Do We Need Global and Local Knowledge of the Citation Network?

S.R. Goldberg[1], H. Anthony and T.S. Evans[2]

[1] *s.r.goldberg@qmul.ac.uk*
Queen Mary University of London, School of Physics and Astronomy, London, E1 4NS, (U.K.)

[2] *t.evans@imperial.ac.uk*
Imperial College London, Centre for Complexity Science & Physics Department, London, SW7 2AZ, (U.K.)

Introduction

Models which reproduce key features of the distribution citations to academic papers have a long history (Price, 1965). One aim is to illustrate if certain simple processes can explain important features. In this paper we focus on the fact that the distribution of citations for papers of a similar age scales primarily with the average number of citations (Radicchi, Fortunato, & Castellano, 2008; Evans, Hopkins & Kaube, 2012), with the shape otherwise largely invariant. In particular the width shows no temporal evolution. Simple multiplicative processes or basic models such as the Price model (Price, 1965) give dramatically different results, typically the distributions become narrower over time. The purpose of this study is to find a simple model which can lead to the observed behaviour of citations over time.

Methods

Consider a set of *N* papers all published in one year with an average number of citations *C*. We take 'reasonably well cited' papers with *c>0.1C* and following Evans, Hopkins and Kaube (2012) we fit the number of papers with *c* citations to a log-normal distribution

$$\frac{n(c)}{N} = \int_{c-0.5}^{c+0.5} \frac{dx}{\sqrt{2\pi}\sigma x} exp\left\{-\frac{(\ln(\mathrm{x}/C)+\sigma^2)^2}{2\sigma^2}\right\}$$

The log-normal form is an effective description and our only interest here is that the σ parameter is a reasonable characterisation of the width of the distribution. We want to find a model which has the correct properties for this width, namely it is roughly constant over time and of the right size. We compare outputs from our models against measurements made on data from the citation network of the hep-th section of the arXiv repository (KDD cup 2003).

We tried three models. In model A, with probability *p* papers are cited in proportion to their current number of citations, Price's cumulative advantage (Price 1965), otherwise the papers cited are chosen uniformly at random. In model B both these probabilities are modified by a factor $\exp((N-t)/\tau)$ for paper number *(N+1)* where τ is a time scale parameter.

Models A and B are based purely on global information – knowledge of the whole network is required. This reflects authors discovering papers using mechanisms other than the bibliographies of papers.

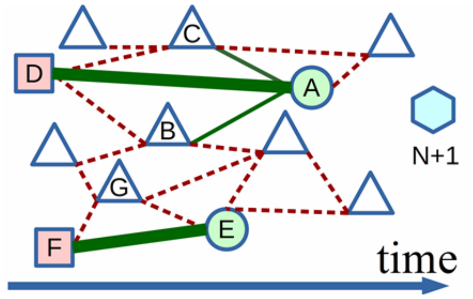

Figure 1. Illustration of Model C. A new paper (hexagon, *N+1*) is set to have four references. The first 'core' paper is chosen, A, using the global process of model B. Then with probability *q*, papers cited by A are also added to the new bibliography. Here B and C are considered (thin solid lines) but only D is added (thick line). The process continues until the required bibliography is complete. Here a second core paper E is chosen and one of its citations, F, is copied. At that point the process stops, paper G is never considered. The new bibliography is A, D, E and F.

For model C we add a second process, which uses only local information, see Figure 1. A set of 'core' papers are chosen as in model B. However each time a core paper is chosen, we examine each of the papers cited by this core paper and with probability *q* we add each to the new bibliography. This random walk from core papers to subsidiary papers is known to generate an effective cumulative attachment (Evans & Saramäki, 2005). In all cases we choose the length of the bibliography from a normal distribution with the same mean, 12.0, and standard deviation, 3.0, as measured in our hep-th data. The models involve a small number of parameters which have to be chosen. One feature we use is the number of zero cited papers and we match that to the proportion found in our results. We also look at the time it takes a paper in our model to reach half its final citations in order to find an optimal τ value. Finally parameter *q* in Model C is set by using an approximate form of transitive reduction (Clough et al., 2014) to estimate the faction of core papers in our data.

