Bipartite Yule Processes in Collections of Journal Papers

Steven A. Morris

samorri@okstate.edu
Electrical and Computer Engineering, Oklahoma State University
202 Eng. So., Stillwater, Oklahoma, 74078, USA

Abstract
Collections of journal papers constitute a series of coupled bipartite networks that tend to exhibit linear growth and preferential attachment as papers are added to the collection. Assuming primary nodes in the first network partition and secondary nodes in the second network partition, the basic bipartite Yule process assumes that as each primary node is added to the network, it links to multiple secondary nodes, and with probability, α, each new link may connect to a newly appearing secondary node. The number of links from a new primary node follows some empirically measured distribution. Links to existing secondary nodes follow a preferential attachment rule. With modifications to adapt to specific networks, bipartite Yule processes simulate networks that can be validated against actual networks using a wide variety of network metrics. The application of bipartite Yule processes to paper-reference networks and paper-author networks is demonstrated and the results compare favorably to networks from actual collections of papers.

Collections of papers as coupled bipartite networks

As shown in Figure 1, a collection of journal papers constitutes a series of coupled bipartite networks (Morris, 2005). As diagrammed in Figure 1, a collection of papers contains six direct bipartite networks: 1) papers to paper authors, 2) papers to references, 3) papers to paper journals, 4) papers to terms, 5) references to reference authors, and 6) references to reference journals. Additionally, there are 15 indirect bipartite networks in collections of papers as defined by the diagram. Examples of interesting indirect networks are paper authors to reference authors, and paper journals to reference journals networks, which can be used for author co-citation analysis (White & Griffith, 1981) and journal co-citation analysis (McCain, 1991) respectively.

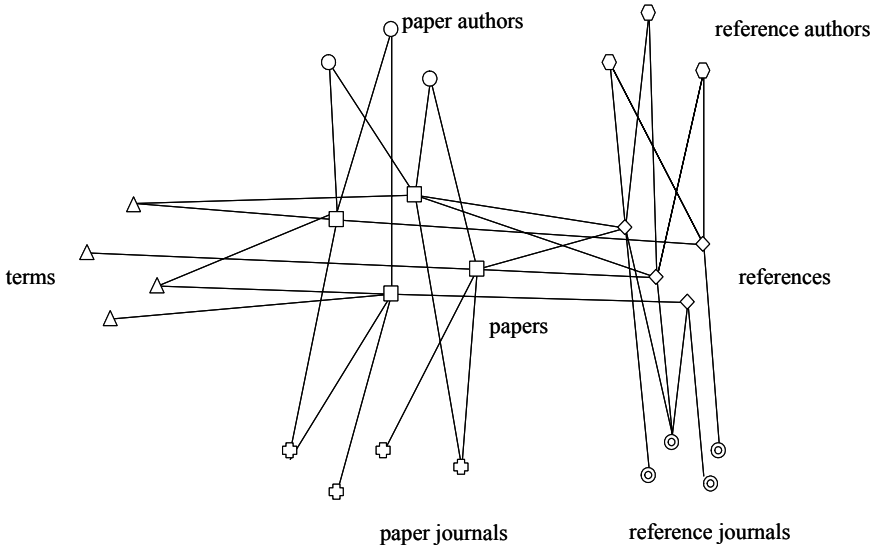

Figure 1. Diagram showing a collection of papers as a series of coupled bipartite networks.

Modeling the growth of these bipartite networks helps characterize the underlying processes driving a research specialty, such as knowledge accretion, researcher productivity, or collaboration processes. Bipartite growth models produce many network metrics, allowing comprehensive validation of models against real collections of papers.

Basic bipartite Yule processes

As originally proposed, Yule processes do not model networks, but simply model the formation of power laws of frequencies of entities. (Price, 1976; Simon, 1955). Yule processes are a mathematical

Proceedings of ISSI 2005, pages 316–321.
Edited by P. Ingwersen and B. Larsen.

expression of "success-breeds-success" phenomena (Newman, 2003), and have been applied to paper per reference distributions (Price, 1976), paper per author distributions (Chen, 1994), and paper per journal distributions (Vukovic, 1998).

