_G = 20$ . These parameters indicate that about a quarter of new papers correspond to the appearance of a new team in the specialty, and that about a tenth of authorships of papers correspond to inter-team collaborations. The mean of authors per paper is about 5 or 6 authors.

**Comparison of simulation to actual data.** Figure 9a shows the comparison of simulated author per paper frequencies against the actual frequencies. While the plots generally agree in shape and scale, there are differences between actual and simulated frequencies, probably from deviations in the actual data from the 1-shifted Poisson distribution. The comparison of papers per author frequencies is shown in Figure 9b. Note that in the figure there is a fairly close match of the model simulated frequencies to the actual frequencies. The esti-

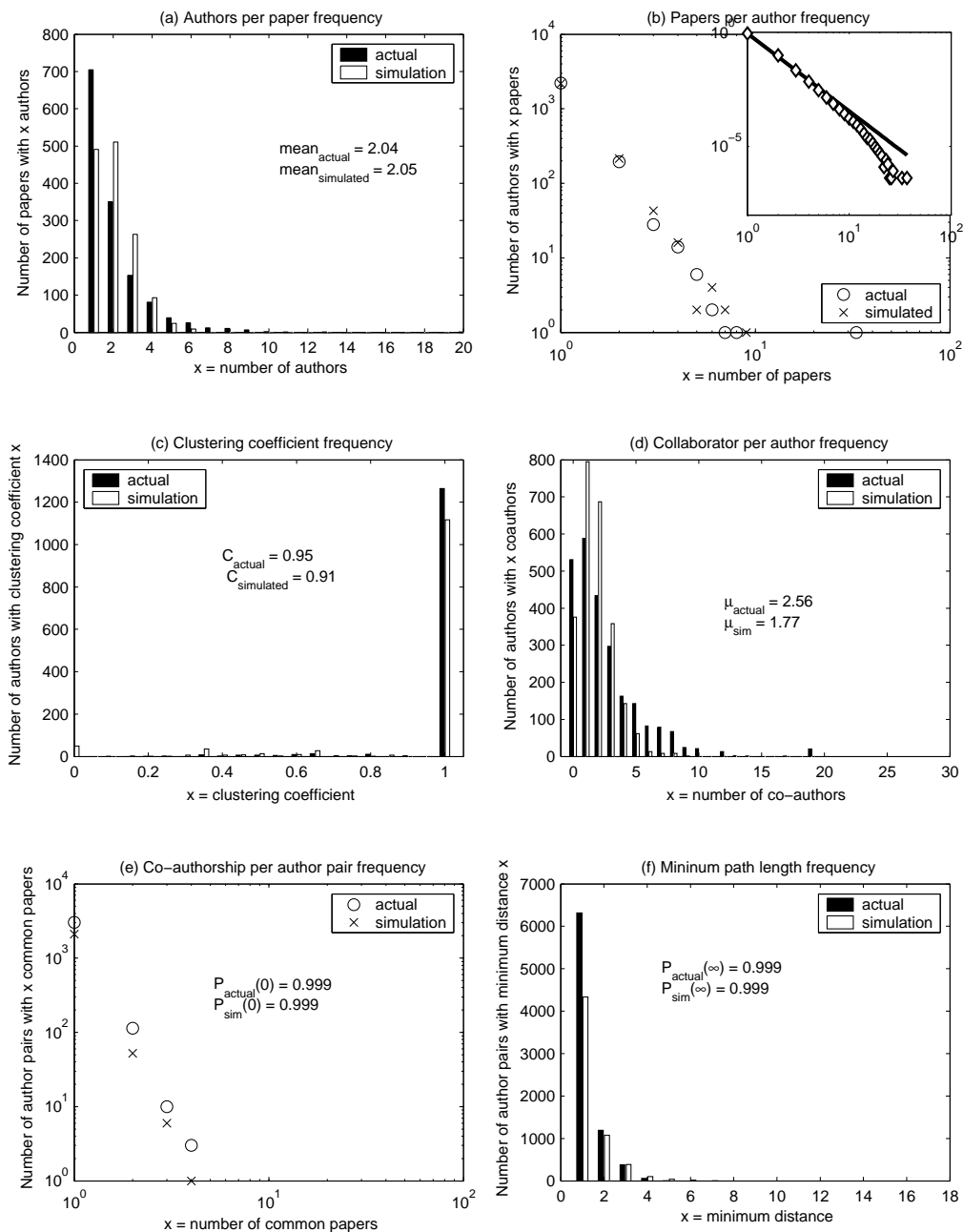


FIG. 8 Comparison of model simulation metrics against actual metrics from a collection of papers covering distance education.

mated power-law exponent for actual data is 2.25, while the estimated power-law exponent for the simulation is 2.15. The inset of Figure 9b shows the model-predicted paper per author distribution, generated by gathering statistics from 1000 simulations. This predicted distribution follows the fitted zeta distribution well, with slight distortions from linearity on the plot. Divergence in the extreme tail of the distribution is presumably caused by truncation of the simulations at 3091 papers. Figure 9c shows a comparison of the coauthor clustering coefficient distribution for the distance education collection. There is general agreement. Figure 9d shows the comparison

of the cooperativity frequencies. There is good agreement in shape and scale of simulated frequencies against actual frequencies, with few specific differences. Figure 9e shows a comparison of the coauthorship distribution from simulated and actual data, presented on a log-log plot. Simulation and actual data on the plot appear similar, with some slight divergence of the slopes. Figure 9f shows a comparison plot of minimum path length distribution for the atrial ablation collection. The general shape of the distributions agree but the scale of the simulated data is magnified and this is especially evident at path lengths above 7 or 8.

