ASIST SIG/MET presentation

An Exploratory Study on Co-word Network Simulation

ABSTRACT

Simulation is a feasible way of understanding regularities in scholarly communication. Previous studies have not explored the simulation of co-word networks. This study attempts to simulate co-word networks according to the dynamics of research fields. Two strategies of keyword selection for papers, random selection and preferential attachment, were compared. Results show that preferential attachment generates co-word networks that are more similar to empirical networks than random selection does. The comparisons between simulation and empirical networks also shed light on the keyword selection and growth in a research field. The findings from this study contribute to methods for co-word network simulation which facilitates the study of underlying mechanisms of the intellectual structure of sciences.

INTRODUCTION

Simulating how science works plays an important role in disclosing the mechanisms of science (Price, 1956). Simulation methods can overcome the limitation of insufficient amount of empirical data, and allow easy parameter adjustments to reveal patterns and regularities that empirical data does not cover. This advantage allows simulation studies to have better generalizability than empirical studies that usually confine their findings to the limited samples observed. Simulation approaches have been applied to collaboration and author productivity (Morris, & Goldstein, 2007), and citation distribution (Goldberg, et al, 2015). However, the simulation of co-word networks has not been discussed in the literature.

Co-word networks have been widely used to reveal themes, structures and development of a field. Studies on co-word networks contribute to our understanding of the intellectual structure of sciences. From the perspective of the dynamics of science, the co-word network of a research field grows as papers are published in the field. The purpose of this study is to explore how to simulate co-word networks according to this process. A generative process of co-word networks is introduced with two strategies of keyword selection for papers. Based on the comparison of the simulated and empirical co-word networks, this study aims to verify the simulation process and shed light on the mechanisms of keyword selection for papers and keyword growth in a research field.

DATA AND METHOD

Empirical Data

Empirical data were collected from three fields, including Information Science & Library (LIS), Sociology (Socio), and Physics, Fluids & Plasma (Phys). Articles from the top 20 journals with the highest impact factors in the three fields for the period January 2006 to December 2015 were downloaded from Web of Science. This resulted in 14,048, 11,978 and 65,603 articles for the three fields, respectively. The KeyWords Plus field of the bibliographic records was used to generate empirical co-word networks. After removing records without KeyWord Plus field, 11,530, 7,166 and 61,301 articles remained for the three fields. The keywords from the KeyWord Plus field of these articles were then used to generate empirical co-word networks. Other term sources, such as author keywords or title words, can also be used. However, there are high percentages of missing data for author keywords, ranging from 42% to 69%, and title words are free-text with many variations.

Simulation Process

The generative process of a co-word network ties to the generation of papers and keyword assignments for the papers. Four iterative steps are followed:

1) Generate a paper p_i incrementally. Since we are only interested in keywords of p_i, not its content, here p_i is just a sequential number.

2) Determine the number of keywords to be assigned to p_i, μ, according to the probability distribution of the number of keywords $\boldsymbol{\gamma}$ observed in empirical data. In this study, the probability distribution $\boldsymbol{\gamma}$ is set as:

$$\mathrm{p}(\mu|\boldsymbol{\gamma}) = \begin{cases} \frac{1-\delta}{9}, & \mu = 1,2,\cdots,9 \\ \delta, & \mu = 10 \end{cases} \qquad (1)$$

where μ is the number of keywords in the paper and δ is the probability that the paper has 10 keywords ($\delta = 0.36$). This is based on the observation that a majority of papers are assigned with 10 keywords and the rest share roughly equal probability of having from 1 to 9 keywords (Table 1).

3) Select each of the keywords for the paper. A keyword for a paper is either selected from existing keywords with a probability of 1-*a* or generated a new keyword with a probability of *a* (damping factor). We set *a* to 0.17, which controls the growth rate of new keywords. Two strategies of keyword selection were compared: random selection (RS), i.e., to pick a keyword uniformly; and preferential attachment (PA) (Simon, 1955), i.e., to select with a probability proportional to the frequency of the keyword. It should be noted that preferential attachment, which can produce power law distributions for item occurrences (Mitzenmacher, 2004), is based on empirical observations that keyword frequency distribution generally follows power law distributions (Liu et al., 2014; Zhang et al.,2008).

$$P(k) = (1 - \alpha)\frac{n_k}{\iota} \qquad (2)$$

where n_k is the current frequency of the keyword *k* and ι is the total frequency of all keywords in the collection.

4) Update the co-word network according to the newly assigned keywords.

The sizes of the simulated networks are controlled by the number of articles. Comparable number of articles are generated for comparisons with empirical data. In total, six simulation networks were produced by following the two strategies.

RESULTS AND DISCUSSION

Figure 1 provides an example of the keyword frequency distributions of the empirical network in the LIS field and the corresponding simulation networks. It shows that PA generates more similar distribution to the empirical network than RS does. Similar situation is observed in the other two fields. Then, seven common network metrics are selected to compare the generated simulation networks with empirical ones (Table 2). The two types of simulation networks have similar numbers of nodes (nn), which can be explained by the same damping factor (α=0.17) that determines the probability of generating new keywords. It is noted that the nn of smaller simulation networks (e.g. LIS and Socio) are a bit smaller than those of the empirical networks, while the nn of larger simulation networks (Phys) is much greater. This suggests the growth rate of new keywords in empirical networks may depend on network size. Our damping factor underestimates the growth rate for smaller networks and overestimates for the larger network. Different strategies of keyword selection account for the differences in edge-related metrics. RS networks have more edges, higher density, average degree and Clustering coefficient, larger diameter and average distance than PA networks. This is because RS selects keywords uniformly, which more likely creates edges between keywords. Small PA networks (LIS, and Socio) have similar number of edges as the corresponding empirical networks, while large PA network has more edges. However, all RS networks overestimate the number of edges. In general, the differences between PA and empirical networks are smaller than those between RS and empirical networks. This suggests PA is a better simulation method than RS for co-word networks.