For a bipartite Yule process, assume a bipartite network where nodes fall into two partitions: 1) primary nodes and 2) secondary nodes. Typically, primary nodes are papers while secondary nodes are entities that are associated with papers, such as authors, references, journals, or terms.

The rules for a basic bipartite Yule process are as follows:

- The network grows by adding primary nodes one at a time.
- When a new primary node is added, it links to N secondary nodes. N is a random deviate drawn from a discrete probability distribution that is a characteristic of the type of network being modeled. For paper-reference networks N is lognormally distributed (Morris, 2004), while for paper-author networks N is 1-shifted Poisson distributed (Goldstein, Morris, & Yen, 2004; Morris, Goldstein, & Deyong, 2004). For paper-journal networks, N is unity, since a paper is only linked to the journal in which it was published. As defined here, a primary entity does not link to any particular secondary entity more than once.
- For each of the N links, there is a probability, α, that it will link to a newly appearing secondary node.
- If a link happens to be to an existing secondary node, the linked node is selected using preferential attachment (Newman, 2004), that is, the probability of linking to a secondary node is proportional to the number of links that the node possesses. This models the success-breeds-success phenomenon, where, for example, references that have received many citations have a higher probability of being cited by newly appearing papers than references with few citations.

The stationary distribution of the link degree of the secondary nodes is a Yule distribution (Johnson, Kotz, & Kemp, 1992; Simon, 1955), a power law whose exponent is $1+1/(1-\alpha)$. The stationary distribution is independent of the distribution of N, but for finite collections of papers the distribution of N profoundly affects the tail of the distribution (Morris, 2004).

Practical bipartite Yule processes

In practice, the basic bipartite Yule process outlined in the proceeding section must be modified to account for the characteristics of the specific type of bipartite network being studied.

Paper-reference Yule process

Figure 2 shows a diagram of a bipartite Yule process modified for the characteristics of paper-reference networks. Such networks are characterized by the accretion of highly cited exemplar references, which are cited at rates far higher than would be predicted by simple preferential attachment (Morris, 2004). These exemplar references tend to appear during the initial growth of the network and their rate of appearance decreases exponentially as papers are added to the collection.

As each paper is added to the collection, it links to a lognormally distributed number of references, as discussed in (Morris, 2004). For each reference cited by a paper, there is a probability α that the citation is to a newly appearing reference. When a new reference appears, there is a small probability that the reference will be a highly attractive exemplar reference. If so, the reference receives a large initial attraction, A_0. Newly created non-exemplar references receive no initial attraction. If a citation is to an existing reference, the probability that an particular existing reference will be cited is proportional to the sum of its attraction plus the number of times it has been cited. A specific reference cannot be cited more than once by a paper. In paper-reference networks, the parameters m and γ usually constrain the ratio of references to papers in the collection to about 20, i.e., a collection of papers usually has about 20 times more references than papers.

Paper-author Yule process

Figure 3 shows a diagram of the basic bipartite Yule process modified for the characteristics of paper-author networks (Goldstein et al., 2004; Morris et al., 2004). In this case the Yule process is applied to

teams of researchers rather than individual researchers. As each paper is added, there is a probability α that the paper will be authored by a new research team. If so, a team of N_G authors is added to the network, but only $N(\lambda)$ appear as authors of the team's first paper, where $N(\lambda)$ is a random deviate drawn from a 1-shifted Poisson distribution whose parameter is λ. If choosing an existing team, the team is chosen using preferential attachment, that is, the probability that a team will author the new paper is proportional to the number of papers that the team has previously published.