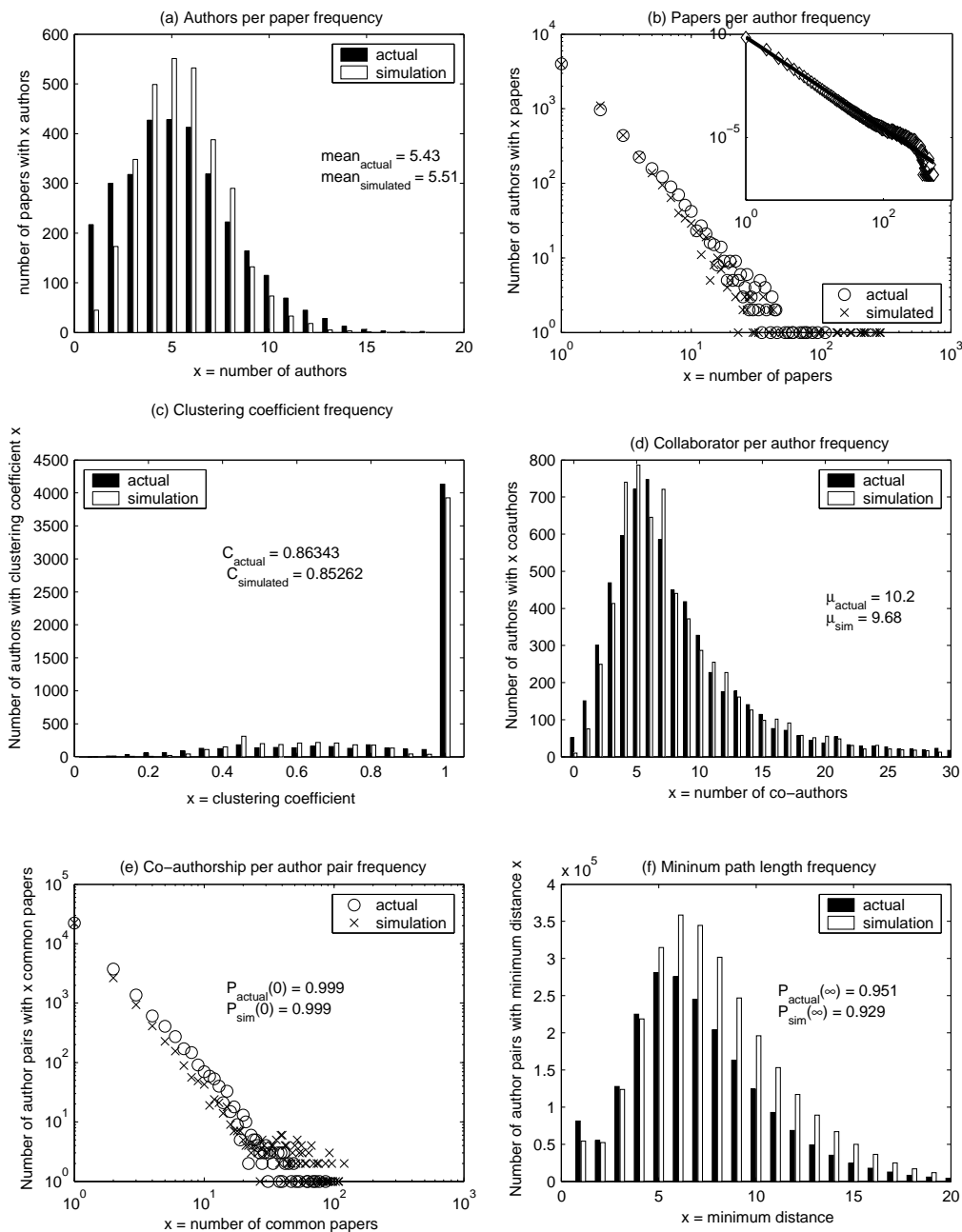


FIG. 9 Comparison of model simulation metrics against actual metrics from a collection of papers covering atrial ablation.

## 7. CONCLUSION

We have introduced a qualitative team-based model of research in a specialty. The scope of this model is assumed to be for specialties where most tasks are unitary and not divisive. In summary the elements of this qualitative model are as follows:

- Researchers work in small teams of less than 20 members.
- The teams work on unitary tasks. Each task is assigned to a team member to be accomplished.

- Team members seek assistance of their co-workers, and co-workers provide assistance at varying levels of participation.
- Upon completion of the task the lead researcher becomes the writer and lead author of a report on the task that will be submitted as a journal paper.
- Co-workers that helped in the task above a participation threshold will be selected by the lead author to be secondary authors. This results in a Poisson distribution of the number of secondary authors

and a 1-shifted Poisson distribution of the number of authors.

- Successful lead researchers will undergo a process of success-breeds-success within their team for assignment of tasks and production of papers.
- Teams undergo a success-breeds-success process for acquiring funding and research projects, and this is reflected in the number of papers authored by successful teams.
- Teams seek outside assistance on tasks when they lack appropriate resources and expertise. This assistance results in random collaborations with outside researchers on tasks.

From this qualitative model we constructed a quantitative mathematical growth model based on a nested success-breeds-success process. Using a fixed team size, a Yule process models the appearance of teams as papers are produced in a specialty's literature, and further models the success-breeds-success process of repeated team authorship. At a lower level, however, within teams the production of papers by authors in the team is modeled by a success-breed-success process. The selection of the number of the subset of team members that will author a paper is drawn from a 1-shifted Poisson distribution. With a fixed, small probability, authors may be chosen from outside an authoring team, modeling random inter-team collaboration and weak ties.

The resulting mathematical model has four parameters: 1)  $\alpha$ , the Yule parameter for creation of new teams, 2)  $\lambda$ , the parameter for the 1-shifted Poisson distribution of authors per paper, 3)  $N_G$ , the fixed size of teams, and 4)  $\beta$ , the probability that an author will be from outside the authoring team. The parameters  $N_G$  and  $\beta$  are stable and constant over the three examples presented here. The parameters  $\alpha$  and  $\lambda$  are most probably negatively correlated.