Results

Both our Models A and B produced long-tailed citation distributions but in both cases the width parameter σ was significantly smaller than that found in our data. However we were able to find a range of parameters where Model C was consistent with our data, for example see Figure 1. In particular the papers produced in one year had fat tails with a width σ which was roughly constant in time.

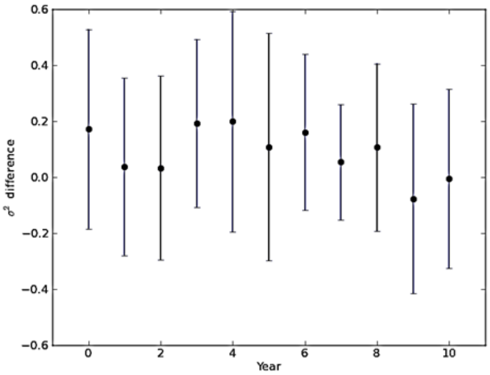

Figure 1. The difference between the width σ of the hep-th data and that found in our Model C for final fitted parameters.

Discussion

We started from the observation that the width of the fat-tailed citations distributions for papers published in one year show some consistent patterns. In particular, in terms of our log-normal width parameter, σ, this width is roughly constant and independent of the age of the papers studied. To keep our work rooted in real citations, we worked with hep-th arXiv data which also shows this characteristic static width.

The difficulty in finding a model which reproduces this key feature was illustrated by results from our first two models: Model A mixed cumulative and uniform random attachment while Model B added a time decay to favour citations to more recent papers. We were unable to find parameter regimes where these models provided good fits to our data.

However our model C with just three parameters was able to produce an accepted fit to the hep-th data over 11 different years, see Figure 1.

The big difference between model C and our earlier attempts is that only in model C was local information as well as global information used to find references for a new paper. We conclude that the citation patterns we see reflect a mixture of local searches of the citation network (reading papers and finding the papers they cite) along with global information providing the recommendation (a chance personal suggestion at a conference perhaps).

Another interesting result is that we find the best fits for our model to our data is when around 70% to 80% of papers cited are ‘subsidiary papers’, papers found from local searches through the bibliographies of other papers. Interestingly similar results have been found seen by Simkin and Roychowdhury (2005) who arrive at a similar model but for different reasons. Namely they suggested that mistakes in bibliographic entries suggest that around 80% of citations are copied (Simkin & Roychowdhury, 2003). In our terminology these would be citations to subsidiary papers so both sets of results are consistent. Further support for this result comes from the transitive reduction analysis of Clough et al. (2014)

Finally we suggest that more work needs to be done to capture the effect of the variation in the length of bibliographies. We used a normal distribution for this aspect. This encodes some fluctuations in this bibliography length, something usually neglected in other models, but the reference distribution should also be fat-tailed. We failed to get good agreement with data when we modelled bibliography length this way.

Acknowledgments

We would like to thank James Clough, James Gollings, and Tamar Loach for sharing their results on related projects.

References

Clough, J.R., Gollings, J., Loach, T.V. & Evans, T.S. (2014). Transitive reduction of citation networks *J. Complex Networks* (to appear) http://dx.doi.org/10.1093/comnet/cnu039.

Evans, T.S; Hopkins, N. & Kaube, B.S. (2012). Universality of Performance Indicators based on Citation and Reference Counts. *Scientometrics*, *93*, 473-495.

Evans, T.S. & Saramäki, J. (2005). Scale-free networks from self-organization. *Phys.Rev. E*, *72*, 026138.

KDD Cup (2003). Network mining and usage log analysis. Retrieved October 1, 2012 from http://www.cs.cornell.edu/projects/kddcup/datasets.html .

Radicchi, F., Fortunato, S. & Castellano, C. (2008). Universality of citation distributions: Towards an objective measure of scientific impact. *PNAS*, *105*, 17268-17272.

Goldberg, S.R., Anthony, H. & Evans, T.S. (2014). Modelling citation networks, Scientometrics (to appear) [*arXiv:*1408.2970].

Simkin M.V. & Roychowdhury V.P. (2003). Read before you cite! *Complex Systems*, *14*, 269-274.

Simkin M.V. & Roychowdhury V.P. (2005) Stochastic modeling of citation slips. *Scientometrics*, *62*, 367-384.