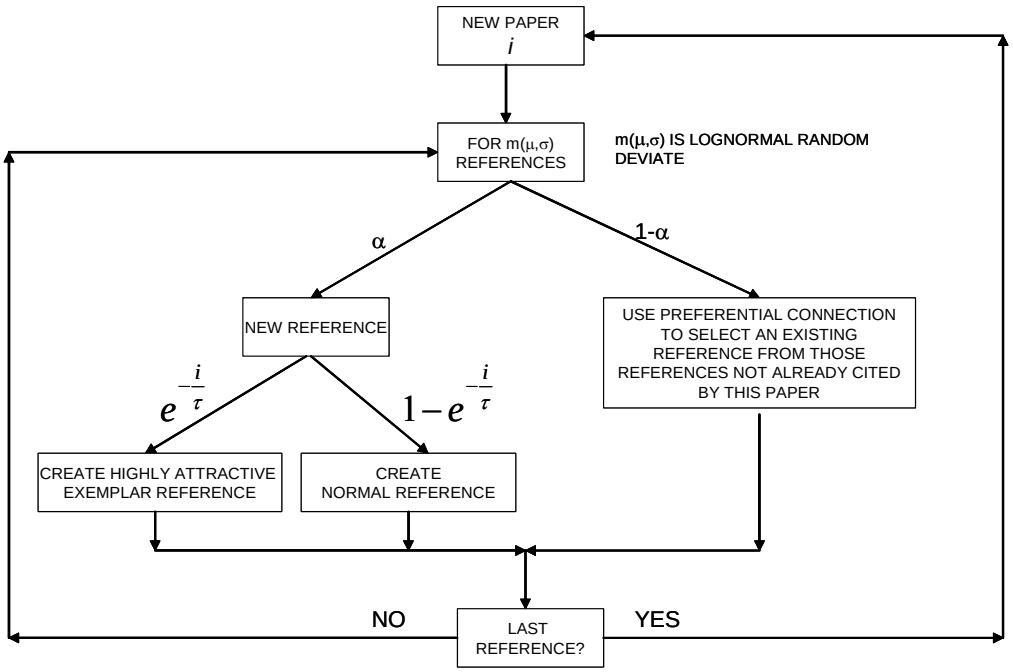

Figure 2. Diagram of a bipartite Yule process for paper-reference networks

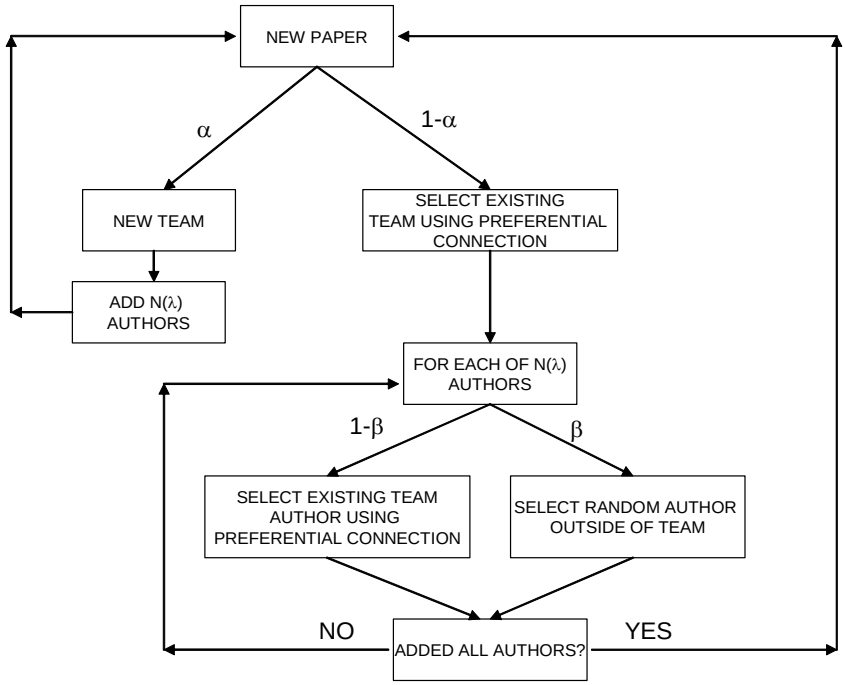

Figure 3. Diagram of a bipartite Yule process for paper-author networks.

When selecting authors for a paper from within an existing team, $N(\lambda)$ authors are chosen and the authors are selected using preferential attachment, that is, the probability of selecting an author is proportional to 1 plus the number of papers that the author has published. Inter-team collaborations

(weak ties) are modeled as random events, when an existing author is to be selected, there is a probability β that the author will be drawn randomly from some other team.