The model is not amenable to the derivation of closed form asymptotic distributions of productivity, but may well prove to be a useful simulation tool. The model can be used to construct data sets of synthetic collections by known teams which can be used to test the ability of clustering algorithms to find teams from coauthorship in collections of papers. The model effectively mimics one of the known facts about collaboration: not all collaborators on a task become authors on the papers that reports the task.

The proposed model was validated on three example collections of papers that covered the range of the scope of the model in terms of intensity of collaboration. The intended scope of the model was for specialties in which most of the tasks were unitary and accomplished by one researcher, with the assistance of others. The three examples ranged from very light collaboration in the distance education collection, through typical collaboration in the complex networks collection, to heavy collaboration in the atrial ablation collection. The simulations

from the proposed model were able to mimic well the 6 metrics used for comparison of actual to simulation.

Having shown the validity of the model through its match to a variety of collections of papers via the proposed quantitative model, the model provides great insight into the social processes of research. The model points to areas of further investigation that can provide more information on the processes of research, particularly on the team-based nature of research and its manifestation in journal literature. Is this team-based process driven principally by academic research and reporting? Does corporate research have the same team characteristics, and the same manifestations in the literature? Does the model extend into high collaboration research fields such as biomedicine with little or no modification? Is fixed team size a reasonable model? Are inter-team collaboration and weak ties really well modeled by random coauthorships? Does the model still mimic actual collections after elimination of transitory "scatter" authors? That is, are the metrics mimicked by the model when core authors only are studied?

As a final note, it is useful to compare the team-based model of paper-author networks proposed here with the exemplar-based model of paper-reference networks proposed by Morris (2005). Both types of networks produce well-known distributions that are power-laws: Lotka's law for paper-author networks, and the reference power-law (Naranan, 1971) for paper-reference networks. Because of the ubiquitous appearance of such power-laws in bibliometrics, it is tempting to look for 'universal processes' that can be applied in a general sense to any bipartite network occurring in collections of papers. For example, both Price (1976), and additionally Egghe and Rousseau (1995) develop models of universal bibliometric processes based on success-breeds-success. Success-breeds-success does appear to be a universal component in bibliometric processes. However, comparison of the team-based paper-author network model to the exemplar-based paper-reference model shows stark and fundamental differences. The team-based process is a social model and describes a fundamental two-tiered hierarchy of success-breeds-success in author productivity, while the exemplar-based process is a linguistic model of concept symbols, where success-breeds-success is more of a background process. Assuming the validity of these two models, considering these stark differences indicates that seeking universal bibliometric processes is somewhat problematic, in that it distracts from noticing the important differences among such processes. Surely, the interesting and meaningful characteristics of these processes are in the details of their fundamental differences.

## 8. ACKNOWLEDGEMENTS

We would like to thank Wolfgang Glänzel of the Hungarian Academy of Sciences for generously providing experimental data on the authors per paper distributions

used for his 2002 paper.