CONCLUSION

This study explores simulating methods for co-word networks according to a generative process of papers and keywords. Two strategies of keyword selection were compared. Evidence showed that preferential attachment is better than random selection. The growth of keyword in co-word networks appears to depend on network size. Future research will further explore other factors, such as keyword decaying and other network dynamics rules, in simulating these observed phenomena.

REFERENCES

Goldberg, S. R., Anthony, H., & Evans, T. S. (2015). Modelling citation networks. *Scientometrics*, *105*(3), 1577-1604.

Liu, L., Qi, X., Xue, J., & Xie, M. (2014). A topology construct and control model with small-world and scale-free concepts for heterogeneous sensor networks. *International Journal of Distributed Sensor Networks*, *10*(3), 374251.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, *1*(2), 226-251.

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

Price, D. J. (1956). The exponential curve of science. *Discovery*, *17*(6), 240-243.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425-440.

Zhang, Z. K., Lü, L., Liu, J. G., & Zhou, T. (2008). Empirical analysis on a keyword-based semantic system. *The European Physical Journal B*, *66*(4), 557-561.

Table 1. Frequency distribution of the number of keywords in the empirical data.

Information Science & Library		Sociology		Physics, Fluids & Plasma	
# of keywords	# of papers	# of keywords	# of papers	# of keywords	# of papers
1	985	1	536	1	3,582
2	985	2	476	2	4,722
3	910	3	495	3	5,455
4	860	4	469	4	5,860
5	830	5	456	5	5,781
6	760	6	400	6	5,379
7	649	7	399	7	4,877
8	584	8	355	8	4,286
9	523	9	356	9	3,677
10	4,444	10	3,224	10	17,682

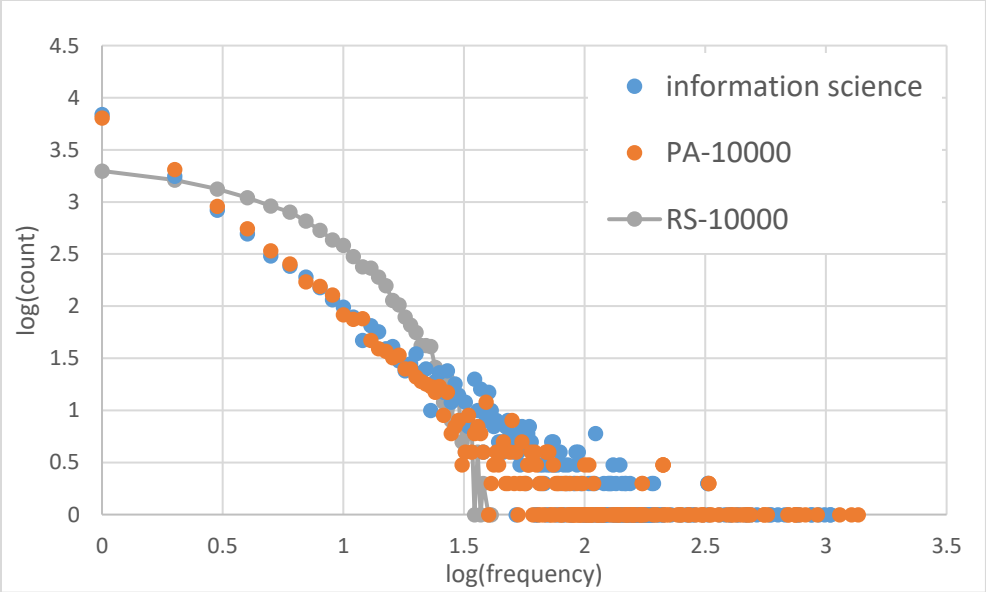

Figure 1. Frequency distribution of the empirical and simulation networks in LIS field.

Table 2. Comparisons between the simulation networks and empirical networks.

Attributes	Socio	PA	RS	LIS	PA	RS	Phys	PA	RS
# of articles	7,238	7,500	7,500	11,530	10,000	10,000	61,301	60,000	60,000
# of nodes (nn)	10,499	8,825	8,840	12,165	11,832	11,578	50,106	70,032	69,446
# of edges (ne)	146,966	131,564	183,296	182,449	181,037	244,724	841,627	1,151,422	1,481,929
Density	0.0027	0.0034	0.0047	0.0025	0.0026	0.0037	0.0007	0.0005	0.0006
Average degree	27.996	29.816	41.470	29.996	30.601	42.274	33.594	32.883	42.679
Clustering Coefficient	0.11081	0.08082	0.12208	0.09825	0.07486	0.11640	0.05688	0.03018	0.09689
Diameter	6	5	6	6	5	7	7	6	8
Average distance	2.81495	2.53248	2.94505	2.76602	2.58106	3.00297	2.84719	2.7216	3.4298