Results

A comparison of simulations to actual bipartite networks from collections of journal papers demonstrates the use of bipartite Yule models. The simulations were performed by using a MATLAB program to execute the algorithms of Figure 2 and Figure 3 for paper-reference networks and author-paper networks respectively. These simulations were used to build adjacency matrices of simulated networks which were compared to actual networks.

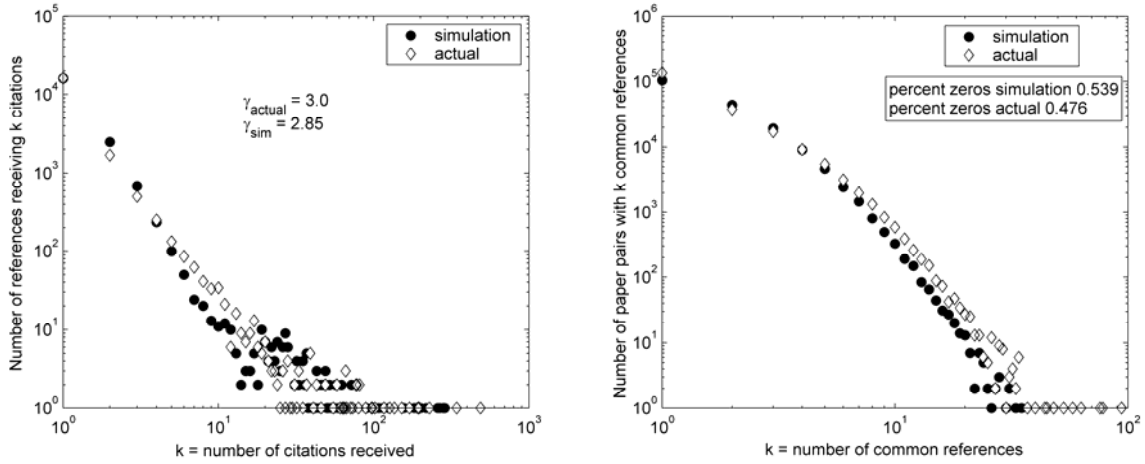

Figure 4. Comparison plots of paper per reference frequency and bibliographic coupling strength frequency from the complex networks paper collection.

Example simulation of paper-reference network

The Yule model for paper-reference networks was tested on a collection of papers that cover the topic of complex networks. This collection was gathered on September 8th, 2003 from ISI's Web of Science product using a series of queries to find all papers that cite key references and authors in the specialty. The collection contains 902 papers with 31355 citations to 19185 references. The Yule parameter, estimated by dividing the number of references by the number of citations to references, is 0.61. The mean references per paper is 34.8. The parameters used for the bipartite simulation of this paper-reference network can be found in (Morris, 2004).

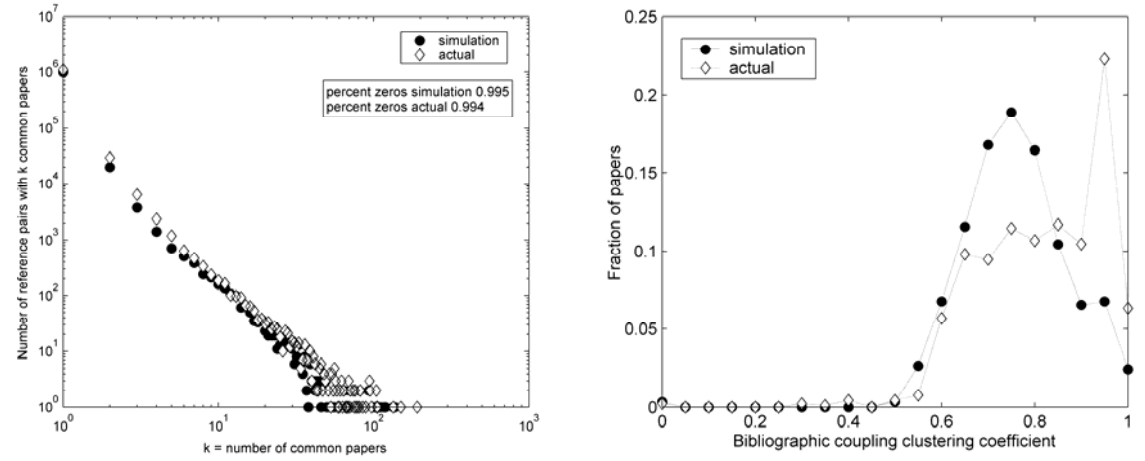

Figure 5. Comparison plots of co-citation strength frequency and bibliographic coupling clustering coefficient distribution from the complex networks paper collection.