## References

- Ajiferuke, I. (1991). A probabalistic model for the distribution of authorships. *Journal of the American Society for Information Science*, 42(4), 279-289.
- Ajiferuke, I., Burell, Q., and Tague, J. (1988). Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics*, 14(5), 421-433.
- Albert, R., & Barabasi, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47-97.
- Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(509-512).
- 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.
- Beaver, D. D. (1978). Studies in scientific collaboration. Part I. The professional origins of scientific co-authorship. *Scientometrics*, 1, 65-84.
- Beaver, D. D. (1979). Studies in scientific collaboration. Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite. *Scientometrics*, 1(2), 133-149.
- Beaver, D. D. (1984). Teamwork: A step toward collaboration? In George Sarton Centennial, *Communication and Cognition* (pp. 449-452). Ghent, Belgium.
- Beaver, D. D. (2001). Reflections on scientific collaboration (and its study): Past, present, and future. *Scientometrics*, 52(3), 365-377.
- Beaver, D. D., & Rosen, R. (1979). Studies in scientific collaboration. Part iii. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, 1(3), 231-245.
- Braun, T., Glanzel, W., & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, 51(3), 499-510.
- Chen, Y. S. (1989). Analysis of Lotka's law: The Simon-Yule approach. *Information Processing & Management*, 25(5), 527-544.
- Chen, Y. S. (1994). The Simon-Yule approach to bibliometric modeling. *Information Processing & Management*, 30(4), 535-556.
- Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices. *Journal of the American Society for Information Science and Technology*, 52(7), 558-569.
- Egghe, L., & Rousseau, R. (1995). Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science*, 46(6), 426-445.
- Fox, M. F. (1983). Publication productivity among scientists: A critical review. *Social Studies of Science*, 13, 285-305.
- Glänzel, W. (2002). Co-authorship patterns and trends in the sciences (1980-1998). A bibliometric study with implications for database indexing and search strategies. *Library Trends*, 50(3), 461-473.
- Godin, B., & Gingras, Y. (2000). The place of universities in the system of knowledge production. *Research Policy*, 29, 273-278.
- Goldstein, M. L., Morris, S. A., & Yen, G. G. (2005). Group-based Yule model for bipartite author-paper networks. *Physical Review E*, 71, 026108.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 778(6), 1360-1380.
- Gupta, B. M., Kumar, S., & Rousseau, R. (1998). Applicability of selected probability distributions to the number of authors per article in theoretical population genetics. *Scientometrics*, 42(3), 325-334.
- Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions* (2nd ed.). New York: John Wiley & Sons.
- Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26, 1-18.
- Kretschmer, H. (2004). Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the web. *Scientometrics*, 60(3), 409-420.
- Laudel, G. (2002). What do we measure by co-authorships? *Research Evaluation*, 11(1), 3-15.
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317-323.
- Mahlck, P., & Persson, O. (2000). Socio-bibliometric mapping of intra-departmental networks. *Scientometrics*, 49(1), 81-91.
- Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363-377.

- Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, co-citation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, 56(12), 1250-1273.
- Morris, S. A., & Yen, G. (2004). Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences of the United States*, 101(suppl. 1), 5291-5296.
- Naranan, S. (1971). Power law relations in science bibliography- a self-consistent interpretation. *Journal of Documentation*, 27(2), 83-97.
- Newman, M. E. J. (2003). The structure and function of complex networks. *Siam Review*, 45(2), 167-256.
- Peters, H. P. F., & VanRaen, A. F. J. (1991). Structuring scientific activity by co-author analysis - an exercise on a university faculty level. *Scientometrics*, 20, 235-255.
- Porter, A. L., Roper, A. T., Mason, T. W., Rossini, F. A., & Banks, J. (1991). *Forecasting and management of technology*. New York: John Wiley and Sons.
- Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5-6), 292-306.
- Price, D. (1986). *Little science, big science- and beyond*. New York: Columbia University Press.
- Price, D., & Beaver, D. D. (1966). Collaboration in an invisible college. *American Psychologist*, 21, 1011-1018.
- Rousseau, R. (1994). The number of authors per article in library and information science can often be described by a simple probability distribution. *Journal of Documentation*, 50(2), 134-141.
- Seglen, P. O., & Aksnes, D. W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research. *Scientometrics*, 49(1), 125-143.
- Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440. Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.
- Subramanyam, K. (1983). Bibliometric studies of research collaboration. *Journal of Information Science*, 6, 33-38.
- Vukovic, V. O. (1998). Simon's generating mechanism: Consequences and their correspondence to empirical facts. *Journal of the American Society for Information Science*, 49(10), 867-880.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440-442.
- White, H. D., & McCain, K. W. (1989). *Bibliometrics*. *Annual Review of Information Science and Technology*, 24, 119-186.
- Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of dr. J. C. Willis, f. R. S. *Philosophical Transactions B*, 213, 21-87.