Figure 4 and Figure 5 show plots comparing network metrics from the actual data to a Yule simulation of network growth. On the left of Figure 4 is a plot of papers per reference frequencies. Maximum likelihood expectation (MLE) estimated power-law exponents are 3.0 for the actual frequencies, and

2.85 for the simulation. The paper-reference Yule process mimics the phenomenon of exceptionally highly cited exemplar references in the extreme lower right of the plot. On the right of Figure 4 is a plot of frequency of bibliographic coupling strength per paper pair. The Yule process-based simulation frequencies match the actual frequencies well. The series of high bibliographic coupling strength pairs in the lower right from actual data corresponds to pairs of review papers with long lists of almost identical references, a phenomenon not modeled by the Yule process. On the left of Figure 5 is a plot of frequency of co-citation strength per reference pair. The simulated frequencies match the actual frequencies well across the whole plot. On the right is a plot of bibliographic coupling clustering coefficient distribution. The simulated distribution matches the shape and scale of the actual data.

Example simulation of a paper-author network.
The Yule model for paper-author networks was tested on three collections of papers representing specialties with a wide range of collaboration intensities. A collection of 1391 papers on the topic of distance learning with 51% single-authored papers represented a specialty with little collaboration. A collection of 900 papers on the topic of complex networks with 21% single-authored papers represented a specialty with typical amount of collaboration. Finally, a collection of 3095 papers on the topic of atrial ablation with 7% single-authored papers represented a specialty with heavy collaboration (Morris et al., 2004). The parameters used for the bipartite simulation of these author-paper networks can be found in (Morris, Goldstein & DeYong, 2004). Figure 6 shows the comparison of Yule model simulations to actual data for these three collections using two metrics: 1) paper per author frequency (Lotka's Law), and 2) collaborating author distribution.

Noting the paper per author frequency plots in the left column, the Yule process produces excellent matches to actual data. The inset plots show Yule model predicted paper per author distributions derived by gathering statistics from 1000 simulations for each collection. A line representing an MLE fitted zeta (pure power-law) distribution is shown in each inset. The Yule model produces excellent fits to the zeta distribution for all three collections, confirming its usefulness as a predictor of Lotka's Law. Note the deviation of the distributions from the zeta distribution in the tail of the distributions is due to truncating the simulations at the number of papers in each collection. Noting the plots in the right column, the Yule model produces good matches of collaborating author frequencies to actual data across the wide range of collaboration intensities represented by the three collections.

References

Chen, Y. S. (1994). The Simon-Yule approach to bibliometric modeling. *Information Processing & Management*, 30(4), 535-556.

Goldstein, M. L., Morris, S. A., & Yen, G. G. (2005). Group-based Yule model for bipartite author-paper networks. *Physical Review E*, 71, 026108.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate discrete distributions (2nd ed.)*. New York: John Wiley & Sons.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317-323.

McCain, K. W. (1991). Mapping economics through the journal literature: an experiment in journal cocitation analysis. *Journal of the American Society for Information Science, 42*(4), 290-296.

Morris, S. A. (2004). 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,* in print.

Morris, S. A. (2005). *Unified mathematical treatment of complex cascaded bipartite networks: the case of collections of journal papers.* Unpublished Dissertation, Oklahoma State University, Stillwater, Oklahoma, U.S.A.

Morris, S. A., Goldstein, M. L., & Deyong, C. F. (2004). Manifestation of research teams in journal literature: A growth model of papers, authors, collaboration, coauthorship, weak ties, and Lotka's Law. *submitted to Journal of the American Society for Information Science and Technology*..

Naranan, S. (1971). Power law relations in science bibliography- a self-consistent interpretation. *Journal of Documentation, 27*(2), 83-97.

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.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42, 425-440.

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.

White, H. D., & Griffith, B. C. (1981). Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32(3), 163-172.

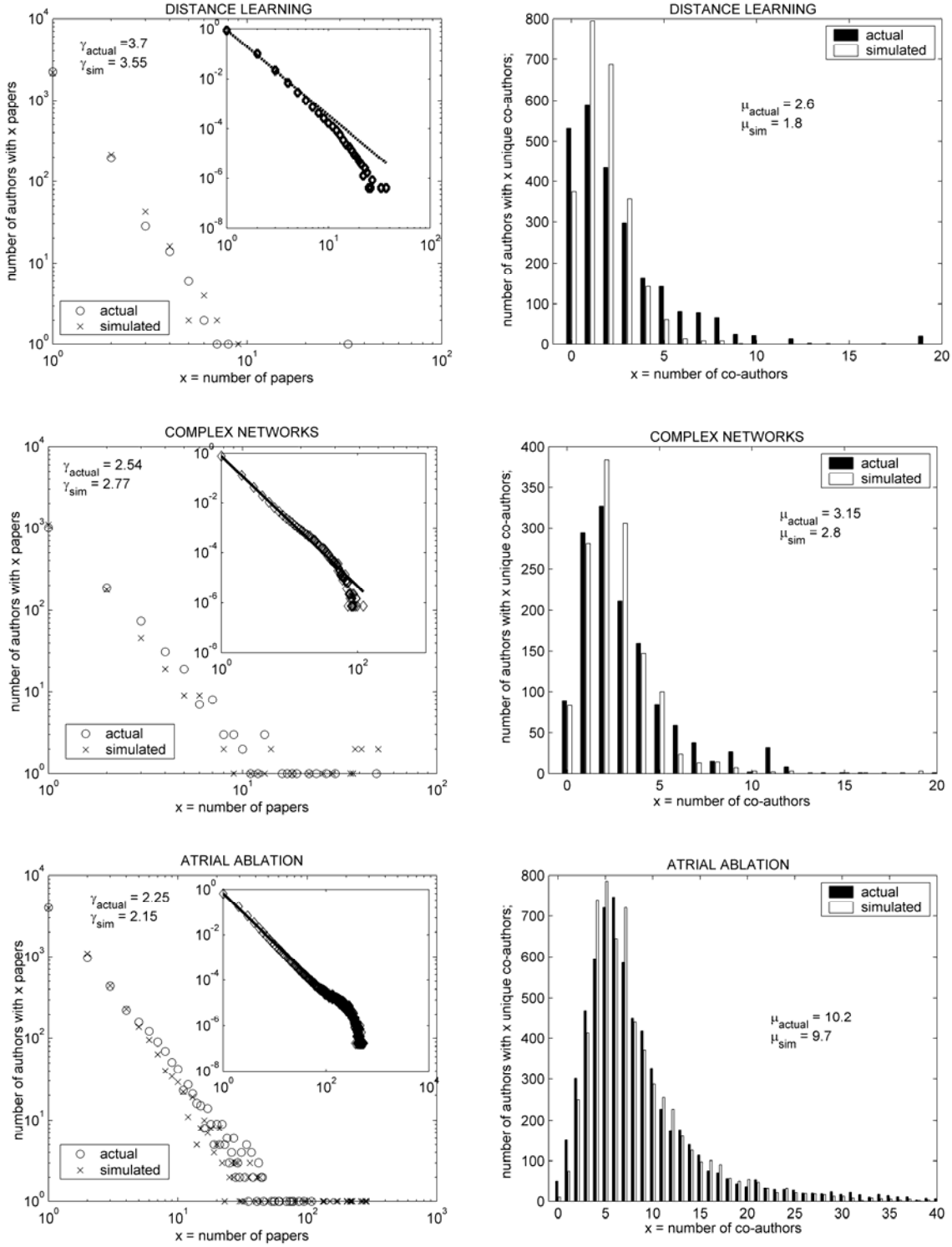

Figure 6. Comparison plots of paper per author frequencies and collaborating author frequencies from three paper collections representing a range of collaboration intensities.